INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Kale to Action- Veggies' Behavior from Transactional Data: Case of Pingo Doce

Carolina de Canilho Pinto Ribeiro Ferreira

Master in Business Analytics

Supervisor:
Professor Nuno Santos, Invited Assistant Professor
ISCTE Business School

October, 2022

BUSINESS
SCHOOL

Department of Quantitative Methods for Management and Economics

Kale to Action- Veggies' Behavior from Transactional Data: Case of Pingo Doce

Carolina de Canilho Pinto Ribeiro Ferreira

Master in Business Analytics

Supervisor:
Professor Nuno Santos, Invited Assistant Professor,
ISCTE Business School

October, 2022

Acknowledgments

This thesis marks the culmination of one cycle and the beginning of a new one.

First, I want to thank my supervisor, Professor Nuno Santos, for all his help and guidance through this project and for all the tips I will keep for life.

To my teacher, Professor Raul Laureano, for all the support and availability throughout my master's degree.

To Tiago Marques and Miguel Pina, from Jerónimo Martins, for making this project possible, and for all the help and availability to answer questions related to this work and the business.

Also, many thanks to my family and friends, for always being there in some way, believing in me, and giving me the support to finish this cycle.

Finally, a special thanks to the unforgettable Encarnação.

Abstract

There has been an increase in the number of people shifting towards plant-based diets and a growing interest in the vegetarian and vegan products market. However, most studies conducted on these individuals (veggies) have used declarative data, missing a behavioral analysis. This can be overcome by using business analytics, which can be applied to a retail context.

Therefore, this project aimed to identify customer segments that are potentially following these diets and are interested in the plant-based product market, as well as to analyze their behavior and profile, using transactional data collected from a six-month period of a Portuguese retailer, Pingo Doce.

To do this, it was first necessary to create a new item's market structure, which included creating a new nomenclature to identify vegetarian and vegan items. Then, a cluster analysis was performed in terms of shopping baskets, using Two-Step Clustering, which allowed the identification of five segments of customers with different levels of interest in plant-based products: *Traditional Omnivores*, *Receptive Omnivore*s, *Convenience and Vegan Sweets*, *Potential Veggies*, and *Veggie Lovers*. Additionally, after their description, a deployment process was defined using a classification technique to create a set of rules to help the retailer classify the same customers, in another timeframe, and new customers, based on the clusters found. To finalize, recommendations were given to Pingo Doce to improve its business in the market of plant-based products and the level of engagement of these clusters with the retailer.

Keywords: plant-based diets, veggies, retail, cluster analysis, customer behavior, profiling

JEL Classification: D12, C38, L81, M31

Resumo

Verifica-se um aumento no número de pessoas que estão a adotar dietas baseadas em plantas e um crescente interesse no mercado dos produtos vegetarianos e veganos. Contudo, a maioria dos estudos realizados sobre estes indivíduos (*veggies*) têm utilizado dados declarativos, faltando-lhes uma análise comportamental. Isto pode ser ultrapassado através da análise de negócios, que pode ser aplicada ao retalho.

Portanto, este projeto visava identificar segmentos de clientes que estão potencialmente a seguir estas dietas e interessados no mercado de produtos vegetais, bem como analisar o seu comportamento e perfil, através da utilização de dados transacionais recolhidos de um período de seis meses de um retalhista português, o Pingo Doce.

Para tal, foi necessário criar primeiro uma nova estrutura mercadológica para os artigos, que incluiu a criação de uma nomenclatura para identificar os artigos vegetarianos e veganos. Depois, realizou-se uma análise de agrupamentos em termos de cabazes de compra, utilizando o *Two-Step Clustering*, permitindo a identificação de cinco segmentos de clientes com diferentes níveis de interesse em produtos vegetais: *Traditional Omnivores, Receptive Omnivores, Convenience and Vegan Sweets, Potential Veggies, and Veggie Lovers*. Além disso, definiu-se um processo de implementação, recorrendo a uma técnica de classificação, para criar um conjunto de regras para ajudar o retalhista a classificar os mesmos clientes, noutro período de tempo, e novos clientes, com base nos segmentos encontrados. Para finalizar, foram dadas recomendações ao Pingo Doce para melhorar o seu negócio no mercado de produtos vegetais e o nível de interação destes segmentos com o retalhista.

Palavras-chave: dietas baseadas em plantas, *veggies*, retalho, análise de agrupamentos, comportamento de clientes, perfilamento

Classificação JEL: D12, C38, L81, M31

Index

Figures Index

Tables Index

List of Abbreviations

CFA	Confirmatory Factor Analysis
EFA	Exploratory Factor Analysis
JM	Jerónimo Martins
LCA	Latent Class Analysis
LPA	Latent Profile Analysis
PCA	Principal Component Analysis
PD	Pingo Doce
SLR	Systematic Literature Review

1. Introduction

1.1. Context

Recently, there has been an increased interest in plant-based diets. The pandemic has nourished the growth of these diets, and it is estimated that if they continue to grow at the current rate, they may replace traditional diets within a century (Vegan Food and Living, 2021). In 2021, sales of vegetarian and vegan products from a German producer known for its animal meat sausages have already exceeded meat sales for the first time (Vegconomist, 2022).

Vegetarianism, which is considered a plant-based diet, is a dietary pattern characterized by the exclusion of meat and fish, although its by-products may also be consumed (Associação Vegetariana Portuguesa, 2021). Individuals who follow this eating pattern, are referred to as *vegetarians*. This is an umbrella term because, depending on the inclusion or exclusion of the consumption of animal by-products, different designations are used to identify them. The most common designations for a vegetarian are *ovo-lacto-vegetarian* (which consumes eggs and dairy products), *ovo-vegetarian* (which includes eggs but excludes dairy products), *lacto-vegetarian* (which includes dairy products but excludes eggs), and, at the extreme, *vegan* (which eliminates any animal products and their derivatives) (Associação Vegetariana Portuguesa, 2021; Healthline, 2017).

According to a second study from Nielsen, between 2007 and 2017 the number of Portuguese vegetarians (lacto-/ovo-/ovo-lacto-vegetarians) quadrupled, and the number of vegans was double the number of vegetarians in the 2007 study (Centro Vegetariano, 2020). However, another group has emerged with great weight, the *flexitarians*. These follow a less strict diet that allows the occasional consumption of meat and fish. In 2021, the veggie community, which includes vegans, vegetarians, and flexitarians, already had more than 1 million Portuguese people over the age of 18, representing 11.9% of the Portuguese adult population (0.5% vegan, 2.1% vegetarian, and 9.3% flexitarian). This corresponds to an increase of 34% compared to 2019, explained especially by the increase in vegetarians (+137%) and flexitarians (+27%) (Lantern, 2021). To facilitate the understanding of the various designations in this research, from here onwards all these individuals who seek to reduce or eliminate the consumption of animal products will be referred to as *veggies.*

1.2. Problems

The challenge of keeping up with these new trends arises and is no exception for retailers. Pingo Doce, which is an insignia belonging to the Jerónimo Martins Group and a reference in food retail, has only recently begun to study the customers who seek to follow plant-based diets. This is partly due to not having any categorization of vegetarian and vegan items that allow it. In a competitive market, it is crucial to be on top of these trends to be in competitive advantage and maintain brand loyalty. Not having good knowledge about these individuals may threaten their business in a growing market. According to Lantern (2021), of the major retailers, Pingo Doce was the fifth retailer that veggies identified to be the best place to find plant-based products, with 28% of veggies considering it to be one of the best.

Additionally, as a topic of growing relevance, the literature related to the analysis of veggie consumers has also increased, with studies being carried out, namely, on the motivations of these individuals (Verain et al., 2022; de Koning et al., 2020; Hielkema & Lund, 2021), the frequency of meat consumption or plant-based alternatives (Grasso et al., 2021; Szejda, et al, 2021) or how some plant-based substitute products are perceived (Sucapane, et al., 2022; Bryant et al., 2019). However, these investigations have two major limitations: the method of data collection and selective bias. Analyses regarding the behavior of people who follow these types of food are mostly declarative studies that are based on surveys or interviews (Niva & Vainio, 2021; Szejda et al.,2021; Koch et al., 2019). This means that what is obtained are only perceptions. In other words, this type of data collection has a certain error associated with it, as the answers given to these surveys may not always correspond to the observed reality. In addition, some studies' samples were not completely representative of the population (Culliford & Bradbury, 2020, Garnett et al., 2020, Gomez-Luciano et al., 2019a).

1.3. Motivations, Relevance, and Contributions

The motivations for this research arise from the fact that plant-based diets are an emerging topic, with good growth prospects, and to which more attention should be given because of the positive impact they can bring. In fact, the importance and visibility that is intended to be given to this topic and the incentive that is intended to be created around the demand for plant-based items meets, directly or indirectly, some Sustainable Development Goals, for instance, *SDG 8- Decent Work and Economic Growth, SDG 12- Responsible production and consumption, SDG 13-Climate action*, among others (Plant-Based Foods Association, 2021).

However, there is currently a gap in the literature on this subject that is verified by the lack of studies using transactional rather than declarative data, and with such a voluminous dataset that allows, more than stating, to observe the behaviors of veggie consumers and characterize them. By using data from one of the largest retail chains operating in Portugal, Pingo Doce, and because it has a loyalty program through the existence of a customer card, there is access to demographic and transactional data from millions of Portuguese individuals.

Therefore, this work will be extremely important for Pingo Doce. Firstly, it will get a market structure adapted to vegetarian and vegan products, which does not currently exist in the company. Secondly, the retailer will receive insights about veggie customers' profiles and behaviors. Consequently, it will enable more effective marketing campaigns and might help the business in decision-making. Their strategic actions may result in a higher satisfaction of customers who seek plant-based products because their needs might be better met, and on a higher level, this may lead to the increase of customer loyalty and sales.

Besides these reasons, this research will also have an impact on the scientific community, namely in Portugal. A study using historical purchasing data from millions of individuals will provide a more realistic view of these consumers who seek to reduce or eliminate animal consumption. Moreover, new knowledge can be acquired, some theories of studies already carried out can be verified (e.g., what they consume/purchase and what is their profile) and this work may serve as a reference for similar studies to be performed in other countries.

1.4 Research Question and Objectives

Considering the problems encountered and what was intended to be investigated, the research question that was raised was as follows:

R.Q. How to generate value in the plant-based food products market?

The main goal was to acquire better knowledge about veggie customers and their shopping patterns, for a period of six months (from September 1, 2021, to February 28, 2022), to generate valuable conclusions and recommendations for Pingo Doce's strategic decisions. For this business objective to be achieved, the main analytical objective consisted in identifying segments and characterizing veggies' customer profiles and their buying patterns, through an analytical technique, namely a clustering technique. More specifically, the following objectives were intended to be met:

O.1. Create a market structure for vegetarian and vegan items.

O.2. Segment customers (to identify them).

O.3. Characterize segments (to know what they buy and what is their profile in terms of demographic characteristics, lifestyle, and average spend vs. frequency).

O.4. Define strategies for each segment to increase frequency or spending.

In parallel, as mentioned in motivations, the goal is also to contribute to the increase of knowledge in the literature on this subject.

1.5 The Company

Jerónimo Martins (JM) Group was founded in 1792 and is headquartered in Portugal. It is also present in Poland and Colombia, mainly operating in the sectors of food distribution and specialized retail. In 2021, the Group with more than 120 thousand employees generated 20.889 thousand million euros (JM, 2022b). According to Deloitte's Global Powers of Retailing 2022 report, the Group is the 18th largest food retailer in Europe and the 32nd in the world (Jerónimo Martins, 2022a).

In the food distribution sector, the Group owns in Poland the Biedronka chain, which is responsible for the majority of the Group's profits. In Portugal, the Group owns the Pingo Doce chain, competing in the supermarket segment, and Recheio (whose target is mainly the Hotel, Restaurant, and Cafeteria sectors), competing in the cash & carry market. In Columbia, the Group is present through a chain of proximity shops, designated as Ara.

In the specialized retail activity, in Poland, the Group owns the Hebe chain (specialized in health and beauty products) and, in Portugal, it owns the Jerónimo coffee shops and the Hussel shops (specialized in chocolates and confectionery).

In the present thesis, the research focuses on the Portuguese food retail market, and more precisely, on Pingo Doce (PD). This chain, which has over 40 years of history and more than 460 stores, generated 4.046 thousand million euros in 2021 (Jerónimo Martins, 2022a).

2. Literature Review

Plant-based diets is a generic term that refers to diets that consist mostly of foods derived from plants, and that can be followed healthily with the inclusion of vegetables, legumes, fruits, nuts, seeds, and whole grains or in an unhealthy way with the inclusion of more processed plant-based products (Gibbs & Cappuccio, 2022). Those who follow a plant-based diet seek to reduce or, in the extreme, eliminate their consumption of animal products. Vegetarianism covers a range of eating patterns, ranging from individuals who choose to eliminate only animal products (ovo-lacto-vegetarians), to those who remove any animal products and their derivatives from their diet (vegans). However, some authors also include in this spectrum *pesco-vegetarians* (or pescatarians), those who avoid meat but consume fish and other shellfish, and *flexitarians*, also known as *semi-vegetarians*, who consume meat and fish sporadically or even once a week (Hargreaves et al., 2021, World Health Organization, 2021). Although these last two types of individuals might not be considered vegetarians (Healthline, 2017), they can all be included in the definition of individuals who follow plant-based diets.

Nevertheless, it is also important to highlight that following a vegan diet is different from *veganism*, another concept that emerges related to this topic. Veganism can be defined as a philosophy or a lifestyle, marked by the non-use of animal products, whether in food, clothing, cosmetics, or other products and materials (Hargreaves et al., 2021; Associação Vegetariana Portuguesa, 2021). Therefore, while vegetarianism corresponds to the practice of following a particular type of diet, veganism corresponds to a way of living that goes beyond food issues.

Reasons for people following a plant-based diet are varied, but three stand out among the Western population: health concerns (which are the most common among the general population), the environment, and animal rights (Hopwood et al., 2021). Individuals who adopt vegetarianism, mainly for ethical reasons and animal welfare, may naturally choose to not consume items that contain any ingredient that has been involved in the death or suffering of animals, such as cheeses that contain rennet of animal origin (which comes from the stomach of animals), gelatin (which is produced from animal skins, tendons, and bones), dyes, animal fats, among others. However, in the study conducted by Lantern (2021), the main reason for Portuguese flexitarians to follow this type of diet is health concerns, while the reason for vegetarians and vegans is the concern for the environment.

In fact, if a plant-based diet is followed healthily, with few processed products, it will be more beneficial to the health of the individual. As seen by Gibbs & Cappuccio (2022), a healthy

plant-based diet has been associated with a lower risk of developing cardiovascular diseases and all-cause mortality, since it may help with weight loss, preventing and treating type 2 diabetes, reducing blood pressure, among other benefits. In addition, the same authors demonstrate that the shift from a meat-based diet to a plant-based diet will be better for the environment as it requires less land and water resources and can reduce greenhouse gas emissions and water pollution. For example, to produce 1 kilogram of beef takes more than 15 thousand liters of water, while to produce 1 kilogram of nuts and pulses, only needs around 9 thousand and 4 thousand liters, respectively (Marie, 2022).

To extract more relevant information from the literature about individuals who seek to follow a plant-based diet (veggies), a systematic literature review (SLR) was conducted. Thus, the methodology followed in the SLR, and its results, are presented in subsequent sections.

2.1. Systematic Review Methodology- Veggies

The purpose of this SLR, which was based on the PRISMA methodology (Page et al., 2021) and carried out from a set of scientific articles in the areas of vegetarianism, behavioral analysis, and data analysis, was to identify how these analyses were conducted and to gain insights about patterns and characteristics already studied about veggies.

More specifically, it was intended to answer the following questions:

1. What was the methodology followed by the study?
2. What are the consumption patterns associated with veggies?
4. What are the purchasing patterns associated with veggies?
5. What is the profile of veggie consumers?

The research strategy followed by the SLR is summarized in Figure 2.1. Initially, the query used to search articles resulted from an iterative process that required the investigation of synonymous terms in different sources, as was the case of a previously conducted systematic review on vegetarianism and consumer behavior (Onwezen et al., 2021). The formulation of the query also involved the review and opinion of experts in the field of vegetarianism, namely from Associação Vegetariana Portuguesa and the European Vegetarian Union. The source of data used in this SLR was the scientific database Web of Science, as it is a widely used, reliable, and the oldest platform to search scientific publications (Birkle et al., 2020). In turn, the final query was applied only to the topic (a designation that represents the set formed by the title, abstract, and keywords of a document) of each document in this database. Next, selection

criteria were used and the documents that were not part of these criteria were excluded. Appendix A contains the query used and the selection criteria for the articles. Lastly, after the selection process, 35 articles were selected and are shown in Appendix B.

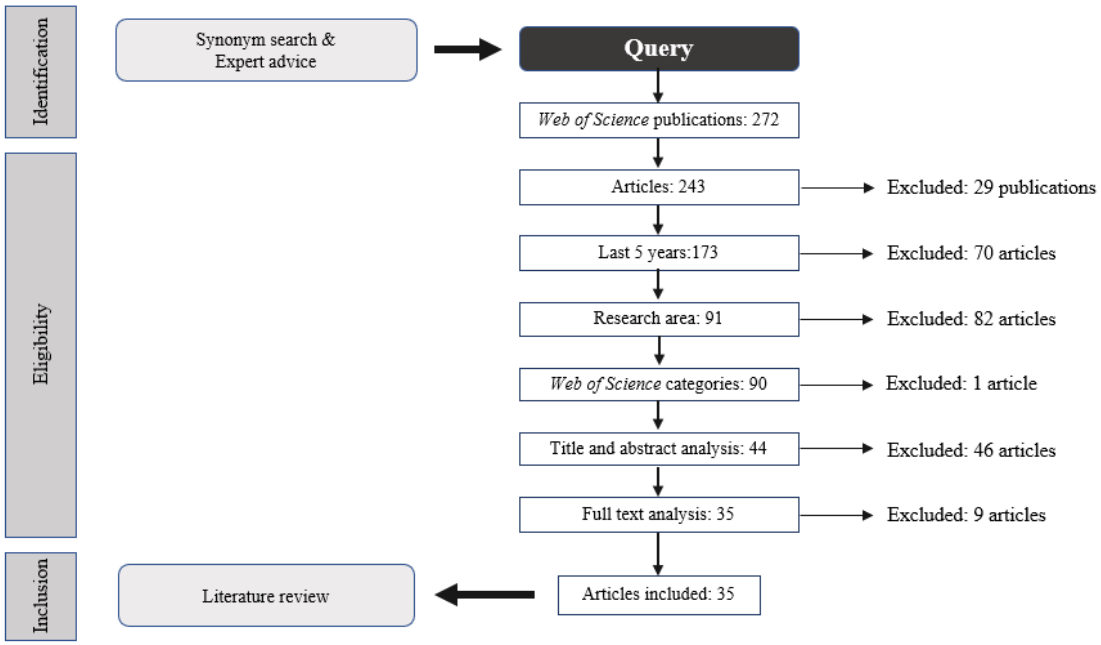

Figure 2.1: Article Selection Process.
Source: Elaborated by the author

Before reviewing the articles, the VOSviewer tool was used to identify which terms appeared most frequently in the abstract and title of these articles. Since it was required that each concept should occur at least 10 times in all the documents, a value that was by default in the software, the result was 24 terms. Of these, the terms *study* and *participant* were removed because these are words that, by their nature, occur frequently in scientific articles and because they did not contribute with relevant information to the topic.

In this way, as shown in Figure 2.2 four clusters were identified. The first cluster, represented by the red color, contained words such as *plant*, *protein*, *product*, *alternative*, *insect*, and *cultured meat*. The second cluster, in green, displayed terms such as *consumer*, *diet*, *meat*, *meat-reducer*, and *vegetarian*. The third cluster, in blue, gathered terms related to *consumption, barriers, and drivers*. Finally, the last one, in yellow, contained the term *product* This figure served as an indication that the articles selected for the SLR were more related to consumers who reduced meat or were vegetarians (green cluster), with alternative products to meat (red and yellow clusters) and with drivers and barriers in consumption patterns (blue cluster).

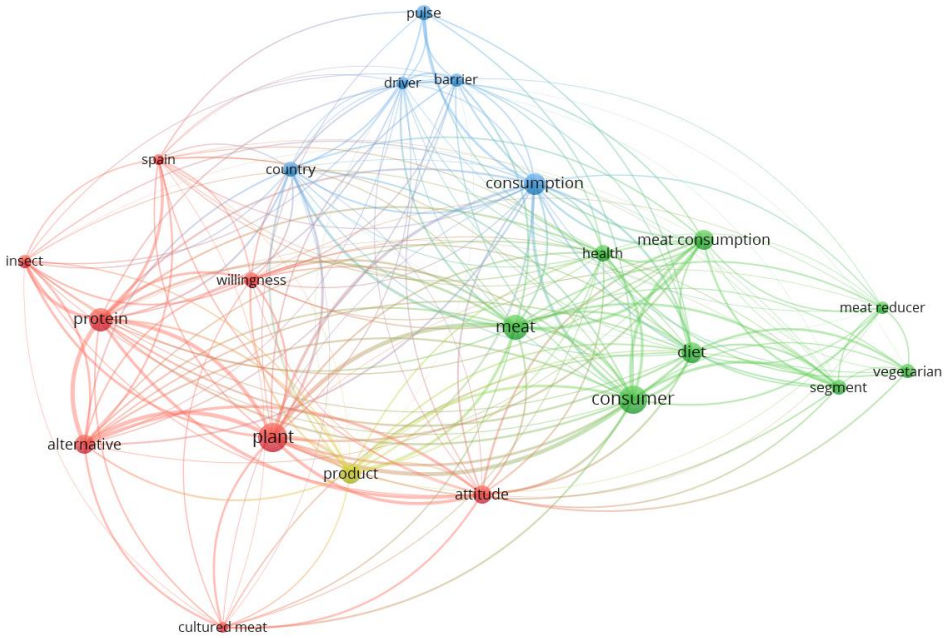

Figure 2.2: Network visualization: - terms in title and abstract
Source: Obtained from VOSViwer

2.2. Results

2.2.1. Methodology of studies on veggies

Appendix C provides information on the methodology followed by the various authors, such as: how data were obtained, sample size, and others. Most of the analyzed articles conducted studies based on online surveys (Verain et al., 2022; Niva & Vainio, 2021; Culliford & Bradbury, 2020) or questionnaires answered in the presence of the interviewer himself (Bullock et al., 2020, Gomez-Luciano et al., 2019b). However, other authors have also used interviews, either face-to-face (Apostolidis & McLeay, 2019) or by telephone (Koch et al., 2019). Yet, it should be noted that studies that are based on questionnaires or interviews, i.e., that use declarative data, do not always reflect people's real behavior. According to Malek & Umberger (2021), the respondents who self-identified as vegetarians and vegans did not show behaviors that strictly corresponded to the type of diet they claimed to follow. In other words, only 52.3% of those who identified themselves as vegans reported never having eaten animal products in the last year, and only 42% of those who identified themselves as vegetarians reported never having eaten meat or fish in the previous year. On the other hand, this data indicates that, although there are different types of diets, people do not always follow them strictly and rigidly, which is a complex issue.

About the type of data used and taking into account that the form of obtaining them was mostly carried out through surveys, it was analyzed sociodemographic data, such as gender, age, education, and income, and other data such as consumption or purchasing habits, interests in certain diets or products, motivations, opinions, and lifestyles, provided by the participants themselves (Culliford & Bradbury, 2020; Bryant & Sanctorum, 2021; Grasso et al., 2019). Nevertheless, behavioral data that resulted from direct observation of choices (Sucapane et al., 2022; Bullock et al., 2020) and product transactions (Andersson & Nelander, 2021; Yang & Dharmasena, 2021) were also studied.

Four studies analyzed the sales of meals delivered in university canteens or cafeterias (Andersson & Nelander, 2021; Morris et al., 2020; Garnett et al., 2020; Garnett et al., 2019). However, the samples showed limited diversity, either because the participants were essentially represented by students or individuals who lived or worked near these cafeterias, or because three of these studies only focused on the sales of meals and not products. Only in the study conducted by Morris et al. (2020), 651 different items sold in a university were analyzed. However, the aim was to analyze common dietary patterns, so it did not focus on the purchase of vegetarian and vegan products and, based on the examples of items given, possibly there were not even other products directly related to vegetarianism like plant-based milk, vegan yogurts and vegan or vegetarian alternatives to meat and fish. Another study, which was conducted in the United States, used transactional data provided by the Nielsen company to analyze the purchase of plant-based milk alternative drinks (Yang & Dharmasen, 2021). Yet, although these data were from sales, the information that was acquired by Nielsen about households' purchases was provided by the families themselves, which may not be completely accurate.

In terms of the type of analysis performed and their respective techniques, several authors performed more descriptive analyses, such as analysis of frequencies, means, and standard deviations (Figueira et al., 2019) but also inferential and predictive analyses (Szejda et al., 2021; Thomas & Bryant, 2021;) in which they made use of, for instance, hypothesis tests (Bryant & Sanctorum, 2021; Culliford & Bradbury, 2020; Malek et al., 2019) linear regressions (Andersson & Nelander, 2021; Yang & Dharmasena, 2021) and logistic regressions (Henn et al., 2022; Pandey et al., 2021), which allowed to predict consumption or purchasing intentions or behaviors.

To identify subgroups in data, latent variable models were also used: latent class analysis (LCA) (Apostolidis & McLeay, 2019) and latent profile analysis (LPA) (Lacroix & Gifford, 2019). Factor analyses were also performed, both in an exploratory (exploratory factor analysis

[EFA]) (Henn et al., 2022; Cheah et al., 2020), and in a confirmatory way (confirmatory factor analysis [CFA]) (Martinelli & De Canio, 2021; de Koning et al., 2020). In addition, it was observed the use of another exploratory technique to reduce the dimensionality of the data, the principal components analysis (PCA) (Malek & Umberger, 2021).

Nevertheless, some authors also made use of Clustering techniques (Verain et al., 2022; Henn et al., 2022; Grasso et al., 2021). However, of the studies that used transactional data, only Morris et al., (2020), used clustering techniques and, more precisely, the K-Means algorithm, to find groups among the individuals analyzed. The remaining studies that analyzed sales data, either at the transaction level, i.e., treating each transaction independently of the others (Andersson & Nelander, 2021; Garnett et al., 2020), at the product level (Yang & Dharmasena, 2021), or both at the transaction and customer level (Garnett et al., 2019), used linear regression techniques.

2.2.2. Veggies' consumption patterns

There has been a growing trend in the consumption of plant-based products and the demand for new and more sustainable plant proteins is expected to continue to increase in the future (Niva & Vainio, 2021, Grasso et al., 2021, Contini et al., 2020). Some consumers are changing their diet to entirely plant-based diets (vegetarians and vegans), others are reducing their meat and fish consumption (flexitarians), and the rest, who despite still consuming animal products, are also increasing the consumption of vegetarian or vegan products (Martinelli & De Canio, 2021). From 2011 to 2019, the number of people from Germany who self-identified as flexitarians increased from 13% to 43% (Verain et al., 2022).

Compared to those who follow an omnivore's diet with no restrictions, those who reduce or eliminate meat consumption, consume dairy products less often, but more often include vegetables, nuts, seeds, legumes, cereals, plant-based milk alternatives, and plant proteins (Malek & Umberger, 2021, Koch et al., 2019). In agreement with these data, Niva and Vainio (2021) found that individuals who do not consume meat and who tend to increase their consumption of plant-based proteins, also consume fish and soy-based products more often.

In general, there is a consensus among authors that the sustainable protein source most preferred by people is plant-based, compared to other substitutes such as single-cell proteins (derived from organisms such as algae, yeasts, and fungi), cultured meat (which results from animal cells and the use of technological processes, without killing them) or insect-based proteins (de Koning et al, 2020, Grasso et al., 2019, Gomez-Luciano et al., 2019a, Slade, 2018).

Although pulses are one of the meat alternatives, not everyone sees them as one. In a study conducted by Figueira et al. (2019) in Australia, although about 37% of participants considered them as a protein source, only 11% considered pulses as an alternative to meat. Nevertheless, the consumption of these foods is not restricted to vegetarians and vegans. In the same study by Figueira et al. (2019), 93% of participants reported consuming legumes, although 16% of respondents reported following a vegetarian or vegan diet. Also, the pandemic generated by COVID-19 triggered a high interest from all consumers, regardless of their diet, leading to these foods even being sold out in supermarkets (Didinger & Thompson, 2021). Nevertheless, veggies are more likely to have a more frequent and varied consumption of pulses compared to omnivores (Henn et al., 2022).

Additionally, according to Culliford & Bradbury (2020), although around 85% of respondents in their study reported consuming plant-based proteins, such as nuts, seeds, beans, and lentils, at least once a week, over 50% reported consuming processed plant-based meat substitutes with equal frequency. These results indicate that consumers are choosing different alternatives or substitutes when replacing meat. In another study, 70% of the respondents considered that pre-cooked plant-based products, i.e., products derived from plants that represent meals already prepared and that require only minimal preparation, such as defrosting or heating, were useful to improve their diet (Contini et al., 2020). Therefore, these two studies suggest consumers' interest in processed vegetarian and vegan products. Nonetheless, vegetarian and vegan products are not always healthy, especially if they are processed. For this reason, it should be noted that the results of Contini et al. (2020) may seem contradictory, in the sense that it was found that health concern is one of the reasons for the interest in convenience vegan products. Indeed, in a study conducted by Bullock et al. (2020), a halo effect was observed in a vegan ice cream regarding its salubriousness, meaning that people considered the product to be healthier than reality just because it was considered vegan.

Regarding vegetable alternatives to dairy products, in a study conducted in the United Kingdom, over 50% of participants reported that they consumed such alternatives at least once a week (Culliford & Bradbury, 2020). Pandey et al. (2021) found, more precisely, that approximately 21% of respondents in their study consumed plant-based yogurts two to three times per week and only 11% consumed them daily, with plant-based alternatives preferred by the majority being soy yogurts. As for cheese alternatives, a study on consumers from different countries and following different diets found that cheeses resulting from precise fermentation (a technological process carried out in a laboratory and which does not involve the use of

animals), that are not yet on the market, will have more market penetration than the current niche of vegan cheeses, with main interest from flexitarians (Thomas & Bryant, 2021).

Finally, it should be noted that in addition to the fact that health, environmental and animal issues motivate the reduction of meat consumption and the increase of plant-based products, social pressure also has a significant and influential impact on the adoption of plant-based diets (Cheah et al., 2020, Malek et al., 2019).

2.2.3. Veggies' purchasing patterns

Plant-based food substitutes are of interest to all consumers, regardless of their diet. That is, vegetarians, vegans, and even omnivores are interested in buying plant-based substitutes (Kopplin & Rausch, 2021). Almost 30% of consumers who are neither vegetarian nor vegan, stated that they regularly buy vegetarian or vegan private-label products (Martinelli & De Canio, 2021). Interestingly, it was found that in China and India, individuals who consume more meat are significantly more likely to buy plant-based meat substitutes than vegetarians and vegans (Bryant et al., 2019). Furthermore, a study at Cambridge University showed that the increase in the offer of vegetarian or vegan options (from 25% to 50%), increased sales of these meals (between 49% and 79%, considering three distinct cafeterias) and substantially reduced meat consumption, even among those who did not follow a plant-based diet (Garnett et al., 2019). Thus, the authors of this study suggest that the selection of these options does not solely depend on preference or randomness, but in part on the proportion of vegetarian or vegan options that are available. As expected, individuals who are more exposed to products are more stimulated to consume them (Contini et al., 2020). Hence, sales of vegetarian or vegan products in a supermarket may be proportionally higher in situations where the assortment is larger than in situations where the offer is smaller.

Likewise, displaying vegetarian and vegan products in a shop first, away from meat products, may also increase sales. A study in another university found that vegetarian options that were placed first in a canteen and were more than 1.5 meters away from meat options, increased monthly sales of vegetarian options by around 6.2 percentage points (Garnett et al., 2020). Similarly, vegetarian options that were placed at the top of a cafeteria menu, compared to situations where meat options appeared at the beginning of the menu, decreased sales of meat options by approximately 6 percentage points and since the number of customers remained the same, there was a positive effect in the sales of fish and vegetarian options (Andersson & Nelander, 2021).

Regarding meat substitute products, it was also found that the protein most likely to be purchased is plant-based, followed by cultured meat and protein from insects (Gomez-Luciano et al., 2019a), which is in line with findings for consumption intentions. However, aversion to new foods and especially lab-produced foods tends to make people less willing to try, buy or pay more for meat alternatives (de Koning et al., 2020). In contrast, in a study conducted by Szejda et al. (2021), in a sample in which about 2% of participants were vegan, 4% vegetarian, and 3% piscivorous, it was found that 59% of respondents were very likely to buy plant-based meat, even though approximately 32% were predisposed to pay more for this product. On the other hand, Slade (2018) suggests that the demand for products that simulate meat (plant-based or cultured meat-based) is price sensitive, meaning that if the price of the products increases the demand will decrease, although the author also found that individuals who have a strong preference for these products (such as veggies are), tend to be less sensitive. In line with these latter results, another study found that meat reducers and vegetarians give less importance to price than meat eaters and that it is especially meat reducers who have lower price sensitivity (Apostolidis & McLeay, 2019).

Regarding plant-based beverages, there has been an increase in the trend of their consumption, which can be justified by the fact that soy milk serves as a substitute for conventional milk (and vice versa), which means that when the price of milk increases, consumers tend to buy soy milk (Yang & Dharmasena, 2021). This result suggests that even individuals who do not follow a plant-based diet may consume these types of drinks. Furthermore, it was found in the same study by Yang & Dharmasena (2021) that, on the one hand, consumers tend to frequently purchase soy, almond, and rice milk together and, on the other hand, consumers are not sensitive to price changes (inelastic demand), meaning that increasing the price of these beverages does not decrease their purchase in the same proportion. Since consumers of plant-based beverages are not price-sensitive and veggies are the main target of this market, in some way, these results support Slade's (2018) findings, mentioned earlier, about these individuals being less price sensitive.

To conclude, it should be noted that just because a person follows a plant-based diet, it does not necessarily mean that they do not buy animal products, as well as the opposite. Approximately 72% of individuals who self-identified as vegetarians in one study reported that they regularly or occasionally purchased meat products for their partner, other family members, or guests (Apostolidis & McLeay, 2019). Therefore, purchasing patterns may differ from consumption patterns.

2.2.4. Veggies' profile

The adoption of plant-based diets tends to differ according to a set of sociodemographic characteristics, such as gender, age, and education (Sucapane et al., 2022, Koch et al., 2019, Gomez-Luciano et al., 2019b). Those who adopt these diets, or consume plant-based products, tend to be female, younger, and with higher levels of education, compared to unrestricted omnivores (Hielkema & Lund, 2021, Malek et al., 2019, Bryant & Sanctorum, 2021, Culliford & Bradbury, 2020). It was also found that those who follow more plant-based diets tend to have a higher income (Pandey et al., 2021, Bryant et al., 2019, Apostolidis & McLeay, 2019, Pfeiler & Egloff, 2018).

However, some studies contradict these claims. According to Grasso et al. (2019), whose study was conducted with data from five European Union countries (United Kingdom, Netherlands, Spain, Poland, and Finland), but among people at least 65 years of age, no evidence was found that gender, as well as age, were predictors of consumption of plant-based protein sources. The same results even suggest that there is a window of opportunity to increase the acceptance and consumption of more sustainable alternatives to meat by these elderly individuals. In parallel, Morris et al. (2020), suggests that young women tend to follow extreme eating patterns, i.e., both healthier (vegetarian) and less healthy (lower consumption of vegetables, salads, and fruits and higher consumption of snacks).

Other authors have found no significant differences in education, as well as housing area, between individuals who consumed meat and those who intended to or had already reduced their consumption (Niva & Vainio, 2021). However, although the difference was not significant, they found that, while meat eaters and meat reducers lived more in big cities, individuals who did not consume meat and had increased their consumption of plant-based proteins lived more in small towns. Contrarily, Kosh et al. (2019) found that no meat consumption is more frequent among those who do not live in small cities and rural areas. Also, in the study of Hielkema and Lund (2021), it was observed that large cities had more percentage of vegetarians and vegans than small cities.

In terms of household size, compared to meat eaters, individuals who reduced or that do not consume meat (veggies) are more likely to live alone (Grasso et al., 2021; Kosh et al., 2019). Moreover, comparing those who exclude meat (vegetarians and vegans) to those who just reduce their consumption (flexitarians), the first group has a higher share of people living alone (Malek & Umberger, 2021).

Concerning lifestyles, vegans give less importance to convenience (Malek & Umberger, 2021). However, in addition to the existence of a previous intention, the availability of products on the market and especially the lack of available time leads individuals who predominantly follow a plant-based diet to consume convenience plant-based products, i.e., products that are pre-cooked and require only minimal preparation (Contini et al., 2020). In any case, it was also found in the study by Contini et al. (2020), that even individuals who had more relaxed lifestyles felt the need to consume these products in occasional situations.

Regarding purchasing patterns, according to Martinelli & De Canio (2021), the intentions and respective frequency of purchase of vegetarian or vegan private-label products do not differ according to age, education, and income. Although vegetarians and vegans tend to be younger, in the study by Szejda et al. (2021), 60% of individuals between the ages of 18 and 27, 62% between the ages of 28 and 41, and 53% between the ages of 42 and 61 were highly likely to purchase plant-based meat alternatives. On the other hand, in the same study by Szejda et al. (2021), although income was the only sociodemographic variable that had an impact on predicting the intention to purchase plant-based meat substitutes, it was interestingly observed that the lower the income, the more likely the intention to purchase these products.

Finally, individuals who are reducing their meat consumption and are more likely to consume or purchase plant-based meat substitutes are more politically liberal and left-leaning (Hielkema & Lund, 2021, Lacroix & Gifford, 2019, Bryant et al., 2019).

3. Methodology

The methodology applied in this thesis was based on the Cross Industry Standard Process for Data Mining (CRISP-DM), as it is widely used in data mining projects, and is organized into six phases: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment (Martínez-Plumed et al., 2019). Since this methodology is iterative, its phases are described in this project in a more convenient way to represent the iterative process that was executed.

Regarding business understanding, as previously mentioned, the business goal of this project was to generate value for PD in the market of plant-based products and more precisely, to acquire better knowledge about veggie customers and the shopping patterns associated with them. This analysis, which was performed on transactions that took place over a six-month period (from September 2021 to February 2022), would therefore be useful for future marketing campaigns. For this objective to be achieved, the analytical objective was to use a clustering technique to identify segments and characterize veggie customer profiles, as well as their buying patterns.

Therefore, the analytical strategy that was followed in this project is described in this chapter. First, the source data (section 3.1) and the framework followed for this data, which included the creation of a new item's market structure (section 3.2), were described. Subsequently, the PD's business was explored, and an overall descriptive analysis was performed (section 3.3). Then, different types of clusters were identified and characterized, including the main cluster analysis in this project, which was based on the customer's basket, and their results were evaluated (section 3.4). After, criteria were defined to profile the customers according to the clusters found based on shopping baskets and whose behaviors were associated with different diets (section 3.5). To finalize, strategic recommendations were given to PD to achieve its business goal, but also to improve its engagement with the clusters found (section 3.6)

3.1. Data Source

The database that was accessed, from JM, was comprised of 5 tables which were named *Customers*, *Customers_Segm*, *Stores*, *Items*, and *Sales*. To cross information between all tables, each of the first four tables, which contextualized the data, could be linked to the *Sales* table

through a unique identifier field, an ID, which was common to them. Although the five tables contained several fields, a quality assessment and pre-selection of the variables that would be important for this research were performed, either based on descriptive analysis, to check for missing values, extreme values, and erroneous data, or based on their value to the business. The tables and the selected fields that were considered in this project are in Appendix D.

The data was collected from a *Teradata* environment using different queries and was subsequently integrated, analyzed, and processed in *IBM SPSS Modeler*, a tool that allows data preparation, exploration, and the use of different algorithms of machine learning. Due to the volume of the data, these extractions and integration were a challenge.

The *Customers* table (Table D1) contained declarative data about the demographic characteristics of individuals who acquired a loyalty card and did their registered on the PD website. But it is important to mention that their data was anonymized, so their identification was done through an ID. Considering only the customers who made purchases during the analyzed period (September 1, 2021, to February 28, 2022), around 3 million individuals constituted this project sample. Since the data in this table was declarative information given by the customers, it was not always filled out or submitted with quality. Therefore, there was access to demographic data of around 88% of them. For instance, regarding locations, despite the poor quality of this data, these variables were not removed for later comparison with the information accessed from the sales table, since the customers' home or work area could be deduced by their behavior in a certain store.

The *Customers_Segm* table (Table D2) included data about a segmentation performed by JM, which has allowed the company to recognize the engagement level that customers have with PD. This one identifies, for a given period, core customers (that make all or most of their purchases at the PD), peripheral customers (not very loyal), and shared customers (that also go to competitors). Around 21% of the analyzed customers were not assigned to any segment for operational reasons.

The *Stores* table (Table D3) contained data such as the location, format, and origin of distinct places, which corresponded not only to opened or closed stores but also to other establishments that JM Group owns or in which it has operated (e.g.: Recheio stores, gas stations, clothing stores). However, filtering the table to the goals of the work, 469 PD stores remained to be analyzed.

The items table included fields that allowed the identification of each item’s market structure, which is a hierarchy of different levels to organize the items by divisions, and allowed the identification of the item's brand. Being an incremental table, items were frequently added

or recoded, and the older ones might not be eliminated. By delimiting the items to the analysis scope, it allowed the volume of data to be reduced to useful information, excluding obsolete items or other item codes generated due to operational reasons. Therefore, around 82.4 thousand distinct code items were considered, which were either purchased or returned by the customers' sample of this project. In terms of quality, it was verified that, even though each item code was unique, approximately 4,7% of these items had similar descriptions.

Finally, the *Sales* table which was the most important because it contained attributes that characterized each transaction, presented the data with the highest level of detail, i.e., each record corresponded to each item that was scanned for a particular transaction, in a specific cash register, at a certain time, in a certain store. However, it was possible to verify some registers with negative sales values, which did not correspond to errors but rather to discounts or returns. For this reason, even filtering sales to the 6 months and for the identified customers, this table had more than 1.136 thousand million records, which corresponded to 53.1 million transactions. For this reason, it was necessary to perform different extractions, aggregating data from the start in different ways, according to the need of what was intended to be analyzed, namely, spending amount, frequency, preferred store, and preferred time to shop. The fields that were selected based on the business and used for these aggregations are illustrated in Table D5. It should be noted that the variable that identified the quantity of each item was not considered since it could be associated with either units or kilograms, depending on the type of item, so it would be complex to find a unified method comparable across the different units.

3.2. Data Framework

To characterize the PD customer, new variables were created such as the customer's current age, but also age ranges, and length of loyalty. Identified outliers in the age variable that would be an error were considered missing values. Also, variables related to their location went through a cleaning process, such as removing meaningless characters, and accents, among others. Then, as there were only identified customers from Portugal and Spain, this data was crossed with a database from CTT (2022), which gathered all existing postal codes in Portugal and respective localities and permitted an identification more accurately of each customer's district and municipality. Similarly, for customers coming from Spain, the information was crossed with data from a website (https://esp.postcodebase.com/) with postal codes and respective regions. This process resulted in the identification of the customer’s district and municipality if they were from Portugal, and their region in the case they were from Spain.

Regarding stores data, a dataset from INE (2014) that contained the identification of the urban-rural typology of each parish, allowed the creation of a new variable that identified the type of area (predominantly urban area, medium urban area, or predominantly rural area) of each store's parish. Additionally, the variable related to the format of stores was recodified to allow a more concise characterization, only based on their format: hiper stores (bigger stores), mega stores (medium stores), supermarkets (smaller stores), and convenience stores.

3.2.1. New items market structure

3.2.1.1. Nomenclature creation

A key step, vital for this study, was the creation of a new nomenclature to classify vegetarian and vegan items. Although the company had some categories that could be directly associated with vegetarianism, such as vegetarian dishes and plant-based milk, they were limited. Therefore, only with a new categorization, it would be possible to identify all the transactions that contained items belonging to a certain class associated with vegetarianism. For this reason, was created a new variable (*Class_Veg*), that could assume four exclusive categories, as illustrated in Table 3.1.

Table 3.1: Vegetarian Nomenclature- Classe_Veg Variable

Classe_Veg	Description
0	Item not classified with new nomenclature
1	Vegetarian item/ Suitable for vegetarians
2	Vegan item/Suitable for vegans
3	Non-vegan and non-vegetarian items/ Neither suitable for vegetarians nor vegans

Since the focus of the analysis was food items fit for human consumption, to classify them as being suitable for vegetarians, suitable for vegans, or neither, the variable *Class_Veg* assumed the value 0 in all items where there was no classification in terms of this nomenclature related to vegetarianism. Thus, the value 0 was attributed to those items that corresponded to non-food products or those that represented customizable products and did not contain enough information, making it impossible to classify them, such as utensils, clothing, pet food, supplements, customizable menus, and beverages. However, some exceptions were considered regarding beverages, like dairy drinks, plant-based drinks, and other hot drinks, such as coffees and teas, which were included in the classification process. Therefore, any other types of beverages such as water, juices, and alcohol were excluded. It should also be noted that, although food supplements are considered food for human consumption, they function as a

complement and not as a substitute in the diet (Martins et al., 2017), besides being difficult to classify as being suitable for vegetarians or vegans (Mingo, 2021), especially based on the information that was accessed. Therefore, the supplements in the form of pills were also not classified according to the nomenclature associated with vegetarianism, assuming the value 0 in the variable *Class_Veg*. To simplify, from now on, these items (*Class_Veg*=0) will be referred to as *non-class*. Also, the items classified as neither suitable for vegans nor vegetarians (*Class_Veg* =3) will be referred to as *non-veg*.

Following a plant-based diet is not about a religion, so what may be acceptable for some individuals may not be for others. Furthermore, to date, there is no legal definition in Europe regarding the use of the terms *suitable for vegetarians* and *suitable for vegans* for product labeling (McElfresh, 2021). Only independent organizations issue certified labels for these types of products. In line with this, and due to the difficulty in obtaining complete information on some items (e.g., the identification of their ingredients and their origin) and for time management reasons, a more flexible and generalized definition for these types of items was considered throughout this research.

Therefore, the classification of vegetarian items was based on the type of diet followed by ovo-lacto-vegetarians (who exclude animal products such as meat and fish but consume their derivatives, such as eggs, dairy products, and honey), but not too strict, while the classification of vegan items was based on the diet followed by vegans. In any case, it was also taken into consideration the definition of products suitable for vegans and suitable for vegetarians, proposed to the European Commission by the European Vegetarian Union, FoodDrinkEurope, and EuroCommerce (McElfresh, 2021).

This classification took into consideration the whole PD's market structure up to the item description, which corresponded to the highest level of detail. Another field considered and of extreme importance was the item brand, because some brands are automatically recognized for selling vegetarian or vegan products (e.g., *Alpro, Shoyce, Green Cuisine*, among others). For situations where it was not clear what classification should be attributed to the item based on its market structure, it was used the internet for this purpose. First, it was investigated whether the item had already been considered vegetarian or vegan, namely, on the official PD website, the official website of the item's brand, in other retailers' websites, forums, and other pages, such as the Abillion website (https://www.abillion.com/), which identifies numerous products as vegan or vegetarian based on users' evaluations, and the Facebook page Achados Veganos (n.d), which contains information about vegan products that exist in Portuguese retail. In addition, or at least when it was not possible to obtain this information, the ingredients of the

items were also identified with the help of the sources mentioned above but also others, such as the Open Food Facts website (https://world-pt.openfoodfacts.org/). It was even necessary to verify the origin of certain ingredients, such as additives (Johnson, 2022).

Therefore, in the analysis of ingredients, several criteria were followed. Items that contained animal components such as meat, fish, seafood, sausages, animal fats, additives, or gelatins of animal origin were classified neither vegan nor vegetarian (*Class_Veg*=3), as was the case of items belonging to divisions such as fishmonger, butchery, but also certain items associated to desserts, such as puddings and mousses.

Items that contained any dairy products, eggs, or honey, but did not include other animal products, such as those mentioned above, were considered vegetarian (*Class_Veg*=1). It is noteworthy that certain items that belonged to the bakery/pastry, restaurant, and take-away divisions, mainly related to dessert items, and that corresponded to already processed products, could include animal gelatin in their composition. However, due to a lack of information regarding its inclusion, these items were classified as vegetarian. On the other hand, although the cheese that contains rennet is not suitable for the strictest vegetarians, it may be accepted by others, since it depends on everyone's choice (Wartenberg, 2018). Therefore, since it was considered a less strict definition for a vegetarian consumer in this work, all cheeses were considered as being suitable for vegetarians, except when they were clearly identified as vegan, or, contrarily, when they contained other animal products such as salmon or ham. To conclude, items such as milk, eggs, cheese, butter, cakes, ice-creams, and certain desserts, were classified as suitable for vegetarians, except when it was possible to identify them as being vegan or non-veg.

Finally, items that did not contain animal products or derivatives, additives, colorings, or other substances known to be of animal origin were considered vegan items (*Class_Veg*=2). This group included items such as vegetables, fruits, legumes, and processed products already identified as vegan. Note that a product considered vegan is also suitable for vegetarians, but not vice-versa (Figure 3.1). Therefore, if the product could be suitable for vegetarians but, in addition, met all the conditions to be classified as vegan, it was classified as vegan *(Class_Veg*=2). Otherwise, it was only considered vegetarians (*Class_Veg*=1).

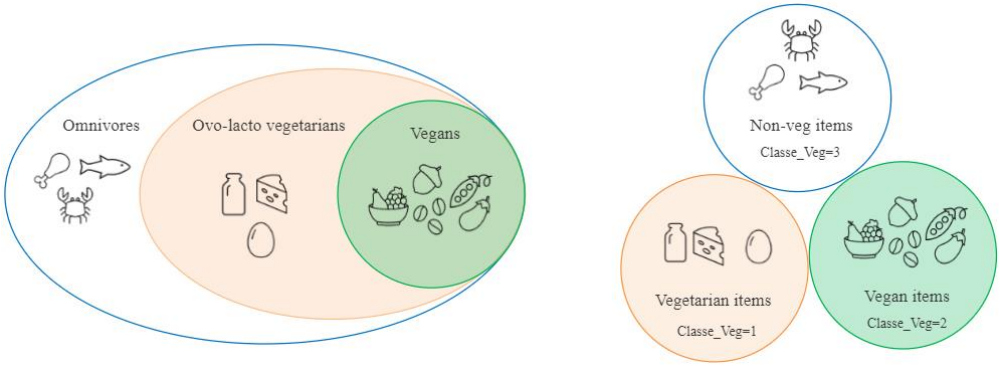

Figure 3.1: The spectrum of diets and the associated items classification
Source: Elaborated by the author

However, there were some limitations. In situations in which an item could be considered vegan but contained an ingredient that had raised doubts due to a lack of knowledge of its origin, it was decided to classify it as only vegetarian. Also, in cases where it was not possible to have access to the list of ingredients or any additional information, making the classification of the items uncertain, they were classified based on the classification made to items with similar characteristics and according to the marketing structure to which they belonged. Additionally, in situations where the item's list of ingredients mentioned that it could contain traces of animals, such as milk or eggs, it was considered that it did not contain them, since their existence was not intentional. If it was considered that a product can only really be labeled as vegetarian or vegan if it is not at all contaminated with certain animal products, it would imply that there would be no products labeled with these terms, because even following the best practices any food can be contaminated by animal products (European Vegetarian Union, 2019).

To conclude, it should be highlighted that the creation of the new product nomenclature related to vegetarianism resulted from an iterative process, which allowed a more correct categorization. Moreover, it should be remarked that it was not removed items with similar market structure descriptions since transactions were identified by the code of the item, which was all unique, and their elimination would lead to a loss of information about those items. Also, they were neither recodified because even with similar descriptions they could represent different items. Instead, one strategy adopted in this project to analyze the variety of the basket of each customer was to count the number of distinct categories (one of the levels of the market structure and often used by JM) instead of counting distinct item codes, since it also provided a more general and realistic view of the purchasing behavior. Nevertheless, items were aggregated in new groups, as will be explained in the following subsection, and it was given

more importance to the amount spent, like Morris et al. (2020) did in their study, and to the number of transactions.

3.2.2.2. Item Groups creation

Product taxonomy is important, and it might affect the results (Griva et al., 2018). Even though PD has its market structure, the way its items are organized does not allow a clear identification of customers who may follow plant-based diets.

For this reason, and always considering the new nomenclature that was mentioned in the previous subsection, the items related to food were aggregated based on previous studies related to vegetarianism (Gallagher, 2022; Abreu, 2021) and on the study conducted by Morris et al., (2020). The remaining items were aggregated according to the original market structure of PD or left isolated when it was no longer appropriate within the structure. The item groups and their content were aggregated in vegetarian, vegan, non-veg, and/or non-class groups, at a lower level, and aggregated independently of the new vegetarian nomenclature, at a higher level. For instance, for the dairy group, it was first identified daily milk, ultra-pasteurized milk, vegan oat milk, vegan soy milk, etc. Then, these were aggregated by *vegetarian milk vs. vegan milk* and other items by *vegetarian yogurts vs. vegan yogurts vs. non-veg yogurts*, and *vegan margarine vs. vegetarian margarine*, among other groups. On a higher level it was distinguished *vegetarian dairy vs. vegan dairy vs. non-veg dairy*, and at the limit as the group of the *dairy*.

Therefore, after the creation of this adapted market structure, the first objective of the project was achieved (O.1 – Create a market structure for vegetarian and vegan items). Based on the item groups created (Appendix E), sales table data was aggregated at the customer level to identify values associated with these new groups for each customer. That is, for each customer, the monthly average expenditure over the six months and the total number of transactions during the whole six-months period on each item group were collected.

3.3. Pingo Doce's Business

The data preparation phase is also important to explore and gain knowledge about data before going to the Modelling phase. Hence, after some transformation of the data, it was possible to characterize PD customers. All subsequent analyses considered only the data from the identified customers who were part of this project sample, but since this is sensitive information, some data are only described in a qualitative format.

Starting with identifying who are PD customers, it was observed that most customers who made purchases in the chosen period were female, and the age range to which the highest number of customers belonged was between 36 and 55 years old, followed by those between 56 and 75 years old. Most of the customers were from Portugal, with a higher concentration of individuals from Lisbon and Porto. In fact, according to Nielsen (2022), 24% of the families live in Lisbon and 17% in Porto. Moreover, of the Spanish customers who were registered and did purchases, most of them were from Galicia followed by customers from Andalusia. Nevertheless, it should be noted that this demographic data might not be completely accurate since it was provided by customers themselves and without prior control.

Relating to where customers preferred to go, it was found that for 76.1% of the analyzed customers, approximately, their favorite store was a supermarket, followed by 14.9% of customers whose favorite store was a mega store (Figure 3.2). This might be explained by the distribution of types per store because the majority of PD's stores are supermarkets, followed by mega stores. Each customer's favorite store was calculated based on the monthly average number of transactions, monthly average value spent, and the monthly number of different items bought.

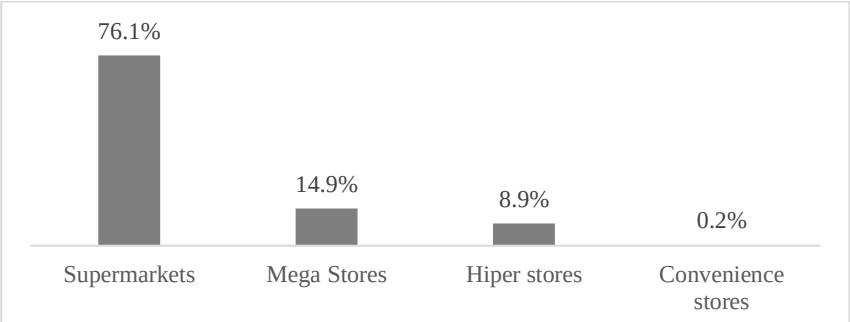

Figure 3.2: Distribution of customers by type of preferred store

For another hand, the plurality of these customers had their preferred store in Lisbon (27%), followed by Porto (17.9%) and Setúbal (8.7%), which is in accordance with the distribution of stores in Portugal. It is also worth mentioning that at least 79.8% of customers preferred to shop in the same district they said they came from. Whereas at least 8.6% of the customers, approx., had their preferred store in a district different from their home, no information was available about the origin of the remaining 11.6% of customers because of their demographic data that was in default. Moreover, comparing the workdays with the weekends, at the weekend more customers preferred to go to hiper stores (+6.8%) or megastores (+3%) (Figure 3.3). In addition,

compared to weekdays, there was a reduction in the number of customers who preferred to shop in big cities, such as Lisbon (-1.28%) and Porto (-0.32%), and an increase in the number of customers who prefer to do their shopping in suburbs or other districts at weekends, such as Santarém (+1.65%), Leiria (+1.47%), Évora (+1.37%), or Setúbal (+1.21%). This may be explained by the fact that people may work in big cities and may live on the periphery since on weekdays purchase in more predominantly urban areas and on weekends purchase in more medium urban areas (Figure 3.4).

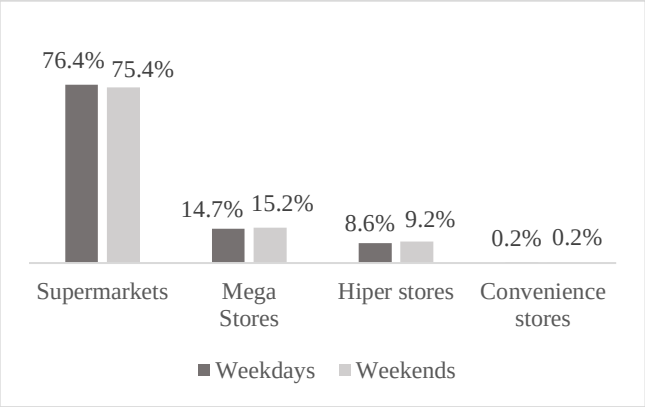

Figure 3.3: Distribution of customers by type of preferred store, by day of the week

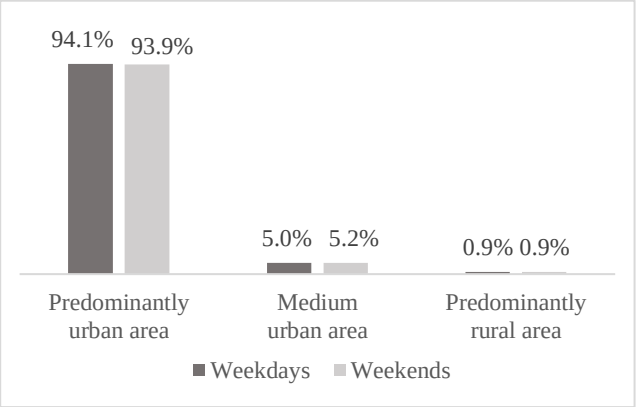

Figure 3.4: Distribution of customers by urban-rural typology of preferred store's parish, by day of the week

Analyzing afterward when customers tend to go to the PD, it was found that the number of days they went to a store was approximately equal to the number of transactions they made, on average, so it could be considered that each transaction represented one visit to the shop. Analyzing the preferred hour to shop (Figure 3.5), which was obtained according to the same criteria as for the preferred store, it was found that, on weekdays, there is a greater preference for shopping between 17:00 and 19:00 (preference of 31.6% of these customers) and, at the weekend, between 10:00 and 12:00 (preference of 32.7% of the customers).

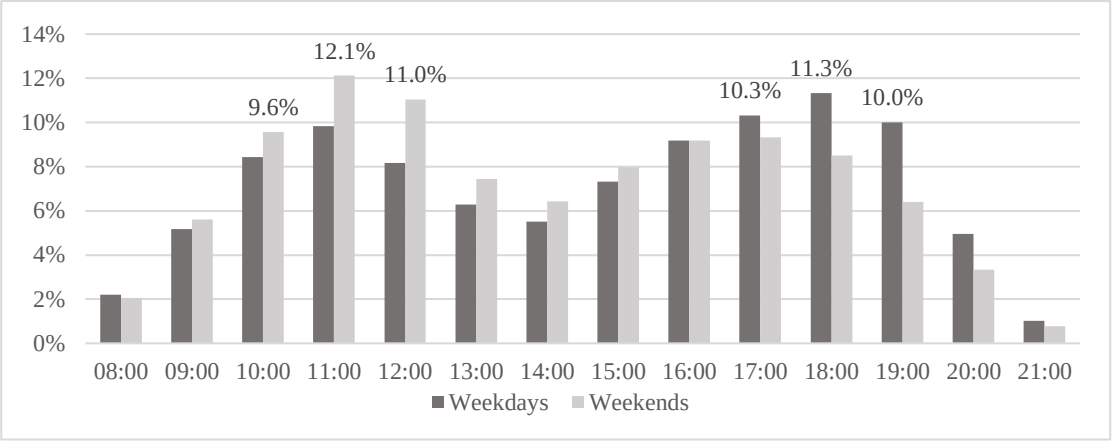

Figure 3.5: Distribution of customers by the preferred hour, by day of the week

In terms of sales, it was concluded that for the 6 months and based only on the customers considered in this investigation, PD had an average monthly sales volume of 271.5 million €, approximately, whereby 23% was from vegan items, and 19% from vegetarian items. Furthermore, on average, 60.4% of the amount billed per month by the PD, from the sale of vegan and vegetarian items, comes from just the top 20% of customers who spend the most on these items, as demonstrated by Figure 3.6.

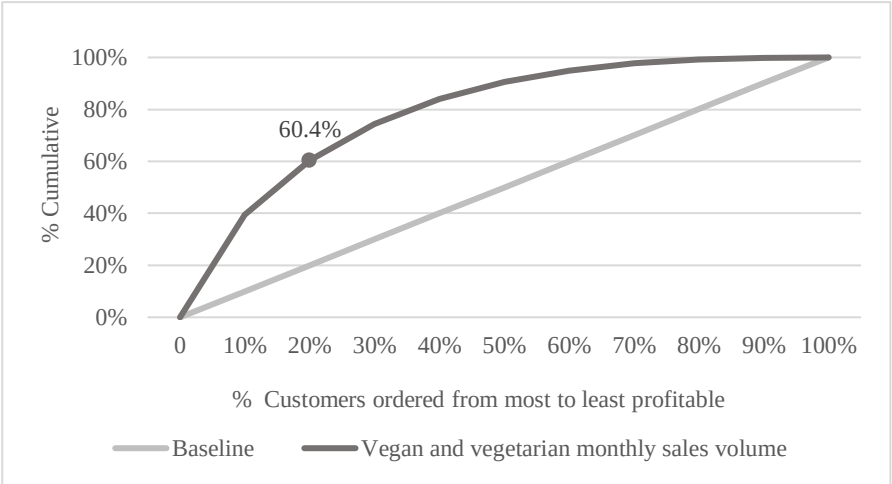

Figure 3.6: Cumulative vegan and vegetarian monthly sales volume

In terms of protein alternatives to meat and fish, from the customers that buy at least once one of these alternatives (item group 1), it can be seen by Figure 3.7 that vegan processed meat (item group 1.2), like hamburgers, falafel, chili, nuggets, among others, is the option that more customers buy (54%), followed by similar but vegetarians options (28%). Also, regarding the

alternatives more processed, in general, vegan alternatives were bought by more customers than vegetarian options, not only for processed meat alternatives (item group 1.2) but also for alternatives to highly processed meat (item group 1.3), such as sausages, bacon, *alheira*, and *farinheira*. Besides that, more customers repeated the purchase of these vegan alternatives than of vegetarian alternatives (Figure 3.8), which might represent that those customers have been liking the vegan choices. For another hand, in general, less processed choices (item group 1.1), such as soy, tofu, and seitan, are bought by fewer customers (16%, 15%, and 8%, respectively). Despite the fact that soy was purchased by more customers compared with tofu (which is made from soy) and seitan, only 19% of the customers that purchased soy bought this more than once. In contrast, seitan, which was the choice purchased by fewer customers, had more who repeated the purchase (33%). This might be explained by the fact that, on one hand, seitan is a less known alternative protein (Byrne, 2019), and, on another hand, it is made from gluten, so the increase in gluten intolerance (Mehmet, 2020) plus gluten-free diets, may lead to people avoiding this product. However, while soy and tofu are a more versatile food, some people do not appreciate their original flavor, so these alternatives require more preparation and cooking skills to make them taste good, whereas seitan is more similar to meat in terms of texture and appearance and can be easily prepared like it. So, new buyers and frequent buyers may prefer seitan.

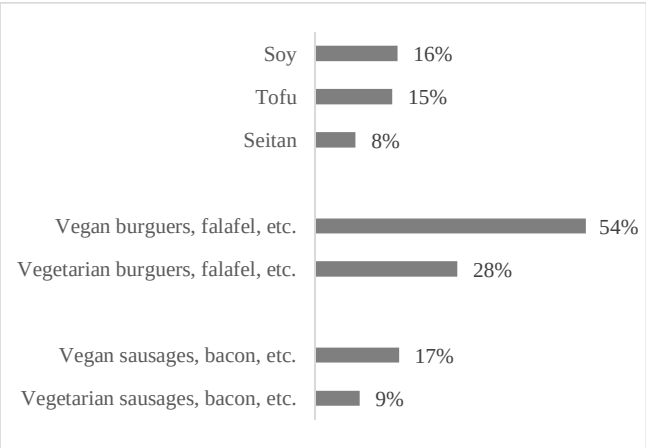

Figure 3.7: Distribution of customers who purchased alternatives proteins to meat & fish (item group 1) considering the whole 6-month period, by alternative

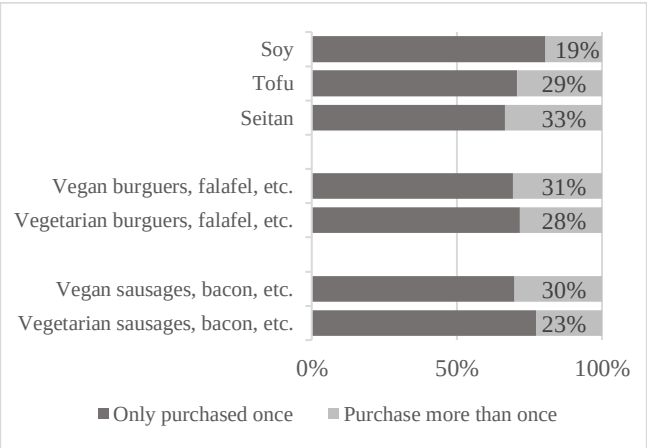

Figure 3.8: Distribution of customers by behavior and alternative protein, considering the whole 6-month period

Regarding vegan dairy alternatives, the most differentiating ones are represented in Figure 3.9 and Figure 3.10. Comparing just the alternatives from the graph, soy milk is the one more acceptable (bought by the 13% of customers that purchased vegan alternatives to dairy (item group 1), and from these, almost half purchased more than once), followed by oat milk (it was

bought at least once by 11% of the customers that purchase vegan alternatives to dairy, where 46% bought more than once). Considering vegan yogurts, from the 7% of customers who bought these items, 39% of customers repeated the purchase. Dairy creams were bought by a similar percentage of these customers (7%) but had more customers that repeated the purchase (42%). Rather, only a smaller percentage of customers purchased vegan ice creams (3%) and vegan cheese (1%). It was expected that ice creams were less purchased since these are a product more eaten in the summer season and the analyzed period was from September to the end of February.

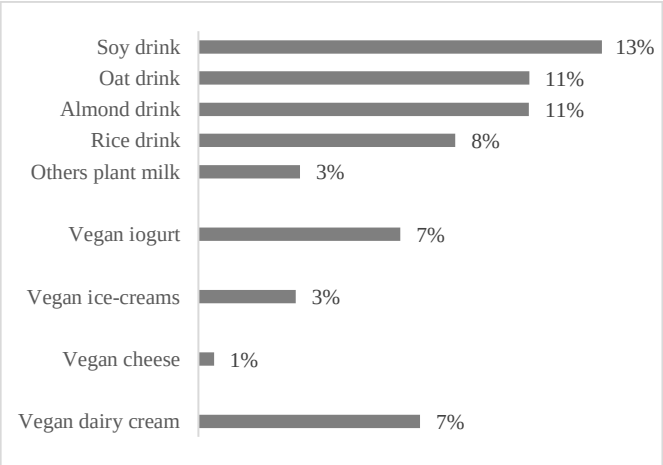

Figure 3.9: Distribution of customers who purchased vegan alternatives to dairy (item group 10) considering the whole 6-month period, by alternative

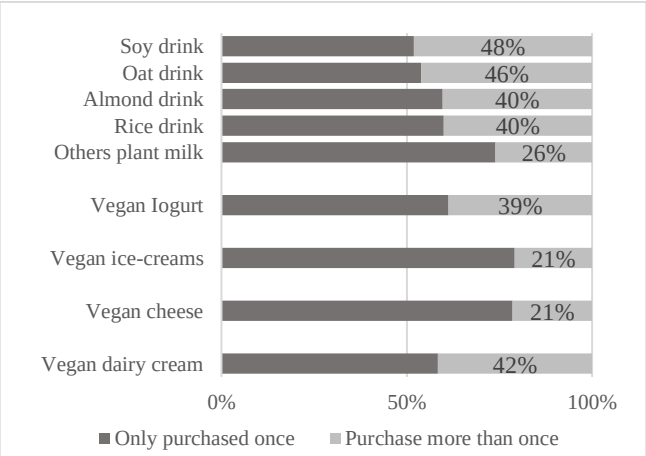

Figure 3.10: Distribution of customers by behavior and vegan dairy alternative, considering the whole 6-month period

Besides, it was also identified that vegan butter was the alternative that had the smallest customer penetration (almost null) in terms of vegan dairy purchases, and that had the smallest percentage of repeated purchases (by only 5% of the customers who had purchased at least once). Conversely, vegan margarine had the highest customer penetration (72%) and the highest percentage of customers who repeated the purchase (52%), which was expected because this ingredient is widely used for cooking and baking and can even be purchased by people who do not seek to follow plant-based diets.

3.4. Clustering Architecture

In this project, it was utilized a clustering technique, which is an unsupervised learning technique, to identify homogeneous groups of customers relative to certain common characteristics, and whose patterns would not be observed otherwise (Morris et al., 2020). The

method adopted was *two-step clustering* as used by other authors (Grasso et al., 2021) since it works efficiently with large data sets, can handle mixed field types (categorical and continuous variables), and can determine automatically the number of optimal clusters (IBM, 2021, Trpkova & Tevdoski, 2009).

The determination of the automatic number of clusters was based on the Bayes information criterion (BIC) (Grasso et al., 2021) and the distance measure was the log-likelihood since categorical variables were used in this project. To use this measure, it is assumed that data follow a normal distribution, but the two-step cluster algorithm gives good results even if this assumption is not met (Trpkova & Tevdoski, 2009). Furthermore, continuous variables were standardized by default, using the z-score method, because this ensures that all variables have equal weights and give better cluster results (Morris et al., 2020; Trpkova & Tevdoski, 2009).

The cluster analyses carried out in this project included four different types of clustering. The basket clustering was the main of this project to meet the second research objective (O.2 – Segment customers) and identify veggies, but three additional clustering types (demographic, lifestyle, and frequency and monetary dimensions) were also performed. These three dimensions were used not only to describe the PD's customers more specifically but also to help in the description of the clusters that would result from the basket clustering.

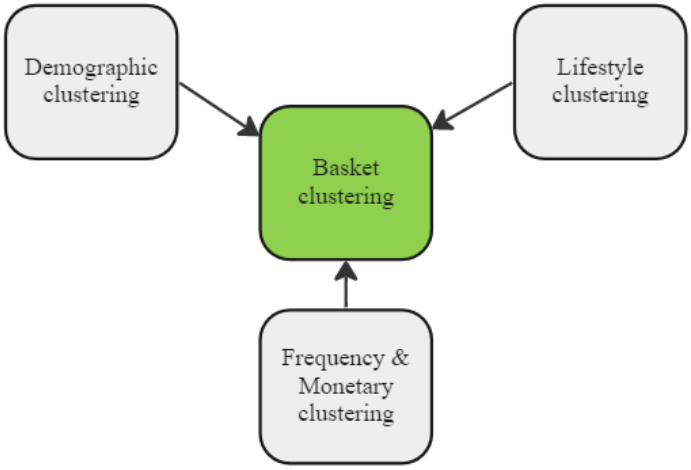

Figure 3.11: Cluster Architecture
Source: Elaborated by the author

3.4.1. Input Variables

As mentioned above, beyond the goal of segmenting customers in terms of the type of products purchased (shopping basket), clustering was initially performed to profile customers based on

other three different dimensions: demographic, lifestyle, and frequency and monetary. Since clustering is an iterative process, it was considered different variables as input and different optimal solutions. The final variables considered as input to characterize customers on these three dimensions are represented in Table F1, from Appendix F.

After, to identify veggie customers, it was performed a cluster analysis based on their baskets. Therefore, it was first created a data set in which each row represented a customer, the columns item groups, and each cell corresponded to the weight of the average monthly amount spent on a given item group in that customer's average monthly expenditure for the six months, i.e., it was the calculated the share of each customer's average monthly expenditure over the six months for the purchase of each specific item group (Equation 1).

$$\text{Share}_{ij} = \frac{\text{Average monthly expenditure on item group } i}{\text{Average monthly expenditure}}\text{, i= item group; j=customer} \quad (1)$$

Next, even though continuous variables were standardized in clustering, these data were first indexed by dividing each item group's share of a customer by the mean of the corresponding item group's share (Equation 2), to allow better performance on clustering.

$$\text{Share\_Index}_{ij} = \frac{\text{Share } ij}{\overline{Share\ i}}\text{, i= item group; j=customer} \quad (2)$$

Since, in general, the shares of specific items more related to the practice of following a plant-based diet, are lower than for other items, this normalization worked as a proxy of the level of importance and preference that each customer was giving to a group of items, comparing to the average. After several tests and based on the works of Gallagher (2022) and Abreu (2021), the twelve item groups considered as input to highlight veggies are represented in Table F2 from Appendix F.

3.4.2. Evaluation

The resulting clustering models were evaluated based on an analytical metric (silhouette coefficient of cohesion and separation) and their business applicability (interpretability and cluster size). The silhouette measure is ranged from -1 to 1, where values between -1 and 0.2 are classified as *Poor*, 0.2 to 0.5 as *Fair*, and 0.5 to 1 as *Good* (Supandi et al., 2022),

For each of the four clustering types that were intended to be realized, it were performed different models. First, it was determined the automatic optimal solution, i.e., the optimal

number of clusters given automatically by the model for each clustering. Then, these models were rebuilt multiple times by pre-selecting in each one a different optimal solution. Figure 3.12 illustrates the silhouette value from 2 to 7 clusters as a solution, to each clustering type.

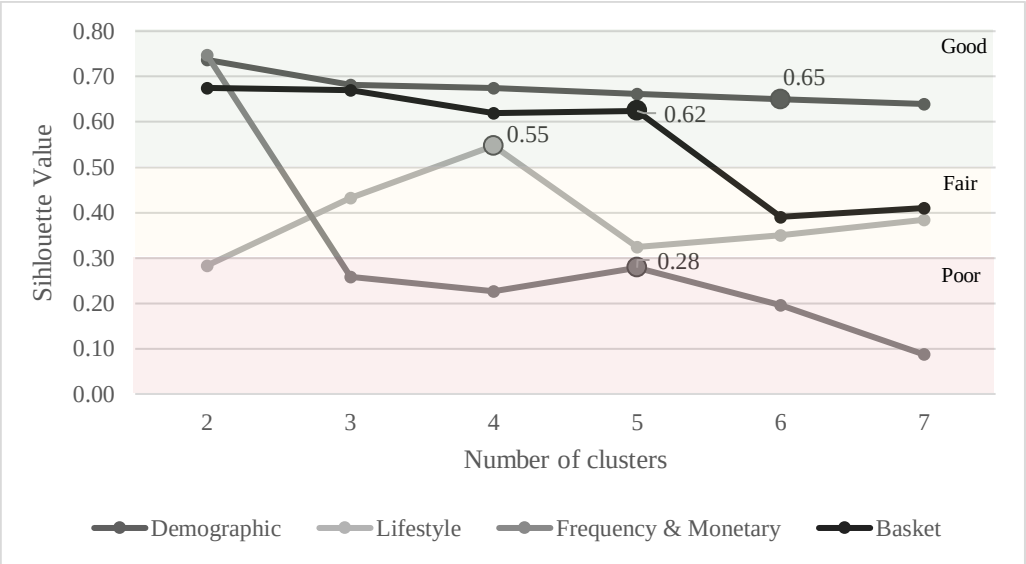

Figure 3.12: Value of silhouette by number of clusters for each type of clustering

The best analytical solution to a clustering problem is not always the best solution for the business. Despite some models with a solution of 2 clusters had the highest value for silhouette (in demographic, frequency and monetary, and basket dimensions), these models were not chosen to permit to have more unequivocal clusters. For instance, for demographic clustering, the silhouette decreased as the number of the optimal solution increased. However, in terms of interpretability, it was chosen the solution with 6 clusters, which was still a good solution (silhouette= 0.7). To cluster customers based on lifestyle dimension, it was chosen the solution with the highest silhouette (0.5), which had identified 4 clusters. Regarding frequency and monetary value, it was chosen the model with the second highest silhouette (0.3), which was yet a solution considered fair and acceptable (Supandi et al., 2022), resulting in 5 clusters.

Finally, in the clustering based on basket data, the most important clustering to identify customers with behavior that might be characteristic of veggies, it was chosen the model with 5 clusters as a solution (which was also the result of the automatic optimal solution) for being more appropriate in terms of cluster size, interpretability, and silhouette value

3.4.3. Results

For data confidentiality reasons, all subsequent results for continuous variables are represented as indices relative to the client's average PD, even for the results from dimensions clustering that were created to characterize the clients and whose continuous variables were in absolute values (e.g., age, average amount spend per transaction and monthly average number of transactions over the six months).

Starting with the results of demographic dimension clustering (Table 3.2), the six clusters identified were the following: *Youngest men* (cluster 3) and *Youngest women* (cluster 4), represented by 7% and 12% of the customers, respectively, and both with an average age which corresponded to 0.6 times the average of the PD customer; *Middle age men* (cluster 2) and *Middle age women* (cluster 5) represented by 14% and 24% of the customers, respectively, who were, on average, 0.94 times below the age average; *Elderly men* (cluster 1) containing 13% of men with an average age 1.35 times the average and *Elderly women* (cluster 6) with 18% of women who had, approximately, an age 1.31 times above the average.

Table 3.2: Results from demographic dimension clustering

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**	**Cluster 6**
Age Index (Mean)	1.35	0.94	0.60	0.60	0.94	1.31
Gender (Mode)	M	M	M	F	F	F
Cluster size %	12.7%	13.7%	7.3%	12.2%	23.6%	18.2%
Cluster name	*Elderly men*	*Middle age men*	*Youngest men*	*Youngest women*	*Middle age women*	*Elderly women*

Regarding lifestyle (Table 3.3), the first cluster was designated as *Not work*, (representing 21% of the identified customers), because showed not having a difference in behavior from weekdays to weekends. In general, they preferred to buy around 11:00 and in the same parish regardless of the day of the week. This behavior might be more associated with students, retirees, or other people that do not work. The second cluster, *Work close-Mega* (24%), whose favorite store was a mega store (medium stores), are the ones that probably work close to their residence since their favorite shop on weekdays is in the same parish as their preferred store on weekends, but while they prefer to shop at the end of the day (18:00) on weekdays, on weekends they prefer to shop at a different hour. The third cluster, *Work close-Super* (41%), the most representative, was similar to customers that belonged to *Work Close-Mega* but with the difference in the fact that they prefer to shop in supermarkets (smaller stores). The fourth cluster was designated as *Work far* (14%) since on weekdays they prefer to shop at 18:00, but on

weekends they prefer to go at a different hour and shop in a parish that is different from the one that they usually prefer to go on weekdays, what may indicate that they work not so close from their residence.

Table 3.3: Results from lifestyle dimension clustering

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
ID_hour_wkday (Mode)	11:00	18:00	18:00	18:00
EqualHour_wkday_wkend (Mode)	1	0	0	0
EqualParish _wkday_wkend (Mode)	1	1	1	0
Format_L1 (Mode)	Supermarket	Mega stores	Supermarket	Supermarket
Cluster size %	20.8%	23.8%	40.9%	14.4%
Cluster name	*Not work*	*Work close-Mega*	*Work close-Super*	*Work far*

In terms of frequency and monetary value, it was found four clusters in the model selected. The first one, represented by 54% of the customers, was nominated as *Products hunting*, since contained the ones that, on average, only spent per transaction 0.64 times the average, and went shopping, on average per month, 0.46 times less than the average, being consequently the cluster that spent less per month (around 0.36 times the average monthly amount, as can be seen in appendix G). For this reason, and because are the ones that bought the smallest number of distinct categories (a hierarchical level from the original PD's market structure), they might just go to PD to buy specific types of products. Furthermore, despite being the most representative cluster, the sales from these customers only represent around 20% of the monthly sales of PD. The second cluster, *Medium interested*, which is the second most profitable cluster, is represented by 23% of customers that go with a similar frequency as the last one but spent more per transaction (1.69 times above the average, on average). The third cluster, *High interested*, is the most profitable even though they only contain 18% of customers. The customers that belong to it go shopping more often per month, (realized 2.17 times more transactions than the average), even though they spent less per month than the average (0.76, on average). Customers from the fourth cluster, *Busy* (representing 1%), are the ones that go less frequently (0.21 times below the average) but spend the biggest amount (5.41 times above the average), so might do all the shops at once. The last one, *Daily* (with 4% of customers), is composed of customers that spent the most per month (3.31 times above the average, approx.), since they go the majority of the days (5.38 times above the average), even though they spent less per transaction (0.55 times below the average), compared with other clusters. In Table 3.4 it is possible to

verify the data results of these clusters and in Figure 3.13 how the indices of mensal frequency and amount spent per transaction are positioned in space.

Table 3.4: Results from frequency and monetary dimension clustering

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Avg_monthly_trx Index (Mean)	0.46	0.62	2.17	0.21	5.38
Avg_amount_trx Index (Mean)	0.64	1.94	0.76	5.41	0.55
Cluster size %	54.4%	22.5%	18.3%	0.9%	3.9%
Cluster name	*Products hunting*	*Medium interessed*	*High interested*	*Busy*	*Daily*

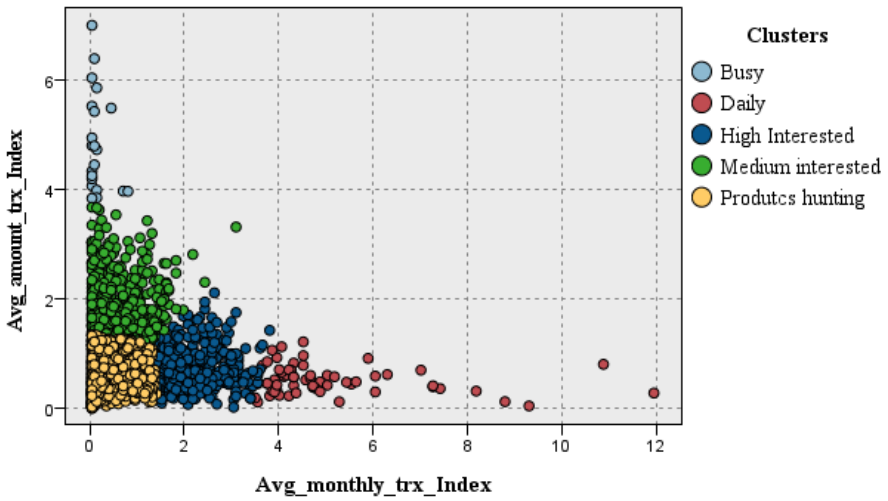

Figure 3.13: Clustering: Index of frequency (avg_monthly_trx_Index) vs. Index of amount spend per transaction (avg_amount_trx_Index)
Source: Obtained from *IBM SPSS Modeler*

Finally, regarding the most important clustering model based on basket data (Table 3.5), cluster 1 was the most representative with 93.2% of customers belonging to it. It was designated as *Traditional Omnivores* since contains the customers whose share of average monthly expenditure for purchasing the item groups considered as input is lower than the average. This corresponds to a total of only 2.3% of these customers' average monthly spending amount that is spent in these item groups, as can be seen in Appendix H. In fact, on average, their share for protein alternatives to meat and fish is the lowest, compared to other clusters (on average, only 0.01% of their average monthly expenditure is spent on these alternative proteins).

The second cluster was designated as *Receptive Omnivore*s and was composed of 2.8% of customers. Even though this is in general the third cluster with more preference for vegan and vegetarian alternatives to meat and fish, their interest is low compared with others. Moreover,

it is only on processed meat alternatives like vegan or vegetarian burgers, beef, nuggets, falafel, etc., that they spend more per month than the average (1.4 times above the average for vegan processed meat alternatives and only 1.01 for vegetarian ones). Nevertheless, without considering most of the protein alternatives, their average monthly shares for the remaining types of food used as input (dairy, biscuits, sweets, creams, etc.) are above the average. Special emphasis goes to the fact that it is the cluster that spends more on vegan cookies, diet bars/balls, and cereal bars (monthly share on these items is 8.4 above the average), more on vegan sauces and creams (monthly share on these items is 5.8 above the average) and it is the second cluster to spend more in vegan ready foods such as soups, vegan noodles, bowls, among others (monthly share on these items is 21.9 above the average), given its average monthly expenditure.

For the third cluster, *Convenience and Vegan Sweets*, represented by 1.8% of the sample, in terms of protein alternatives only the amount spend on vegetarian alternatives to highly processed meats (such as vegetarian sausages, ham, *alheira*, etc.) has a more significant weight in the average monthly expenditure of these customers. The percentage that is spent per month on these vegetarian items is approximately 12.6 times the average, and for this reason, is the second cluster that gives more importance to this kind of product. Yet, they are not so interested in other protein alternatives. Conversely, this is the cluster that most prefers to buy vegetarian ready food, vegan snacks, and vegan sweets and desserts, since the ratios of their average monthly expenditure shares on these food groups relative to the averages are the highest (shares are respectively 15.5 times, 9.4 times, and 15 times above the average), which means that is the one that gives more preference to these items which includes, for instance, vegetarian pizzas, vegetarian noodles, vegetarian pasta, vegan chips, vegan jelly, vegan chocolate, etc. For another hand, they are the ones who spend the least on vegan alternatives to milk and dairy, like plant-based beverages, vegan cheese, yogurts, butter, etc., based on their average monthly expenditure.

The fourth cluster, *Potential Veggies*, represented by 1.9% of the sample, has customers that are potentially interested in plant-based diets, since customers in this cluster are distinguished by having more preference for almost all the food groups used as inputs, even in vegetarian and vegan alternative proteins to meat and fish, but less comparing to cluster 5. Customers from this cluster spend on these protein alternatives 2.3% of their average monthly expenditure, on average. Nevertheless, this is the cluster that spends more on plant-based milk (7.4%, on average).

The fifth cluster, *Veggie Lovers*, even though is the least representative, with only 0.3% of the sample, is the most important in terms of veggie behavior. The customers that belong to this cluster spend more than the average per month on all food groups, but it is especially noteworthy that they differ completely from the others in terms of the consumption of protein alternatives to meat and fish. The share expenditure on these alternatives is at a very high number above the average (it ranges from 87 to 192.35 times above the average, depending on the five alternatives considered). Besides, it should be noted that this cluster has higher ratios for vegetarian protein alternatives (*MA_vegetarian_Share_Index, PM_vegetarian_Share_Index*) than for vegan proteins (*MA_vegan_Share_Index, PM_vegan_Share_Index*), while in *Potential Veggies* is the opposite. *Veggie Lovers* is also the cluster that has the highest share per month of vegan ready foods. Nevertheless, although they give more importance than the average for all item groups considered in this clustering, especially for vegetarian and vegan proteins, they spend less on the remaining groups compared with *Potential Veggies*, *Convenience and Vegan Sweets* or *Receptive Omnivore*s, depending on the item group considered (alternative to dairy, vegan sweets and desserts, vegan biscuits and cakes or vegan snacks).

Table 3.5: Results from basket clustering

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**Cluster 5**
MA_Tofu_Soy_Seitan_Share_Index (Mean)	0.22	0.78	0.28	22.26	118.1
MA_vegetarian_Share_ Share_Index (Mean)	0.17	1.01	0.22	24.04	120.53
MA_vegan_Index Share_Index (Mean)	0.23	1.4	0.32	25.45	87.09
PM_vegetarian_Share_Index (Mean)	0.14	0.35	12.57	3.89	192.35
PM_vegan_Share_Index (Mean)	0.13	0.47	0.28	23.85	140.42
Dairy_vegan_Share_Index (Mean)	0.72	1.62	0.64	13.27	5.36
Biscuits_Cakes_vegan_Share_Index (Mean)	0.77	8.35	0.68	1.36	3.15
Sweet_Dessert_vegan_Share_Index (Mean)	0.72	1.18	15.03	1.16	1.45
Snacks_vegan_ Share_Index (Mean)	0.83	1.24	9.37	1.09	2.28
Sauces_Creams_vegan_Share_Index (Mean)	0.85	5.84	0.74	1.3	2.46
ReadyFoods_vegetarian_Share_Index (Mean)	0.67	1.47	15.53	2.02	7.72
ReadyFoods_vegan_Share_Index (Mean)	0.18	21.91	0.11	0.94	68.53
Cluster size %	93.2%	2.8%	1.8%	1.9%	0.3%
Cluster name	*Traditional Omnivores*	*Receptive Omnivores*	*Convenience and Vegan Sweets*	*Potential Veggies*	*Veggie Lovers*

3.4.4. Clusters' description

To better understand the shopping behavior of each basket cluster, it was determined the monthly average share of each cluster for different item groups. Table 3.6 represents the groups

that bring more value to the analysis. Additionally, to profile the customers of this basket clustering, the results of the previous clustering, which was performed for the three dimensions, and also the data from *Customers_Segm* table were combined (Figure 3.14 to Figure 3.17), meeting the third research goal (O.3 – Characterize segments)

Table 3.6: Average monthly share spent by 1st item groups and by basket cluster

	Traditional Omnivores	***Receptive Omnivores***	***Convenience and Vegan Sweets***	***Potential Veggies***	***Veggie Lovers***
Protein Alternatives to Meat & Fish	0.02%	0.11%	0.06%	2.31%	10.44%
Fish	11.62%	7.13%	5.35%	8.81%	5.45%
Meat	14.60%	9.22%	7.43%	7.57%	4.81%
Ready Food	1.63%	3.00%	7.49%	2.06%	6.28%
Snacks	0.88%	1.24%	7.44%	1.01%	1.96%
Fruit	5.23%	6.21%	5.07%	6.62%	5.99%
Vegetables	4.07%	5.18%	3.78%	5.75%	6.81%
Pulses	0.49%	0.56%	0.39%	0.64%	0.78%
Nuts	0.47%	0.69%	0.74%	0.74%	0.78%
Dairy & dairy alternatives	9.56%	11.26%	8.83%	16.95%	11.09%
Refined Carbohydrates	4.39%	5.13%	4.61%	5.03%	5.06%
High Fiber Carbohydrates	0.29%	0.66%	0.43%	0.85%	0.98%
Biscuits & Cakes	4.48%	8.35%	5.30%	4.14%	5.04%
Sweets & Desserts	2.89%	3.25%	9.87%	2.80%	2.86%
Sauces & Creams	0.93%	4.00%	0.98%	1.23%	2.09%
Seeds	0.03%	0.06%	0.04%	0.09%	0.13%
Kids	0.32%	0.25%	0.19%	0.23%	0.19%
Supplements	0.05%	0.12%	0.05%	0.13%	0.19%
Yeast	0.00%	0.00%	0.00%	0.01%	0.02%
Drinks	2.41%	2.37%	3.17%	1.76%	1.88%
Alcohol	6.12%	4.29%	4.58%	4.08%	3.39%
Take-Away	3.14%	3.10%	5.04%	2.57%	2.74%
Restaurant	0.01%	0.01%	0.00%	0.01%	0.01%
Pets	1.78%	1.55%	1.41%	2.10%	2.07%
Bazar	2.69%	1.98%	1.71%	2.00%	1.79%
Home Hygiene	5.03%	4.47%	3.30%	4.59%	3.66%
Personal Hygiene	6.81%	6.86%	5.63%	6.82%	5.95%
Textile	0.43%	0.25%	0.19%	0.30%	0.27%
Gast station fuel	0.59%	0.20%	0.10%	0.27%	0.15%

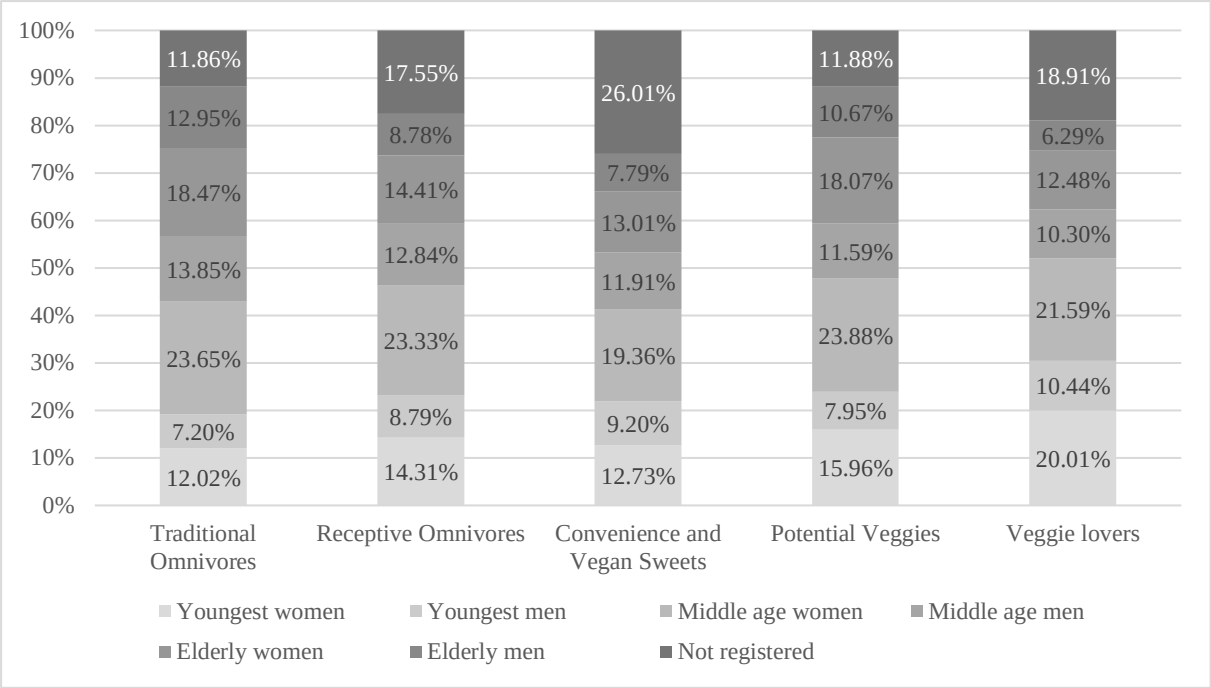

Figure 3.14: Distribution of the clusters from the demographic dimension by basket clusters

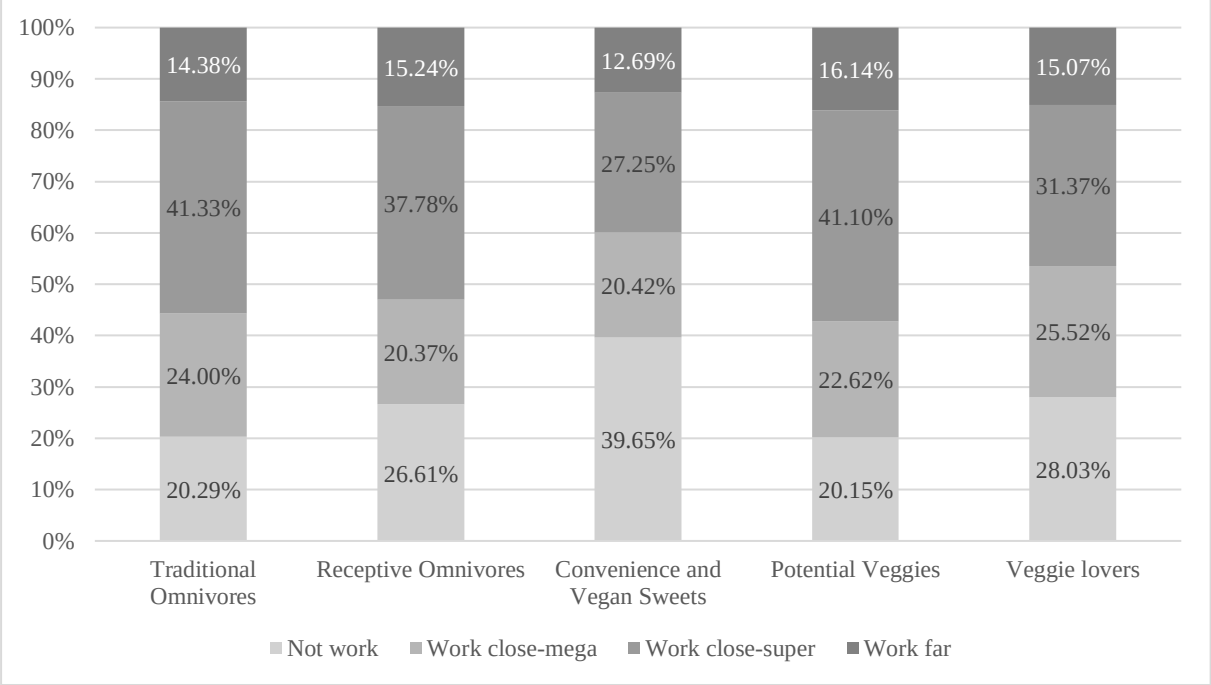

Figure 3.15: Distribution of the clusters from the lifestyle dimension by basket clusters

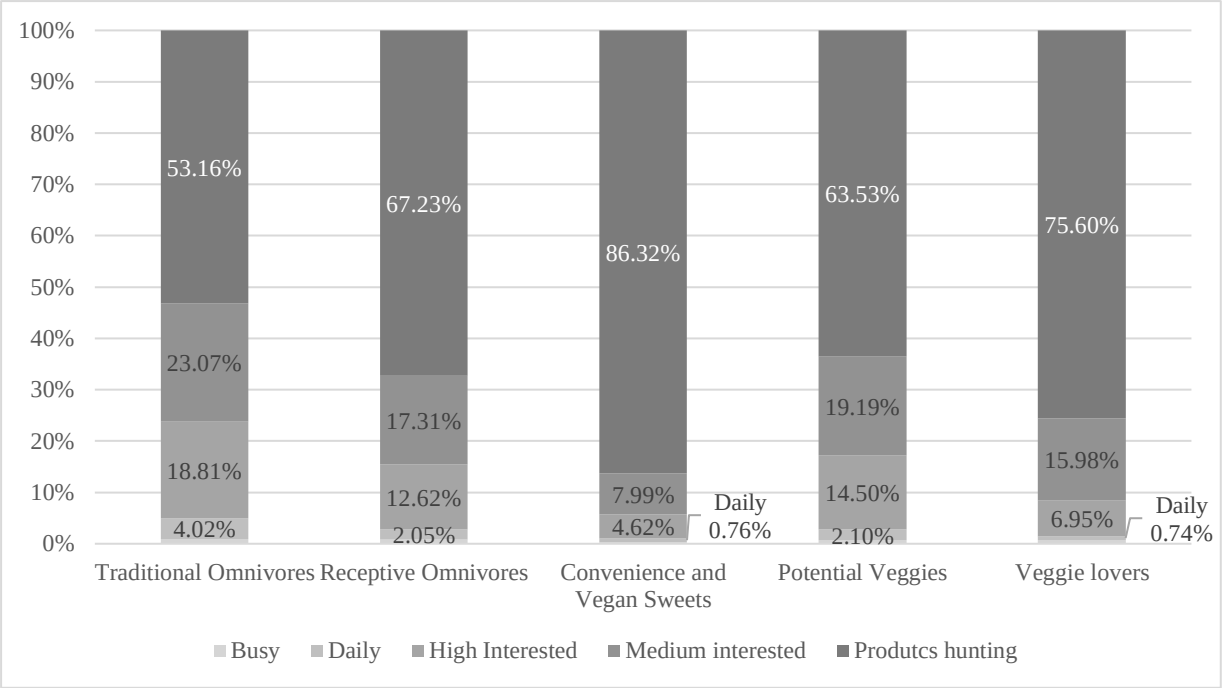

Figure 3.16: Distribution of the clusters of the frequency and monetary dimension by basket clusters [1]

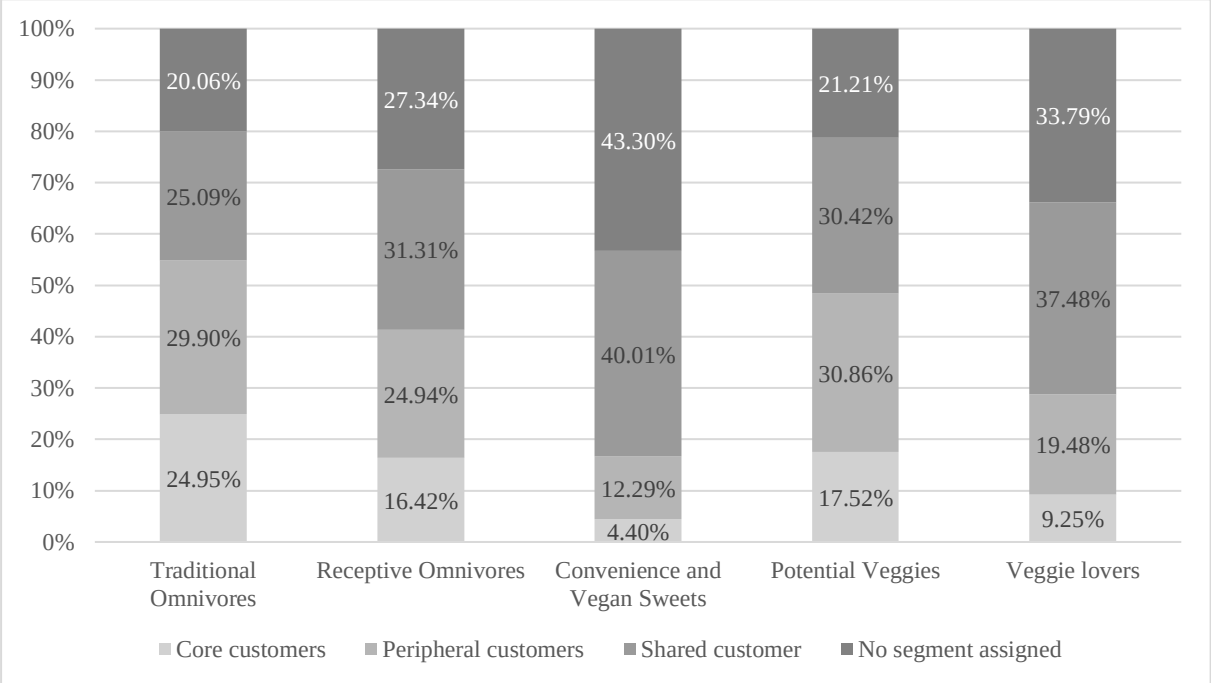

Figure 3.17: Distribution of the segmentation of PD customers by basket clusters

[1] Percentage of *Busy* customers is not displayed on the graph because of their lower value but corresponds to the difference between 100% and the sum of total percentages displayed for each column

Like it was expected, *Veggie Lovers* had the lowest monthly share expenditure on meat (customers from this cluster only spend, on average, 4.8% of their monthly average spending amount on meat), and the percentage of what they monthly spend on fish (5.5%, on average) is higher than on meat. However, where they spend the most, which even exceeds the percentage that is spent monthly on meat and fish, is in the protein alternatives (10.4%) This cluster also contains the customers that have on average, the highest percentages of the amount spend per month on vegetables, pulses, seeds, high-fiber carbs (which include wholegrains, wild rice, oat, etc.), and supplements. Regarding demographic characteristics, the plurality of customers that belong to this cluster are middle age women (21.6%) but is right followed by youngest women (20.0%), while the least percentage of customers are elderly men (6.3%). In fact, from all the clusters it is the one that has more youngers, both men and women. Nevertheless, the demographic characteristics of 18.9% of Veggie Lovers are not known because they did not register. Furthermore, it has more customers that might work close to home and prefer to shop in supermarkets (31.4%), followed by the ones that might not work (28%). Regarding frequency and spending amount, 75.6% of *Veggie Lovers* belong to *Products hunting*, whereas only 7.0% are in the *High interested* and 0.7% in the *Daily* cluster. Regarding the segmentation that is done by PD, 37.5% of *Veggie Lovers* are shared customers (considered less valuable for PD, i.e., it did not buy so many items and spent so much money) and 33.8% were not assigned to a segment.

The *Potential Veggies* even though is the second cluster that spends, on average, more per month on fish (8.8%), is the second that spends less on meat (7.6%) and the share of fish is still higher than meat, similarly to *Veggie Lovers*. It is also the second cluster with the highest average monthly expenditure share on vegetables, pulses, seeds, and high-fiber carbs, and the one with the highest average monthly share on fruit (5.51%) and eggs (1%), and items for pets. This one has more middle-aged women (23.9%) and is followed by elderly women (18.1%). However, is the second cluster with more younger customers, especially younger women (16%). Regarding lifestyles, the percentage of customers that belongs to *not work* cluster (22.8%) and *work close-super* (40.95%) is higher than in *Veggie Lovers*. *Potential Veggies* has also the highest percentage of customers that work far from home (16.41%) compared to all clusters. Finally, in this cluster, 67.8% of customers belong to *Product Hunting*, follows by *Medium interested* (16.8%), and *High interested* (12.6%) clusters, which follows the segmentation from PD, since the plurality are peripheral customers (30.9%), followed by the shared ones (30.4%).

Contrarily, *Traditional Omnivores* are the cluster that spends more per month on meat (14.6%) and fish (11.6%) on average. This cluster also spends more on bazaar items (such as home appliances, technology, books, etc.), hygiene home items, clothes, pharmacy items, and alcohol. In this cluster, like in all clusters except for *Veggies Lovers* and *Convenience and Vegan Sweets*, there are more middle aged women (23.7%) and elderly women (18.5%) but are followed by middle age men (13.9%). Regarding gender, this is the cluster with more percentage of men, while in terms of age the youngest are the least represented. Yet, there is no demographic information from about 11.9% of the customers of this cluster. Compared with *Veggie Lovers*, this cluster has fewer customers that might not work (20.3%) and more that might work close to home and prefer supermarkets (41.5%). Compared to all other clusters this is the one with the lowest percentage of customers that are *Product Hunting* (53.2%) and with the highest percentage in *Medium interested* (23.1%), *High interested* (18.8%), and *Daily* (4%) clusters. Considering the segmentation of PD, this cluster is divided almost equally by segment. Although the plurality is peripheral customers (29.9%), followed by shared customers (25.1%), this is the cluster with the highest number of core customers (25%).

The *Receptive Omnivores* are the second cluster with the highest average monthly expenditure on meat (9.22%), which is in turn higher than their monthly share on fish (7.13%). Yet, they are the third cluster, after *Veggie Lovers* and *Potential Veggies*, that have higher monthly average shares of vegetables, seeds, pulses, and high-fiber carbs. This is also the cluster whose refined carbs share, and personal hygiene share is the highest. Regarding demographic characteristics, like most clusters, it has more women in middle age (23.3%), and in elderly age (14.4%), however, it is not known the information about 17.6% of the customers from this cluster. Considering lifestyle, like *Omnivores* and *Potential Veggies*, it has more customers who might work close to home and prefer supermarkets (36.8%), followed by the ones that might not work (27.4%). Lastly, like the other clusters, most of these customers belong to *Products hunting* (69.6%), followed by the *Medium interested* (17.9%), and *High interested* (10.4%). Likewise, more customers are shared customers (31.3%), followed by customers without a segment assigned (27.3%), and by the core customers (24.9%).

Finally, the *Convenience and Vegan Sweets* cluster have already the highest share expenditure either on vegan, vegetarian, or non-veg convenience items (ready foods and snacks), sweets and desserts, and items that are sold in the takeaway area. It is also the cluster that spends more on drinks, such as fruit juices and soft drinks, and the second one to have a high share on alcoholic drinks. Regarding demographic characteristics, it is the only cluster where the plurality is customers that did not register (26%). Middle age women are the second

biggest group (19.4%), but their percentage is the lowest compared to all the basket clusters. Then, is followed by the youngest women (12.9%) and middle age man (11.2%). The least is the elderly men (8.5%). Considering lifestyle and compared to other clusters, the *Convenience and Vegan Sweets* cluster has the highest percentage of customers that might not work (39.7%), and the lowest of customers that might work far from home (12.7%). For another hand, this is the cluster with the highest percentage of customers that belong to *Products hunting* cluster (86.3%), which is in line with the segmentation of PD. Almost the majority of the customers from this cluster (46.9%) had not a segment assigned and the other 40% were considered shared customers. Analyzing the shopping behavior and lifestyle of these customers, it is observed that this cluster likes to buy more convenience items so it might be expected that the customers whose demographic information is not known are also younger.

Subsequently, it was verified how the transactions that had been analyzed for the PD customers, were distributed by each cluster. As expected, in general, *Veggie Lovers* had more customers who purchased at least once any of the three types of protein alternatives to meat & fish considered in this work (item group 1) (84.7%), followed by *Potential Veggies* (62.6%) (Figure 3.18). But curiously, while 46.4% of these customers from *Veggie Lovers* repeated the purchase, it was the *Potential Veggies* who had more percentage of repeat buyers of these items (51.2%) (Figure 3.19).

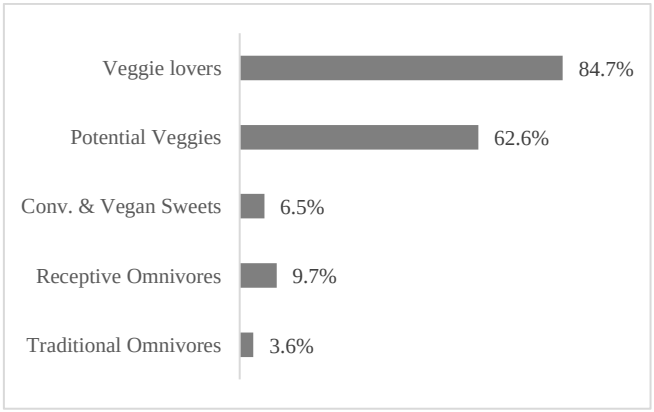

Figure 3.18: Distribution of customers who purchased at least once a protein alternative to meat & fish, by basket clustering, considering the whole 6-month period

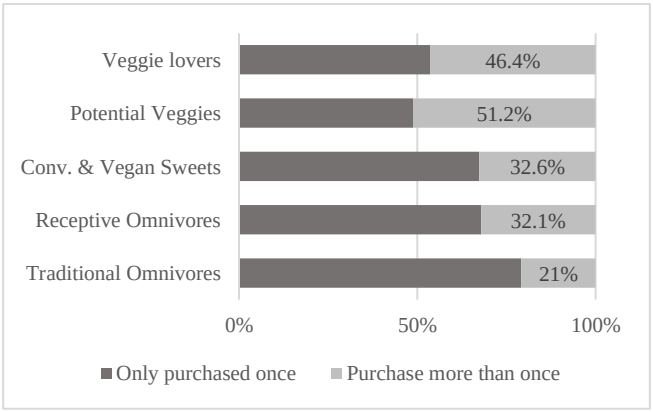

Figure 3.19: Distribution of customers by basket clustering and behavior related to protein alternative to meat & fish, considering the whole 6-month period

Furthermore, it was previously observed that, compared to soy and tofu, seitan was the option with the highest percentage of repeat customers. Now, it could be verified that this is the alternative repeated by more customers in all clusters, except in the *Convenience and Vegan Sweets* cluster (Appendix I). In other words, it is not only among those who are more interested

in plant-based products (veggies) that there is a higher percentage of repeat customers for seitan than for tofu and soy but also among omnivores. Conversely, the option that more customers from *Convenience and Vegan Sweets* repeated was tofu (22.6%), while seitan had the least percentage of customers who repeated the purchase (14.7%). The pattern repeats for processed meat alternatives (vegetarian/vegan nuggets, hamburgers, chili, etc.) and alternatives to highly processed meat (vegetarian/vegan *alheira*, sausage, etc), with the *Veggie Lovers* cluster having the highest number of customers buying at least once these alternatives, but being the *Potential Veggies* those who repeat this purchase the most, although these percentages for the two clusters are close. The only exception goes to the vegetarian alternatives to highly processed meat, in which it was found that more customers from the *Veggie Lovers* repeated the purchase (45.4%) than in the *Potential Veggies* (34.4%).

For vegan dairy alternatives, as expected, *Potential Veggies* had in general more customers who purchased at least once these alternatives, and had also a higher percentage of repeated purchasers, compared to other clusters. The exceptions are for vegan cheese (since it was *Veggie Lovers* that had the highest percentage of customers that purchased it at least once and also that repeated the purchase) and for vegan ice cream (because it was bought by more customers from *Veggie Lovers* (5.4%) than from *Potential Veggies* (5.3%), even though this last cluster had higher percentage of repeat buyers (33%) than *Veggie Lovers* (32.3%)).

3.5. Profiling and Deployment

After customer segments were identified in this project, it was necessary to create a model that would allow JM to label new customers or even the same customers but in a different period, (since their behavior may change) considering the identified clusters. Depending on the technique, iterations number, and dataset where is applied, clustering can lead to different results. To overcome this problem, a decision tree model (Yadav, 2019) was created that used original input variables from basket clustering as predictors and the cluster labels as predicted classes (Waisakurnia, 2020; User11852, 2017). It should be noted that this model was created for a descriptive rather than predictive purpose, to allow the creation of a set of rules that would classify customers based on the previous basket clustering.

First, to estimate models, the sample was partitioned into a train and a test sample, considering different proportions (60% vs 40%; 70% vs 30%; 80% vs 20, respectively) (Bronshtein, 2017). After, since data was unbalanced, it was balanced by reducing records from the classes (clusters) with higher weight in the sample, as it is recommended by several

researchers (Stewart, 2020), to all clusters have equal proportions. Given that this was only meant to be an estimation phase, both sets were balanced, including the test set to check how the model would behave for similar data to that submitted in the training phase. The two most common algorithms to create decision tree models were considered in this analysis, CART, and CHAID (Ramzai, 2020), and to which special attention was given, but the C5.0 and QUEST algorithms were also checked (Pahn, 2020).

Different models were then generated, across the different partitions and algorithms, considering different stop criterion rules, namely, the maximum depth of the tree and the minimum number of records for a *parent* node and *child* node (Ramzai, 2020; Yadav, 2019). The selection of the best model was based on the parsimony principle (a simpler model, i.e. with fewer rules, easier to implement by the company and that could be more possible to generalize) (Iluemeo, 2021), and on analytic metrics through the confusion matrix, such as accuracy, sensitivity, specificity, and precision (Markham, 2014). As previously mentioned in this work, the best statistical model is not always the best model for the business. The goal is to have a model that describes these customers but that is also able to generalize its classification to these and future customers.

In general, the CART algorithm generated better models, with fewer rules (less complex) and higher values on the analytic metrics, than the others. Thus, a model generated by this algorithm, which was created by using a partition with 70% for the train and 30% for the test, was selected. This final model (appendix J) had 24 rules and correctly classified (overall accuracy) 89% of customers from the train set and 89.7% from the test set. In the train set, of the classifications that the model was making, the percentage of customers that the model was getting right (precision) varied between 79.7% and 98%, according to the classes considered. On the other hand, of the customers that belonged to each class, the percentage that was correctly classified, i.e., the percentage of customers that the model was actually identifying (sensitivity) ranged from 69.6% to 97.2%. Conversely, the percentage of customers that were correctly predicted as not belonging to each class (specificity), ranged from 94.2% to 99.5%. For the test set, the results were not very different. This model was preferred to others that had better overall results but were more complex, and to others that were less complex but had lower values for analytic metrics.

Afterward, the selected model was applied to the original dataset (which was unbalanced) and its results were evaluated. The model correctly classified 96.9% of the customers (overall accuracy). Considering the different classes, sensitivity ranged between 69.9% and 97.3%, specificity between 98.1% and 99.8%, and only precision had lower values, from 48.6% to

99.9%. It is important to remember that this classification was based on an exploratory and subjective process that is clustering, so predictions that did not correctly match the identified clusters were not necessarily a very serious problem. The goal was to find a set of rules that would make sense from a business perspective, going along with the identified clusters, for the business to use to classify its customers and identify the customers more similar to what is a veggie.

Lastly, it must be noted that the Deployment phase, in addition to the delivery of a model to JM, also corresponds to the delivery of this thesis and the report of the results, so that the company can take actions that will generate value for its business.

3.6. Recommendations & Strategy

After all these analyses, to meet the fourth research objective (O.4 – Define strategies for each segment to increase the frequency or spending) and to support recommendations with data, it was seen how the centroid (mean) of the identified clusters positioned in space regarding their monthly frequency (measured by the number of transactions) and amount spent per transaction. As can be seen in Figure 4.1, compared to the average, *Traditional Omnivores* are the cluster with the highest frequency and expenditure, followed by *Potential Veggies* and *Receptive Omnivore*s. In the opposite quadrant, *Convenience and Vegan Sweets* are the ones with the lowest frequency and amount spent. *Veggie Lovers* have a frequency lower than the average (but higher than *Convenience and Vegan Sweets*) and approximately an average amount spent per transaction.

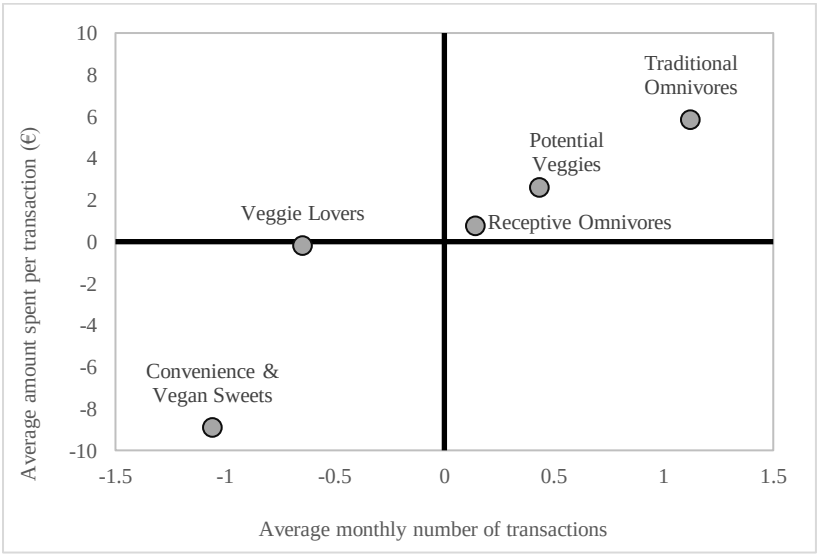

Figure 4.1: Cluster Matrix- Frequency vs amount spent

To define strategies for some clusters to increase their frequency or amount spent (depending on the case), it was distinguished within each cluster the customers who had higher frequency/amount spent than the rest of the customers in the same cluster, in order to identify what led the customers in that cluster to go/buy more than the others. Then, it was also created decision tree models following the same methodology as it was described in the deployment section (balance the data, partition in train and test sets, and application of the model to the original data set). However, the targets of the models were flag variables (higher vs lower frequency/amount spent than the average, depending on the case, for each cluster), and the variables tested as input were indices of the average monthly share of the item groups from the 1st group list of item (Appendix E), calculated as a ratio of the share to the mean of each item group (but with exception of the item groups that included the same variables used in the basket clustering) and it was also considered other variables related with demography and lifestyle. The purpose was to identify the rules and key attributes that differentiate the customers and maximized the higher frequency/amount spend within clusters.

Regarding recommendations around veggie customers and the market of plant-based products, *Veggie Lovers* have an interest in all types of vegan and vegetarian products so they could be a target for marketing campaigns for all these products, but especially for plant-based alternatives to meat and fish, because this is the cluster with the most interest in these protein alternatives. However, these customers go less frequently to PD, so the retailer should beware of this. The lower frequency of this cluster might explain why despite of containing more customers who purchase alternatives to meat and fish, the percentage of customers who repeat the purchase of these items, is lower compared to *Potential Veggies*. Thus, if PD takes action to increase the frequency of these customers, sales of these products will most probably raise. For this reason, in this case, it was created a decision tree model whose target assumed two values (1- if customers from *Veggie Lovers* had a higher frequency than the average of the cluster; 0- if not) and it was created models. Thus, it was possible to verify that customers that purchased at least something of condiments, but whose share of these items was not more than 0.87 times the average share of this item group, or customers that bought at least something from take-way, represented 7.8% of the *Veggie Lovers* and had 6.3 more chances than the others of purchase with a frequency higher than the average. Similarly, it was found a rule which was the following: Veggie Lovers that purchase anything of condiments but whose average share is not more than 1.55 times the average and buy at least something of take-away items, bazar items and nuts, then are 12.4 times more likely to be frequent buyers, even though this rule only represent around 5% of the customers from this clusters. Then one strategy that PD may follow

to increase the frequency of *Veggie Lovers* is to give discounts to these customers on condiments, products from takeaway, bazar, and nuts.

Considering *Potential Veggies*, PD should have special attention to them to maintain their loyalty because, regarding the veggies, is the best cluster that is positioned. Its customers have already higher frequency and amount spent, even though they could be even better, like *Traditional Omnivores*. Since these customers shops more frequently and spent more, they might be more likely to be receptive to PD campaigns. The *Potential Veggies* have an interest in these protein alternatives to meat and fish, although not as much as the *Veggie Lovers*, but are more loyal than them, so PD can also consider *Potential Veggies* as a target to plant protein campaigns. Moreover, if PD wants to create a campaign for vegan alternatives to dairy (plant-based beverages, vegan yogurts, vegan ice cream, vegan margarine, etc.) this cluster is the best target, since these customers have the highest interest in dairy alternatives.

If PD wants to do a campaign regarding sweets and snacks, customers from *Convenience and Vegan Sweets* are the best target. Even though, in general, they are not so interested in proteins alternatives, except for alternatives to highly processed meat, is the cluster whose percentage of the average monthly spending on sweets and snacks like vegan jelly, vegan gummies, or vegan chips made from, for instance, green pea, hummus, lentils, among others, is the highest. On the other hand, this cluster has the highest percentage of customers that might not be so interested to purchase in PD. So maybe they should not be a priority to PD when the chain wants to impact customers to purchase these items. Yet, PD may offer limited-time offers on the items that they like, which creates a sense of urgency and possibly makes them return.

Even though omnivores are not the focus of this project, recommendations can also be made concerning them. *Traditional Omnivores* are, in general, the best customers, so PD should give particular attention to them as well. For *Receptive Omnivores*, the main goal is to increase the amount that they spend. Using also a decision tree model but having now as a target the fact that a *Receptive Omnivores* could have an amount spent higher than the average of the group or a lower spending amount, it was found a rule that represented around 3% of these customers and showed that if the monthly share of a customer for fish is between 0.7 and 1.7 times the average, the bazar share no higher than 0.88, and the share for pulses, personal hygiene, and meat is, respectively, higher than 0, 0.54 and 0.86 times the average share of these item groups, then customers have 2.3 more chances to spend more amount than the average of the cluster. Hence, PD could give coupons on these items which may attract the customers that spend less but have similar characteristics, since they belong to the same cluster of the *Receptive Omnivores* who spend more.

4. Conclusions and Discussion

Recent studies have shown that plant-based diets are becoming more popular among the population, and Portugal is not an exception. With the growing interest in plant-based products and the increasing competition among retailers to offer these alternatives, this project aimed to help the PD chain to gain more knowledge about its customers to increase its competitiveness in this market sector. At the same time, this project allowed a comparison of actual behavior based on transactional data with declarative data from previous studies.

Firstly, the creation of a nomenclature to identify vegetarian and vegan articles and the adaptation of the marketing structure with the creation of new item groups was extremely crucial to reveal behaviors associated with veggies.

Secondly, based on a clustering technique, the PD customers most likely to be veggies or to be interested in plant-based products, as well as those who might follow an omnivore diet, were identified, and profiled based on their basket mix and on three dimensions: demographic, lifestyle and frequency and monetary. The results from this work are curious since different levels of interest were found for different vegetarian and vegan items. Depending on the plant-based products that PD wants to promote, different customers should be considered as a principal target.

Even with the possibility that customers might not do all the shops in the PD chain, *Veggie Lovers* is the most similar cluster for what are real veggie consumers. These customers follow a purchasing pattern that is characteristic of individuals who reduce or try to eliminate meat consumption, with more preference for vegetables, seeds, pulses, and high-fiber carbohydrates than other clusters, as was analyzed by several authors (Malek & Umberger, 2021, Niva and Vainio, 2021; Koch et al., 2019). Nevertheless, meat still has a weight in their monthly average, although this share is the smallest compared to other clusters. However, this is in line with the findings of Apostolidis and McLeay (2019), who say that vegetarians purchase regularly or occasionally meat products for family or guests. In fact, this cluster seemed to include more vegans and vegetarians than flexitarians (because had a higher average monthly expenditure share of protein alternatives than of meat and fish). However, it might include more vegetarian consumers since their preference for vegetarian protein alternatives was higher (higher average monthly expenditure shares) than for vegan alternatives.

Potential Veggies is the cluster that contains customers that are also interested in plant-based proteins, but not so much as *Veggie Lovers*. Also, although the cluster of *Potential Veggies* seems to have more customers who might be vegan than the *Veggie Lovers* since they

have a higher preference for vegan protein alternatives, it must have a higher proportion of flexitarians, because meat and fish have a higher share on their monthly expenditure and have lower share on vegetables, seeds, or pulses, compared to *Veggie Lovers*. So, they might just be at the beginning of a journey to change their eating patterns or are just open to try new these new alternatives. On another hand, this is the cluster that, in general, contains a higher percentage of customers who repeat at least once the purchase of these alternatives. However, this might be explained by the fact that *Potential Veggies* go more frequently to PD than *Veggie Lovers*.

Customers from *Convenience and Vegan Sweets* were found to have the most interest in vegan sweets, sauces, creams to spread, and ready meals, but they also have interest in items of the same type, that are not only vegan. In general, they go casually to the PD to buy more convenience or processed items, going less and spending less than the other clusters.

Finally, it was found the clusters of *Traditional Omnivores* and *Receptive Omnivores*, which corresponds to the majority of the sample of this project and represented behaviors of customers with less interest in these alternatives and who follow an omnivore diet, with more meat and fish. Despite the very low interest of *Receptive Omnivores* in plant-based products, it was slightly higher than the one of *Traditional Omnivores*.

Regarding other findings, it was possible to verify in this project that, even though women were more present in all clusters, it is young people, and especially younger women (or at least members of their families) that spend more of their monthly expenditure on plant-based products. This is in line with the literature since vegetarians and vegans are more women and youngers (Hielkema & Lund, 2021, Malek et al., 2019). Additionally, Lantern (2021) found out that within veggies more people have a cat or a dog, than within omnivores. Likewise, *Potential Veggies* and *Veggie Lovers* are the clusters that, interestingly, have a higher monthly average share expenditure on food or accessorizing for pets, than other clusters.

To conclude, it was possible to verify different levels of interest in plant-based products between PD's customers and to confirm that what has been found and declared in previous studies by flexitarians, vegetarians and vegans can be also observed through transactional and demographic data of this Portuguese retailer. Moreover, it was possible to point out some guidelines for the initial research question of this project (R.Q. How to generate value in the plant-based food products market?). Recommendations were made and strategic actions were defined that the PD should take to generate value in its plant-based products market and to improve the loyalty of some of the clusters found.

5.1. Research Limitations and Future Work

As with most studies, this project also has some limitations. Firstly, it should be noted that due to limitations in the access of complete information about each item (e.g., not possible to clearly identify the item by the description or no information available on the internet about the item's ingredients) there may be items whose classification in terms of vegetarian nomenclature were not completely correct (although validated by PD).

Another limitation is the fact that besides each loyalty card has a single person as titular, other members of the family might also use it to shop. Furthermore, customers may not do all their shopping at this retailer. According to Nielsen (2022), in 2021 PD was the second chain where households spent more, corresponding to 23.7% of their expense from that year. For this reason, it must be noted that all the conclusions from this project were only based on behavioral data that the analyzed customers had at PD (truncated distribution). Moreover, the analysis was limited to the offer of vegetarian and vegan products that existed in the analyzed period. Hence, based on shopping behavior, it is not possible to say with certainty that a customer is a veggie consumer, but rather that maybe he or his family exhibit characteristic patterns of those who follow plant-based diets.

Also, since only transactional data were available, in the absence of attitudinal data, opinions, preferences, and lifestyle, analyses may face the problem of omitted variables. Therefore, the potential next steps are for PD to develop an exhaustive market study to have these data, such as opinions, preferences, and lifestyle and create models to map/classify the loyal customer base, also considering these attributes.

For future work, PD should apply the classification model that was created to the same customers in another time frame, to compare how their behavior changes (including creating predictive models for these segment migrations), and to new customers, to identify their profiles. The classification should be performed regularly, based on aggregated data from six-month periods. After some time, when PD think that data has changed significantly, they could repeat clustering and consequent creation of the classification model.

Also, the identified customers that are more receptive to plant-food products should be targets of more analysis. It could be created a model to predict if customers will adhere well to market campaigns of plant-based products. It could be also interesting to identify different types of veggies based on the way they prefer to purchase specific plant-based items (e.g., fresh fruit *vs* packed fruit, canned pulses *vs* dried pulses, frozen vegan/vegetarian alternatives to meat and fish *vs* refrigerated *vs* neither, etc.) or even include the identification of non-food items as being

suitable for customers that follow the veganism (like clothes, creams, detergents, among others items that are vegan).

Moreover, similar works that want to try to identify veggie customers in a retail environment could repeat this project in other supermarkets. Also, they could include other variables and use other analytic techniques.

5. References

Abreu, T (2021). *Ranking dos retalhistas*. Associação Vegetariana Portuguesa. https://avp.org.pt/wp-content/uploads/2021/10/RankingVeg2021_Estudo.pdf

Achados Veganos (n.d.). *Home* [Facebook page]. Facebook. Retrieved February 07, 2022, from https://pt-pt.facebook.com/achadosveganos/

Andersson, O., & Nelander, L. (2021). Nudge the Lunch: A Field Experiment Testing Menu-Primacy Effects on Lunch Choices. *Games*. 12(1), Article 2. https://doi.org/10.3390/g12010002

Apostolidis, C., & McLeay, F. (2019). To meat or not to meat? Comparing empowered meat consumers' and anti-consumers' preferences for sustainability labels. *Food Quality and Preference*. 77, 109-122. https://doi.org/10.1016/j.foodqual.2019.04.008

Associação Vegetariana Portuguesa. (2021, March 6) *O que é o vegetarianismo*?. https://www.avp.org.pt/o-que-e-o-vegetarianismo/

Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*. 1(1), 363–376. https://doi.org/10.1162/qss_a_00018

Bronshetein, A. (2017, May 17). *Train/Test Split and Cross Validation in Python*. Towards Data Science. https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6

Bryant, C., & Sanctorum, H. (2021). Alternative proteins, evolving attitudes: Comparing consumer attitudes to plant-based and cultured meat in Belgium in two consecutive years. *Appetite*. 161, Article 105161. https://doi.org/10.1016/j.appet.2021.105161

Bryant, C., Szejda, K., Parekh, N., Desphande, V., & Tse, B. (2019). A Survey of Consumer Perceptions of Plant-Based and Clean Meat in the USA, India, and China. *Frontiers in Sustainable Food Systems*. 3, Article 11. https://doi.org/10.3389/fsufs.2019.00011

Bullock, K., Lahne, J., & Pope, L. (2020). Investigating the role of health halos and reactance in ice cream choice. *Food Quality and Preference*. 80, Article 103826. https://doi.org/10.1016/j.foodqual.2019.103826

Byrne, C. (2019, November 20). *What's the Difference Between Tofu, Tempeh and Seitan?*. My Fitness Pal. https://blog.myfitnesspal.com/whats-the-difference-between-tofu-tempeh-and-seitan/

Centro Vegetariano (2020, December 12). *120 000 vegetarianos - Número quadruplica em 10 anos*. https://www.centrovegetariano.org/Article-620-Numero-vegetarianos-quadruplica-10-anos-Portugal.html

Cheah, I., Shimul, A. S., Liang, J. H., & Phau, I. (2020). Drivers and barriers toward reducing meat consumption. *Appetite*. 149, Article 104636. https://doi.org/10.1016/j.appet.2020.104636

Contini, C., Boncinelli, F., Marone, E., Scozzafava, G., & Casini, L. (2020). Drivers of plant-based convenience foods consumption: Results of a multicomponent extension of the theory of planned behaviour. *Food Quality and Preference*. 84, Article 103931. https://doi.org/10.1016/j.foodqual.2020.103931

CTT (2022, May 26). *Códigos Postais- Faça download dos ficheiros* [Data set]. https://www.ctt.pt/feapl_2/app/restricted/postalCodeSearch/postalCodeDownloadFiles.jspx

Culliford, A., & Bradbury, J. (2020). A cross-sectional survey of the readiness of consumers to adopt an environmentally sustainable diet. *Nutrition Journal*. 19(1), Article 138. https://doi.org/10.1186/s12937-020-00644-7

de Koning, W., Dean, D., Vriesekoop, F., Aguiar, L. K., Anderson, M., Mongondry, P., Oppong-Gyamfi, M., Urbano, B., Luciano, C. A. G., Jiang, B., Hao, W., Eastwick, E.,

Jiang, Z., & Boereboom, A. (2020). Drivers and Inhibitors in the Acceptance of Meat Alternatives: The Case of Plant and Insect-Based Proteins. *Foods*. 9(9), Article 1292. https://doi.org/10.3390/foods9091292

Didinger, C., & Thompson, H. (2021). Motivating Pulse-Centric Eating Patterns to Benefit Human and Environmental Well-Being. *Nutrients*. 12(11), Article 3500. https://doi.org/10.3390/nu12113500

European Vegetarian Union (2019). *Traces of animal substances in vegan/vegetarian food* [Position statement]. https://www.euroveg.eu/wp-content/uploads/2021/02/072019_EVU_PP_Traces.pdf

Figueira, N., Curtain, F., Beck, E., & Grafenauer, S. (2019). Consumer Understanding and Culinary Use of Legumes in Australia. *Nutrients.* 11(7), Article 1575. https://doi.org/10.3390/nu11071575

Gallagher C. T., Hanley P., & Lane, K. E. (2022). Pattern analysis of vegan eating reveals healthy and unhealthy patterns within the vegan diet. *Public Health Nutrition*. 25(5), 1310-1320. https://doi.org/10.1017/S136898002100197X

Garnett, E. E., Balmford, A., Sandbrook, C., Pilling, M. A., & Marteau, T. M. (2019). Impact of increasing vegetarian availability on meal selection and sales in cafeterias. *Proceedings of the National Academy of Sciences of the United States of America*. 116(42), 20923-20929. https://doi.org/10.1073/pnas.1907207116

Garnett, E. E., Marteau, T. M., Sandbrook, C., Pilling, M. A., & Balmford, A. (2020). Order of meals at the counter and distance between options affect student cafeteria vegetarian sales. *Nature Food*. 1(8), 485-488. https://doi.org/10.1038/s43016-020-0132-8

Gibbs, J., & Cappuccio, F.P. (2022). Plant-Based Dietary Patterns for Human and Planetary Health. *Nutrients*. 14(8), 1614. https://doi.org/10.3390/nu14081614

Gomez-Luciano, C. A., de Aguiar, L. K., Vriesekoop, F., & Urbano, B. (2019a). Consumers' willingness to purchase three alternatives to meat proteins in the United Kingdom, Spain, Brazil and the Dominican Republic. *Food Quality and Preference*. 78, Article 103732. https://doi.org/10.1016/j.foodqual.2019.103732

Gomez-Luciano, C. A., Vriesekoop, F., & Urbano, B. (2019b). Towards Food Security of Alternative Dietary Proteins: A Comparison Between Spain and the Dominican Republic. *Amfiteatru Economic*. 21(51), 393-407. DOI: 10.24818/EA/2019/51/393

Grasso, A. C., Hung, Y., Olthof, M. R., Brouwer, I. A., Verbeke, W. (2021). Understanding meat consumption in later life: A segmentation of older consumers in the EU. *Food Quality and Preference.* 93, Article 104242. https://doi.org/10.1016/j.foodqual.2021.104242

Grasso, A. C., Hung, Y., Olthof, M. R., Verbeke, W., & Brouwer, I. A. (2019). Older Consumers' Readiness to Accept Alternative, More Sustainable Protein Sources in the European Union. *Nutrients*. 11(8), Article 1904. https://doi.org/10.3390/nu11081904

Griva, A., Bardaki, C., Pramatari, K., & Papakiriakopoulos, D. (2018). Retail business analytics: Customer visit segmentation using market basket data. *Expert Systems with Applications*. 100, 1-16. https://doi.org/10.1016/j.eswa.2018.01.029

Hargreaves, S. M., Raposo, A., Saraiva, A., & Zandonadi, R.P. (2021). Vegetarian Diet: An Overview through the Perspective of Quality of Life Domains. *International Journal of Environmental Research and Public Health.* 18(8), 4067. https://doi.org/10.3390/ijerph18084067

Healthline. (2017, June 4). *Vegan vs. Vegetarian: What's the Difference?*. https://www.healthline.com/nutrition/vegan-vs-vegetarian

Henn, K., Goddyn, H., Olsen, S. B., & Bredie, W. L. P. (2022). Identifying behavioral and verainattitudinal barriers and drivers to promote consumption of pulses: A quantitative survey across five European countries. *Food Quality and Preference*. 98, Article 104455. https://doi.org/10.1016/j.foodqual.2021.104455

Hielkema, M. H., & Lund T. B. (2021). Reducing meat consumption in meat-loving Denmark: Exploring willingness, behavior, barriers and drivers. *Food Quality and Preference*. 93, Article 104257. https://doi.org/10.1016/j.foodqual.2021.104257

Hopwood, C. J., Bleidorn, W., Schwaba, T., & Chen, S. (2020). Health, environmental, and animal rights motives for vegetarian eating. *Plos One*. 15(4), e0230609. https://doi.org/10.1371/journal.pone.0230609

IBM. (2021, March 4). *TwoStep Cluster Node*. https://www.ibm.com/docs/vi/spss-modeler/18.0.0?topic=models-twostep-cluster-node

Ilumeo. (2021, August 17). *Parcimônia e o segredo de bons modelos de data science.* https://ilumeo.com.br/todos-posts/2021/08/17/parcimonia-e-o-segredo-de-bons-modelos-de-data-science

INE. (2014). *Tipologia de áreas urbanas* (Version V03486) [Data set]. https://smi.ine.pt/Versao/Detalhes/3486#Correspond%C3%AAncias

Jerónimo Martins (2022a, March 22). *Jerónimo Martins ascende ao 49.º lugar entre os maiores retalhistas do Mundo*. https://www.jeronimomartins.com/pt/press_releases/pr_20220302_1_pt/

Jerónimo Martins (2022b, March 28). *Apresentação Institucional* [PowerPoint slides]. https://www.jeronimomartins.com/wp-content/uploads/01-DOCUMENTS/Corporate-Presentations/JM_Apresentacao-Institucional_MAR2022.pdf

Johnson, K. (2021, May 16). *Which E Numbers Are Vegan?*. Elated Vegan Supplement. https://elatedvegansupplements.com/which-e-numbers-are-vegan/

Koch, F., Heuer, T., Krems, C., & Claupein, E. (2019). Meat consumers and non-meat consumers in Germany: a characterisation based on results of the German National Nutrition Survey II. *Journal of Nutritional Science*. 8, Article e21. https://doi.org/10.1017/jns.2019.17

Kopplin, C. S., & Rausch, T. M. (2021). Above and beyond meat: the role of consumers' dietary behavior for the purchase of plant-based food substitutes. *Review of Managerial Science*. https://doi.org/10.1007/s11846-021-00480-x

Lacroix, K., & Gifford, R. (2019). Reducing meat consumption: Identifying group-specific inhibitors using latent profile analysis. *Appetite*. 138, 233-241. https://doi.org/10.1016/j.appet.2019.04.002

Lantern (2021, November). *The Green Revolution Portugal 2021* [PowerPoint slides]. Webflow.https://uploads-ssl.webflow.com/5a6862c39aae84000168e863/618ced72b10f1c8646891c8d_Reporte%20The%20Green%20Revolution%20Portugal_final.pptx.pdf

Malek, L., & Umberger, W. J. (2021). Distinguishing meat reducers from unrestricted Omnivores, vegetarians and vegans: A comprehensive comparison of Australian consumers. *Food Quality and Preference*. 88, Article 104081. https://doi.org/10.1016/j.foodqual.2020.104081

Malek, L., Umberger, W. J., & Goddard, E. (2019). Committed vs. uncommitted meat eaters: Understanding willingness to change protein consumption. *Appetite*. 138, 115-126. https://doi.org/10.1016/j.appet.2019.03.024

Marie, A. (2022, February 19). *Water Footprint Of Food List*. HEALabel. https://www.healabel.com/water-footprint-of-food-list/

Markham, K. (2014, March 25). *Simple guide to confusion matrix terminology*. Data School. https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

Martinelli, E., & De Canio, F. (2021). Purchasing veg private labels? A comparison between occasional and regular buyers. *Journal of Retailing and Consumer Services*. 63, Article 102748. https://doi.org/10.1016/j.jretconser.2021.102748

Martínez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernández-Orallo, J., Kull, M., Lachiche, N., Ramírez-Quintana, M. J., & Flach, P. (2019). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*. 33(8), 3048-3061. doi: 10.1109/TKDE.2019.2962680.

Martins, A. S., Ponte, A. L., Mousinho, C., Bragança, F., Hergy, F., Guerra, L., Pedro, M., Silva, M., Duarte, S., & Araújo, V. (2017). *Suplementos alimentares: O que são e como notificar reações adversas.* Infarmed. 21(3). https://www.infarmed.pt/documents/15786/1983294/Boletim%2Bde%2BFarmacovigil%FF%FFncia%2C%2BVolume%2B21%2C%2Bn%FF%FF3%2C%2Bmar%FF%FFo%2Bde%2B2017/89d99edd-fb8c-4042-8a38-8d1bc5a555c7

McElfresh, J. (2021, September 16). *EuroCommerce joins EVU and FoodDrinkEurope in their call for EU harmonised vegan and vegetarian definitions*. European Vegetarian Union. https://www.euroveg.eu/eurocommerce-evu-fooddrinkeurope-call-for-harmonised-vegan-and-vegetarian-definitions/

Mehmet, S. (2020, August 12). *Study reveals why wheat and gluten intolerance is becoming more common*. New Food. https://www.newfoodmagazine.com/news/115778/study-reveals-why-wheat-and-gluten-intolerance-is-becoming-more-common/

Mingo, M. (2021, December 21*). Veganos podem tomar medicamentos e suplementos em cápsula?*. Boa Forma. https://boaforma.abril.com.br/equilibrio/veganos-medicamentos-esuplementos/

Morris, M. A., Wilkins, E. L., Galazoula, M., Clark, S. D., & Birkin, M. (2020). Assessing diet in a university student population: a longitudinal food card transaction data approach. *British Journal of Nutrition*. 123(12), 1406-1414. https://doi.org/10.1017/S0007114520000823

Nielsen (2022). Anuário Nielsen Food 2021. *Biblioteca Iscte*. https://catalogo.biblioteca.iscte-iul.pt/cgi-bin/koha/opac-detail.pl?biblionumber=115200

Niva, M., & Vainio, A. (2021). Towards more environmentally sustainable diets? Changes in the consumption of beef and plant-and insect-based protein products in consumer groups in Finland. *Meat Science*. 182, Article 108635. https://doi.org/10.1016/j.meatsci.2021.108635

Onwezen, M. C., Bouwman, E. P., Bouwman, E. P., & Reinders, M. J. (2021). A systematic review on consumer acceptance of alternative proteins: Pulses, algae, insects, plant-based meat alternatives, and cultured meat. *Appetite*. 159, Article 105058. https://doi.org/10.1016/j.appet.2020.105058

Page M. J., McKenzie J. E., Bossuyt P. M., Boutron I., Hoffmann T. C., Mulrow C. D., Shamseer L., Tetzlaff J. M., Akl E. A., Brennan S. E., Chou R., Glanville J., Grimshaw J. M., Hróbjartsson A., Lalu M. M., Li T., Loder E. W., Mayo-Wilson E., McDonald S., & Moher D. (2020). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *International Journal of Surgery*. 88, Article 105906. https://doi.org/10.1016/j.ijsu.2021.105906

Pahn, J. (2020, March 21). *A Short Introduction To Decision Trees.* Level Up Coding. https://levelup.gitconnected.com/a-short-introduction-to-decision-trees-9481c36d2c52

Pandey, S., Ritz, C., & Perez-Cueto, F. J. A. (2021). An Application of the Theory of Planned Behaviour to Predict Intention to Consume Plant-Based Yogurt Alternatives. *Foods*. 10(1), 148. https://doi.org/10.3390/foods10010148

Pfeiler, T. M., & Egloff, B. (2018). Examining the Veggie personality: Results from a representative German sample. *Appetite.* 120, 246-255. https://doi.org/10.1016/j.appet.2017.09.005

Plant-Based Foods Association (2021, September 20). *PBFA urges UN to prioritize plant-based diets*. https://www.plantbasedfoods.org/pbfa-urges-un-to-prioritize-plant-based-diets/

Público. (2022, July 22). Morreu Encarnação Sousa, a mulher que tinha o título de mais velha em Portugal: tinha 113 anos e 122 dias. https://www.publico.pt/2022/07/22/impar/noticia/morreu-encarnacao-sousa-mulher-titulo-velha-portugal-113-anos-122-dias-2014668

Ramzai, J. (2020, June 19). *Simple guide for Top 2 types of Decision Trees: CHAID & CART*. Towards Science Data. https://towardsdatascience.com/clearly-explained-top-2-types-of-decision-trees-chaid-cart-8695e441e73e

Slade, P. (2018). If you build it, will they eat it? Consumer preferences for plant-based and cultured meat burgers. *Appetite*. 125, 428-437. https://doi.org/10.1016/j.appet.2018.02.030

Stewart, M. (2020, July 20). *Guide to Classification on Imbalanced Datasets*. Towards Data Science. https://towardsdatascience.com/guide-to-classification-on-imbalanced-datasets-d6653aa5fa23

Sucapane, D., Roux, C., & Sobol, K. (2022). Exploring how product descriptors and packaging colors impact consumers' perceptions of plant-based meat alternative products. *Appetite*. 167, Article 105590. https://doi.org/10.1016/j.appet.2021.105590

Supandi, A., Saefuddin, A., Sulvianti, I. D. (2022). Two step Cluster Application to Classify Villages in Kabupaten Madiun Based on Village Potential Data. *Journal of Statistics*. 11(2), 12-26. https://doi.org/10.29244/xplore.v10i1.272

Szejda, K., Stumpe, M., Raal, L., & Tapscott, C. E. (2021). South African Consumer Adoption of Plant-Based and Cultivated Meat: A Segmentation Study. *Frontiers in Sustainable Food Systems*. 5, Article 744199. https://doi.org/10.3389/fsufs.2021.744199

Thomas, O. Z., & Bryant, C. (2021). Don't Have a Cow, Man: Consumer Acceptance of Animal-Free Dairy Products in Five Countries. *Frontiers in Sustainable Food Systems*. 5, Article 678491. https://doi.org/10.3389/fsufs.2021.678491

Trpkova, M. & Tevdoski, D. (2009). Twostep cluster analysis: Segmentation of the largest companies in Macedonia. *Proceedings of the Challenges for Analysis of the Economy, the Businesses, and Social Progress International Scientific Conference*, Hungary, 302-318 http://hdl.handle.net/20.500.12188/2921

User11852. (2017, January 29). Assigning new points to a clustering algorithm is always a bit perplexing because the results of a clustering algorithm are [Comment on the post "How to assign new data to an existing clustering"]. *Stack Exchange*. https://stats.stackexchange.com/questions/258711/how-to-assign-new-data-to-an-existing-clustering

Vegan Food and Living. (2021, April 6). *Research suggests plant-based diets could replace the traditional diet within 100 years*. https://www.veganfoodandliving.com/news/plant-based-diets-replace-traditional-diet/

Vegconomist (2022, May 2). *Germany's Rügenwalder Mühle Sold More Meat-Free Than Meat Products in 2021*. https://vegconomist.com/company-news/facts-figures/germanys-rugenwalder-muhle-sold-more-meat-free-than-meat-products-in-2021/

Verain, M. C. D., Dagevos, H., & Jaspers, P. (2022). Flexitarianism in the Netherlands in the 2010 decade: Shifts, consumer segments and motives. *Food Quality and Preference*. 96, Article 104445. https://doi.org/10.1016/j.foodqual.2021.104445

Waisakurnia, W. (2020, June 13). *The Easiest Way to Interpret Clustering Result*. Towards Data Science. https://towardsdatascience.com/the-easiest-way-to-interpret-clustering-result-8137e488a127

Wartenberg, L. (2018, July 28). *Do Vegetarians Eat Cheese?*. Healthline. https://www.healthline.com/nutrition/do-vegetarians-eat-cheese

World Health Organization. (2021). *Plant-based diets and their impact on health, sustainability and the environment: a review of the evidence: WHO European Office for the Prevention*

and Control of Noncommunicable Diseases. Regional Office for Europe, World Health Organization. https://apps.who.int/iris/handle/10665/349086.

Yadav, A. (2019, January 11). *Decision Trees*. Towards Data Science. https://towardsdatascience.com/decision-trees-d07e0f420175

Yang, T. Y., & Dharmasena, S. (2021). U.S. Consumer Demand for Plant-Based Milk Alternative Beverages: Hedonic Metric Augmented Barten's Synthetic Model. *Foods.* 10(2), Article 265. https://doi.org/10.3390/foods10020265

6. Appendix

Appendix A- Protocol of the Systematic Literature Review

Query

(("plant-based" or vegetarian* or vegan* or veggie* or "veg" or flexitarian* or "meat-free" or "meat substitute" or "ovo-lacto*" or "lacto-ovo*" or "*ovo-vegetarian*" or "lacto-vegetarian*" or "*pesco-vegetarian*" or "semi-vegetarian*") and (consumer* or customer* or customer* or buyer* or shopper*)

and (behav* or pattern*or consum* or preference* or buy* or shop* or purchas*)

and ("data mining" or analytic* or "machine learning" or "deep learning" or *Cluster** or segment* or classificat* or predict* or regression* or association*"market basket analys" or model* or "customer analytic*" or "shopping mission*"))

or

(("plant-based" or vegetarian* or vegan* or veggie* or "veg" or flexitarian* or "meat-free" or "meat substitute" or "ovo-lacto*" or "lacto-ovo*" or "*ovo-vegetarian*" or "lacto-vegetarian*" or "*pesco-vegetarian*" or "semi-vegetarian*")

and (transact* or sale*)

and (food or product* or article*))

Selection Criteria

Inclusion criteria:

- Type of document: Articles
- Period of search: 2018-2022
- Research area:
  - Behavioral Science
  - Business Economics
  - Computer Science
  - Food Science Technology
  - Mathematics
  - Mathematical Methods in Social Sciences
  - Nutrition Dietetics
  - Science Technology
  - Social Sciences
- Web of Science Categories:
  - Behavioral Sciences
  - Business
  - Computer Science Theory Methods
  - Economics
  - Food Science Technology
  - Management
  - Mathematics Interdisciplinary Applications
  - Multidisciplinary Sciences
  - Nutrition Dietetics
  - Social Sciences Interdisciplinary
  - Social Sciences Mathematical Methods

Exclusion criteria:

- Articles not related to the subject of the project or with the selection criteria
- Articles that consist of a systematic review
- Articles only related to cultured-meat or insect-meat and not also plant-based meat
- Articles not related to the purchase or consumption behavior related to plant-based products or meat reduction
- Articles whose study is not made at the level of individuals, items, or transactions

Appendix B- Articles included in Systematic Review Literature

ID	Year	Title	Authors	Journal	Quartile
1	2022	Identifying behavioral and attitudinal barriers and drivers to promote consumption of pulses: A quantitative survey across five European countries	Henn et al.	Food Quality and Preference	Q1
2	2022	Flexitarianism in the Netherlands in the 2010 decade: Shifts, consumer segments and motives	Verain et al.	Food Quality and Preference	Q1
3	2022	Exploring how product descriptors and packaging colors impact consumers' perceptions of plant-based meat alternative products	Sucapane et al	Appetite	Q1
4	2021	Towards more environmentally sustainable diets? Changes in the consumption of beef and plant- and insect-based protein products in consumer groups in Finland	Niva & Vainio	Meat Science	Q1
5	2021	Purchasing veg private labels? A comparison between occasional and regular buyers	Martinelli & De Canio	Journal of Retailing and Consumer Services	Q1
6	2021	South African Consumer Adoption of Plant-Based and Cultivated Meat: A Segmentation Study	Szejda et al.	Frontiers in Sustainable Food Systems	Q2
7	2021	Understanding meat consumption in later life: A segmentation of older consumers in the EU	Grasso et al.	Food Quality and Preference.	Q1
8	2021	Reducing meat consumption in meat-loving Denmark: Exploring willingness, behavior, barriers and drivers	Hielkema & Lund	Food Quality and Preference	Q1
9	2021	Above and beyond meat: the role of consumers' dietary behavior for the purchase of plant-based food substitutes	Kopplin & Rausch	Review of Managerial Science	Q1
10	2021	Don't Have a Cow, Man: Consumer Acceptance of Animal-Free Dairy Products in Five Countries	Thomas & Bryant	Frontiers in Sustainable Food Systems	Q2
11	2021	Alternative proteins, evolving attitudes: Comparing consumer attitudes to plant-based and cultured meat in Belgium in two consecutive years	Bryant & Sanctorum	Appetite	Q1
12	2021	Nudge the Lunch: A Field Experiment Testing Menu-Primacy Effects on Lunch Choices	Andersson & Nelander	Games	Q3
13	2021	Distinguishing meat reducers from unrestricted omnivores, vegetarians and vegans: A comprehensive comparison of Australian consumers	Malek & Umberger	Food Quality and Preference	Q1
14	2021	U.S. Consumer Demand for Plant-Based Milk Alternative Beverages: Hedonic Metric Augmented Barten's Synthetic Model	Yang & Dharmasena	Foods	Q1
15	2021	An Application of the Theory of Planned Behaviour to Predict Intention to Consume Plant-Based Yogurt Alternatives	Pandey et al.	Foods	Q2
16	2020	A cross-sectional survey of the readiness of consumers to adopt an environmentally sustainable diet	Culliford & Bradbury	Nutrition Journal	Q3
17	2020	Motivating Pulse-Centric Eating Patterns to Benefit Human and Environmental Well-Being	Didinger & Thompson	Nutrients	Q1
18	2020	Drivers and Inhibitors in the Acceptance of Meat Alternatives: The Case of Plant and Insect-Based Proteins	de Koning et al.	Foods	Q2

Appendix B- Articles included in Systematic Review Literature (continued)

ID	Year	Title	Authors	Authors	Quartile
19	2020	Drivers of plant-based convenience foods consumption: Results of a multicomponent extension of the theory of planned behaviour	Contini et al.	Food Quality and Preference	Q1
20	2020	Order of meals at the counter and distance between options affect student cafeteria vegetarian sales	Garnett et al.	Nature Food	Q1
21	2020	Assessing diet in a university student population: a longitudinal food card transaction data approach	Morris et al.	Journal of Nutrition	Q1
22	2020	Drivers and barriers toward reducing meat consumption	Cheah et al.	Appetite	Q1
23	2020	Investigating the role of health halos and reactance in ice cream choice	Bullock et al.	Food Quality and Preference	Q1
24	2019a	Consumers' willingness to purchase three alternatives to meat proteins in the United Kingdom, Spain, Brazil and the Dominican Republic	Gomez-Luciano et al.	Amfiteatru Economic	Q1
25	2019	Impact of increasing vegetarian availability on meal selection and sales in cafeterias	Garnett et al.	Proceedings of the National Academy of Sciences of the United States of America	Q1
26	2019	To meat or not to meat? Comparing empowered meat consumers' and anti-consumers' preferences for sustainability labels	Apostolidis & McLeay	Food Quality and Preference.	Q1
27	2019	Older Consumers' Readiness to Accept Alternative, More Sustainable Protein Sources in the European Union	Grasso et al.	Nutrients	Q1
28	2019	Consumer Understanding and Culinary Use of Legumes in Australia	Figueira et al.	Nutrients	Q1
29	2019	Committed vs. uncommitted meat eaters: Understanding willingness to change protein consumption	Malek et al.	Appetite	Q1
30	2019	Reducing meat consumption: Identifying group-specific inhibitors using latent profile analysis	Lacroix & Gifford	Appetite	Q1
31	2019	Meat consumers and non-meat consumers in Germany: a characterisation based on results of the German National Nutrition Survey II	Koch et al.	Journal of Nutritional Science	Q1
32	2019b	Towards Food Security of Alternative Dietary Proteins: A Comparison Between Spain and the Dominican Republic	Gomez-Luciano et al.	Food Quality and Preference	Q2
33	2019	A Survey of Consumer Perceptions of Plant-Based and Clean Meat in the USA, India, and China	Bryant et al.	Frontiers in Sustainable Food Systems	Q2
34	2018	If you build it, will they eat it? Consumer preferences for plant-based and cultured meat burgers	Slade	Appetite	Q1
35	2018	Examining the Veggie personality: Results from a representative German sample	Pfeiler & Egloff	Appetite	Q1

Appendix C- Methodology followed by articles of the Systematic Review Literature

ID	Data Source	Country	Type of Data	Sample	Type of Analysis (e.g. Analytical Technique)
1	Online survey	Germany (G), Spain (ES), Denmark (D), Poland (P), UK	Socio-demographic, consumption habits, barriers, motivations, etc.	4916 participants (980-G, 972-D, 979-ES, 985-P, 1000-UK)	Descriptive (Clustering techniques, EFA, etc.) Predictive (Classification technique)
2	Online surveys	Germany	Socio-demographic, consumption habits, motivations, etc.	1253 participants - Year 2011, 1979 participants - Year 2019	Descriptive (Clustering Techniques, EFA, etc.)
3	Online experimental study	Canada, USA	Socio-demographic, preferences, opinions, etc.	149 participants - Pilot Study, 148- Study 1, 274- Study 2	Descriptive/ Hypothesis Tests
4	Online survey	Finland	Socio-demographic, consumption habits, motivations, intentions	1000 participants	Descriptive (LCA) Predictive (Classification technique)
5	Online survey	Italy	Socio-demographic, shopping habits, intentions	269 participants	Descriptive (CFA, SEM, etc.)
6	Online survey	South Africa	Socio-demographic, consumption, purchasing intentions, etc.	1087 participants	Descriptive Predictive (Regression technique)
7	Online survey	Finland (F), Poland (P), Netherlands (NL), Spain (ES), UK	Socio-demographic, physical activity habits, attitudes, preferences, consumption habits	2478 participants (493-F, 498-P, 501-ES, 498-NL, 488-UK)	Descriptive (Clustering techniques)/ Hypothesis Tests Predictive (Classification technique)
8	Online survey	Denmark	Socio-demographic, consumption habits, intentions, motivations, barriers	1005 participants	Descriptive Predictive (Classification technique)
9	Online survey	N/D	Socio-demographic, consumption habits, allergies, motivations, purchasing intentions	1363 participants	Descriptive (SEM, Analysis of Necessary Conditions)
10	Online survey	Brazil (BR) Germany (G), India (I), UK, USA	Demographic, consumption habits, purchasing intentions, opinions	5054 participants (1020-BR, 1051-G, 825-I, 1249-UK, 1009-USA)	Descriptive Predictive (Regression technique)
11	Online survey	Belgium	Socio-demographic, consumption habits, purchase intentions, motivations, opinions	2001 participants (1001-Year 2019, 1000-Year 2020)	Descriptive/Hypothesis tests Predictive (Regression technique)
12	Experimental study	Sweden	Transactional	7968 transactions	Descriptive Predictive (Regression technique)
13	Online survey	Australia	Socio-demographic, consumption and purchasing habits, perceptions, etc.	2797 participants	Descriptive/Hypothesis tests Predictive (Classification technique)
14	Observational study	USA	Transactional	7 items (1008 observations)	Predictive (Regression technique)
15	Online survey	Denmark	Socio-demographic, consumption and shopping habits, perceptions, etc.	265 participants	Descriptive (PCA, SEM) Predictive (Classification technique)
16	Online survey	UK	Socio-demographic, consumption habits, perceptions, intentions, motivations	442 participants	Descriptive/Hypothesis tests Predictive (Regression technique)
17	Not defined	Not defined	Not defined	Not defined	Not defined
18	Online and printed survey	Spain (ES), France (FR), UK, China (CN), USA, Brazil (BR), New Zealand (NZ), Netherlands (NL), Dominican Republic (DR)	Demographics, perceptions intentions of purchase, and consumption	3091 participants (210-ES, 484-FR, 366-UK, 571-CN, 539-USA, 216-BR, 268-NZ, 231-NL, 206-DR)	Descriptive (CFA, SEM)

Appendix C- Methodology followed by articles of the Systematic Review Literature (continued)

ID	Data Source	Country	Type of Data	Sample	Type of Analysis (e.g. Analytical Technique)
19	Online survey	Italy	Socio-demographic, consumption habits, intentions, perceptions, preferences	600 participants	Descriptive (CFA, SEM)
20	Experimental study	UK	Transactional	105143 transactions	Predictive (Regression technique)
21	Observational study	United Kingdom	Demographic, transactional	795 individuals 651 items 107723 transactions	Descriptive (Clustering techniques)
22	Online survey	Australia	Socio-demographic, consumption habits, intentions, perceptions, barriers, motivations	298 participants	Descriptive (EFA, CFA, SEM)
23	Experimental Study- Face-to-Face survey	USA	Demographics, intentions, preferences, knowledge, choices	223 participants	Predictive (Classification and Regression techniques)
24	Online and printed survey	UK, Spain (ES), Brazil (BR), Dominican Republic (DR)	Demographics, perceptions, preferences	729 participants (180-UK, 200-ES, 216-BR, 133-DR)	Descriptive (PCA)/ Hypothesis tests Predictive (Classification technique)
25	Observational and experimental study	UK	Transactional	940 participants 94 644 sales	Descriptive/Confidence Intervals Predictive (Regression technique)
26	Experimental study- face-to-face interview	UK	Socio-demographic, consumption habits, purchasing habits, intentions, motivations, etc.	600 participants	Descriptive (LCA) Predictive (Classification technique)
27	Online survey	UK, Netherlands, Spain, Poland, Finland	Socio-demographic, consumption habits, motivations, etc.	2478 participants	Descriptive (EFA) Predictive (Classification technique)
28	Online survey	Australia	Socio-demographic, consumption habits, purchasing habits, preferences, knowledge	505 participants	Descriptive
29	Online survey	Australia	Socio-demographic, consumption habits, motivations, perceptions, etc.	287 participants	Descriptive (EFA)/ Hypothesis tests
30	Online survey	Canada	Demographics, consumption habits, intentions, preferences	355 participants	Descriptive (LPA)/ Hypothesis tests
31	Personal/telephone interview and 24-h recall	Germany	Socio-demographic, consumption habits, health data	12915 participants	Descriptive/ Confidence Intervals Predictive (Classification and Regression techniques)
32	Online and face-to-face survey	Spain (ES), Dominican Republic (DR)	Socio-demographic, consumption habits, beliefs, intentions, perceptions, opinions	401 participants (200-ES, 201-DR)	Descriptive (PCA)/ Hypothesis Tests Prescriptive (Multi-criteria Decision Method)
33	Online survey	China (C), India (I), USA	Socio-demographic, consumption habits, familiarity, intentions, preferences	3030 participants (1019-C, 1024-I, 987- USA)	Descriptive/ Hypothesis Tests Predictive (Regression technique)
34	Experimental study	Not defined	Socio-demographic, consumption habits, preferences, motivations	533 participants	Descriptive Predictive (Classification technique)
35	Online survey	Germany	Socio-demographic, personality, attitudes	4496 participants - Study 1 5125 participants - Study 2	Descriptive/ Hypothesis Tests Predictive (Classification techniques)

Appendix D- Tables from Pingo Doce's database

Table D1: *Customers* Table

Variable	Type	Description
ID_Customer	Numeric	Customer ID
Age_registration	Numeric	Age at the time of registration
Gender	Categorical	Gender (M/F)
Country	Categorical	Country of origin
Province	String	District of customer
City	String	Municipality of customer
Cp4	Categorical	First 4 numbers of the postal code
Cp3	Categorical	Last 3 numbers of the postal code
RegistrationDate	Date	Registration date

Table D2: *Customers_Segm* Table

Variable	Type	Description
ID_Customer	Categorical	Customer ID
Segment_Name	Categorical	Segment identified by JM

Table D3: *Stores* Table

Variable	Type	Description
ID_store	Categorical	Store ID
Desc_store	Categorical	Store description
Cod_organization	Categorical	Organization code
Desc_organization	Categorical	Organization description
Cod_postal4	Categorical	First 4 numbers of the postal code
Cod_postal3	Categorical	Last 3 numbers of the postal code
Cod_parish	Categorical	Civil parish code
Desc_parish	Categorical	Civil parish description
Cod_municipality	Categorical	Municipality code
Desc_ municipality	Categorical	Municipality description
Cod_district	Categorical	District code
Desc_district	Categorical	District description
Format	Categorical	Store type

Table D 4: *Items* Table

Variable	Type	Description
ID_item	Categorical	Item code (level 1)
Desc_item	Categorical	Item description (level 1)
Cod_hier_level2	Categorical	Hierarchy level 2 code
Desc_ hier_level2	Categorical	Hierarchy level 2 description
…	…	…
Cod_ hier_level5	Categorical	Hierarchy level 5 code
Desc_ hier_level2	Categorical	Hierarchy level 5 description
Cod_brand	Categorical	Brand code
Desc_brand	Categorical	Brand description
Type_brand	Categorical	Private label identification

Table D5: *Sales* Table

Variable	Type	Description
ID_transacction	Categorical	Transaction ID
ID_date	Date	Date of transaction
ID_hour	Categorical	Hour of transaction
ID_store	Categorical	Store ID
ID_item	Categorical	Item ID
ID_customer	Categorical	Customer ID
Value_pvp	Categorical	Sale value of the item

Appendix E- Item Groups

1st Item Groups	2nd Item Groups	Definition and Content
1. Protein Alternatives to Meat & Fish	**1.1. Soja, Tofu & Seitan (vegan)**	Soja, Tofu and Seitan **(*vegan*)**
	1.2. Processed Meat Alternatives (vegan/vegetarian)	Vegan/vegetarian burgers, beefs, nuggets, schnitzel, etc. ***(vegan/vegetarian)***
	1.3. Alternatives to highly processed meat (vegan/vegetarian)	Vegan/vegetarian sausages, *alheira*, chorizo, ham, bacon, etc. ***(vegan/vegetarian)***
2. Fish	**2.1 Fish (nveg)**	Fresh fish and shellfish from fishery ***(non-veg)***
		Processed Fish- pre-processed frozen fish, tinned fish ***(non-veg)***
		Other processed fish- Smoked fish ***(non-veg)***
3. Meat	**3.1 Meat (nveg)**	Fresh Meat- beef, pork, poultry ***(non-veg)***
		Processed Meat- nuggets, frozen burgers, tinned meat ***(non-veg)***
		Highly Processed Meat- sausages, chorizo, *alheira*, *farinheira*, etc. ***(non-veg)***
4. Ready Food	**4.1 Ready Food (vegan/vegetarian/non-veg)**	Frozen/fresh ready meals (bowls, lasagne, pizza, noodles, pastas, etc.), sandwich, wraps, ready prepared *vegetarian*, french fries, ready rices, salads, fresh/instant soup etc. ***(vegan/vegetarian/non-veg)***
5. Snacks	**5.1 Snacks (vegan/vegetarian/non-veg)**	Potato chips, other chips (lentil chips, tortilla chips, green pea chips, etc.), popcorn, cheese balls, mixed nuts, fried chickpea, fried corn, nuts with toppings (chocolate, honey etc.) ***(vegan/vegetarian/non-veg)***
6. Fruit	**6.1 Fruit (vegan/vegetarian)**	Fresh fruit- apple, pear, orange, pineapple, coconut, etc. ***(vegan)***
		Frozen fruit- wild berries, açaí, passion fruit pulp, etc. ***(vegan)***
		Canned fruit- apricot, litchi, mango, etc. ***(vegan)***
		Dried fruit- date, plum, raisin, etc. ***(vegan/vegetarian)***
		Candied fruit- mixed fruit ***(vegan/vegetarian)***
		Fruit pouches ***(vegan/vegetarian)***
7. Vegetables	**7.1 Vegetables (vegan/vegetarian)**	Fresh vegetables- Mushroom, tomato, onion, garlic, green salads ***(vegan)***
		Frozen vegetables- Mushrooms, green beans, Mix vegetables ***(vegan/vegetarian)***
		Canned vegetables- Pickles, olives, mushrooms, corn, etc. ***(vegan/vegetarian)***
		Potatos- Fresh potatos, frozen potatos ***(vegan)***
8. Pulses	**8.1 Pulses (vegan)**	Dried/canned green pea, chickpea, bean, lentils, lupine ***(vegan)***
9. Nuts	**9.1 Nuts (vegan)**	Almonds, walnuts, cashews, hazelnuts, peanuts, pine nuts, etc. ***(vegan)***

Appendix E- Item Groups (continued)

1st Item Groups	2nd Item Groups	Definition and Content
10. Dairy	**10.1 Dairy (vegan/vegetarian/non-veg)**	Milk- Plant-based milk (soy milk, oat milk, rice milk, almond milk and others), milk ***(vegan/vegetarian)***
		Yoghurts ***(vegan/vegetarian/non-veg)***
		Cheese ***(vegan/vegetarian/non-veg)***
		Creams ***(vegan/vegetarian)***
		Butter ***(vegan/vegetarian)***
		Margarine ***(vegan/vegetarian)***
		Ice-cream ***(vegan/vegetarian)***
		Drinks- Refrigerated cappuccino, latte, etc. ***(vegan/vegetarian)***
		Others dairy products- coconut milk, coconut cream, condensed milk ***(vegan/vegetarian)***
11.. Refined Carbohydrates	**11.1. Refined Carbohydrates (vegan/vegetarian/non-veg)**	Refined grains- White bread, white rice, pasta, cereals, lasagna pasta, puff pastry, sugary cereals, etc. ***(vegan/vegetarian/non-veg)***
		Other refined products- Starch flavor, ferment, potato fecula ***(vegan)***
12. High Fiber Carbohydrates	**12.1 High Fiber Carbohydrates (vegan/vegetarian/non-veg)**	Whole grains- Wholemeal bread, brown bread, wild rice, wholemeal rice, wholemeal pasta, porridge, wholegrain cereals, etc. ***(vegan/vegetarian)***
		Other high fiber products- brans, germen, chickpea flavor, etc. ***(vegan)***
13. Biscuits & Cakes	**13. 1 Biscuits & Cakes (vegan/vegetarian/non-veg)**	Biscuits- cookies, galletas, cereal bars, diet bars, diet balls ***(vegan/vegetarian/non-veg)***
		Cakes- packaged cakes (milk bread, waffles, sweet pies, etc.), pastry (milk bread, sponge cake, croissants, etc.) ***(vegan/vegetarian/non-veg)***
14. Sweets & Desserts	**14.1 Sweets & Desserts (vegan/vegetarian/non-veg)**	Sweets- Candies (gums, caramels, chupa chups, chewing gums, etc.), chocolates, etc. ***(vegan/vegetarian/non-veg)***
		Desserts- frozen/instant/refrigerated desserts (jelly, profiteroles, etc.), cake flavoring, dyes, cocoa, etc. ***(vegan/vegetarian/non-veg)***
15. Sauces & Creams	**15.1 Sauces & Creams (vegan/vegetarian/non-veg)**	Sauces- soy sauce, vinegar, ketchup, mayo, piri-piri, pesto, barbecue sauce, tomato sauce ***(vegan/vegetarian/non-veg)***
		Pates, creams to spread- peanut butter, chocolate cream, guacamole, hummus, pate, jam ***(vegan/vegetarian/non-veg)***
		Toppings- chocolate, strawberry, vanilla, caramel ***(vegetarian/non-veg)***
16. Fats & Oils	**16.1 Fats & Oils (vegan /non-veg)**	Olive oil, sunflower oil (***vegan/non-veg***)
17. Condiments	**17.1 Condiments (vegan/vegetarian/non-veg)**	Spices- paprika, pepper, ginger, garlic, etc. ***(vegan)***
		Condiments- broth, lemon juice seasoning, etc. ***(vegan/vegetarian/non-veg)***
		Salt ***(vegan)***

Appendix E- Item Groups (continued)

1st Item Groups	2nd Item Groups	Definition and Content
18. Sugar & sweetener	**18.1 Sugar & sweetener (vegan/vegetarian)**	Sugar- White sugar, brown sugar, stevia, etc. ***(vegan)***
		Honey and syrups- honey, agave, maple syrup, etc. ***(vegan/vegetarian)***
19. Seeds	**19.1 Seeds (vegan)**	Mix seeds, chia seeds, pumpkin seeds, sesame seeds, etc. ***(vegan)***
20. Eggs	**20.1 Eggs (vegetarian)**	Eggs, egg white ***(vegetarian)***
21. Kids	**21.1 Kids (vegan/vegetarian/non-veg)**	Baby food, baby snacks, baby meals, children's milk ***(vegan/vegetarian/non-veg)***
22. Supplements	**22.1 Supplements (vegan/non-class)**	Whey, supplements ***(non-class)***
23. Yeast	**23.1 Yeast (vegan)**	Yeast ***(vegan)***
24. Hot Drinks	**24.1 Hot Drinks (non-class)**	Tea and infusions ***(vegan/vegetarian)***
		Coffee- Coffee beans, Soluble coffee) ***(vegan)***
		Chocolate Drinks ***(vegetarian)***
25. Drinks	**25.1 Drinks (non-class)**	Soft Drinks- Ice tea, coca, fruit juice… ***(non-class)***
26.. Alcohol	**26.1. Alcohol (non-class)**	Beer ***(non-class)***
		Wine ***(non-class)***
		Other Spirits ***(non-class)***
27. Water	**27.1 Water (non-class)**	Water ***(non-class)***
28.. Take-Away	**28.1. Take-Away (vegan/vegetarian/non-veg)**	Take-Away restaurant- Food, drink, and accessories sold in take-away area ***(vegan/vegetarian/non-veg)***
		Take-Away restaurant-
29. Restaurant	**29.1 Restaurant (vegan/vegetarian/non-veg)**	Food, drink, and accessories sold in restaurant area ***(vegan/vegetarian/non-veg)***
30. Pets	**30.1 Pets (non-class)**	Pets food, accessories ***(non-class)***
31. Bazar	**31.1 Bazar (non-class)**	Magazines, journals, books, home appliances, baby toys, technology, home accessories, garden accessories etc. ***(non-class)***
32. Flowers	**32.1 Flowers (non-class)**	Natural flowers, natural plants ***(non-class)***
33. Home Hygiene	**33.1 Home Hygiene (non-class)**	Detergents, aromatic candles, bags, insecticides, mops, etc. ***(non-class)***
34. Personal Hygiene	**34.1 Personal Hygiene (non-class)**	Perfumes, shampoos, intimate products, etc. ***(non-class)***
35. Textile	**35.1 Textile (non-class)**	Clothes, sheets, towels, etc. ***(non-class)***
36. Parapharmacy	**36.1 Parapharmacy (non-class)**	Pharmaceutical products ***(non-class)***
37. Gas station fuel	**37.1 Gas station fuel (non-class)**	Fuel

Appendix F- Input variables used in clustering

Table F1: Input variables used by each dimension clustering

	Variable	Type	Description
1. Demographic Dimension	Age	Continuous	Current age
	Gender	Categorical	Gender
2. Lifestyle Dimension	ID_hour_wkday	Categorical	Preferred hour to shop on weekdays
	EqualHour_wkday_wkend	Categorical	Flag variable: (1) if the preferred hour to shop on weekdays is equal to the preferred hour on weekends; (0) if hours are different
	EqualParih _wkday_wkend	Categorical	Flag variable: (1) if the parish of the preferred store on weekdays is equal to the parish of the preferred store on weekends; (0) if parishes are different
	Format _L1	Categorical	Format of the preferred store
3. Frequency and Monetary Dimension	Avg_monthly_trx	Continuous	Average monthly number of transactions
	Avg_amount_trx	Continuous	Average amount spent per transaction

Table F2: Input variables used in clustering based on the purchasing behavior

	Variable	Type	Description
4. Basket	MA_Tofu_Soy_Seitan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on items from food group 1.1 (tofu, seitan, and soy) and the average group share
	MA_vegetarian_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegetarian items from food group 1.2 (vegetarian processed meat alternatives- e.g., vegetarian hamburger, nuggets, samosas, chili, pies, etc.) and the average group share
	MA_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 1.2 (vegan meat processed alternatives- e.g., vegan hamburger, falafel, nuggets, jaca, etc.) and the average group share
	PM_vegetarian_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegetarian items from food group 1.3 (vegetarian alternatives to highly processed meats- e.g., vegetarian sausages, *alheira*, chorizo, ham) and the average group share
	PM_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 1.3 (vegan alternatives to highly processed meats- e.g.: vegan sausages, *alheira*, *farinheira*, bacon, etc.) and the average group share
	Dairy_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 10.1 (alternatives to dairy items- e.g.: plant-based milk, vegan cheese, yogurts, butter, etc.) and the average group share
	Biscuits_Cakes_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 10.1 (vegan cookies, cereal bars, diet bars/balls) and the average group share
	Sweet_Dessert_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 11.1 (vegan chocolate, gummies, jelly, energetic balls, etc.) and the average group share
	Snacks_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure spent on vegan items from food group 4.2 (vegan crisps, popcorns, nuts to snack) and the average group share
	Sauces_Creams_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 12.1 (mustard, soy sauce, piri-piri, tomato sauce, peanut butter, fruit jam, etc.) and the average group share
	ReadyFoods_vegetarian_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegetarian items from food group 4.1 (soups, pizza, lasagne, pastas, salads, etc.) and the average group share
	ReadyFoods_vegan_Share_Index	Continuous	Ratio between the share of average monthly expenditure on vegan items from food group 4.1 (noodles, bowls, soups, baked beans in sauces, canned vegetable bolognese, etc.) and the average group share

Appendix G- Descriptive statistics by cluster from lifestyle and monetary dimension

	Products hunting	*Medium interessed*	*High interested*	*Busy*	*Daily*
Index of the average monthly amount spent compared to the average (mean)	0.36	1.35	1.96	1.34	3.31
Index of the number of distinct categories[2] purchased compared to the average (mean)	0.60	1.21	1.7	0.89	2.05
Monthly sales volume % (mean)	19.6%	30.5%	35.8%	1.2%	12.8%

Appendix H- Average monthly expenditure share on item groups used as input by each basket cluster

Variables	*Traditional Omnivores*	*Potential Veggies*	*Convenience and Vegan Sweets*	*Receptive Omnivores*	*Veggie Lovers*
MA_Tofu_Soy_Seitan_Share_Index	0.00%	0.40%	0.01%	0.01%	2.13%
MA_vegetarian_Share_Index	0.00%	0.47%	0.00%	0.02%	2.33%
MA_vegan_Share_Index	0.01%	1.20%	0.02%	0.07%	4.12%
PM_vegetarian_Share_Index	0.00%	0.01%	0.03%	0.00%	0.48%
PM_vegan_Share_Index	0.00%	0.24%	0.00%	0.00%	1.39%
Dairy_vegan_Share_Index	0.40%	7.41%	0.36%	0.90%	2.99%
Biscuits_Cakes_vegan_Share_Index	0.34%	0.60%	0.30%	3.68%	1.39%
Sweet_Dessert_vegan_Share_Index	0.28%	0.45%	5.88%	0.46%	0.57%
Snacks_vegan_Share_Index	0.60%	0.91%	0.52%	4.11%	1.73%
Sauces_Creams_vegan_Share_Index	0.47%	0.61%	5.23%	0.69%	1.27%
ReadyFoods_vegetarian_Share_Index	0.18%	0.54%	4.14%	0.39%	2.06%
ReadyFoods_vegan_Share_Index	0.01%	0.03%	0.00%	0.76%	2.39%
Total average share per month	**2.30%**	**12.87%**	**16.49%**	**11.10%**	**22.84%**

[2] Categories from original Pingo Doce's market structure

Appendix I- Descriptive statistics for each basket cluster considering transactions

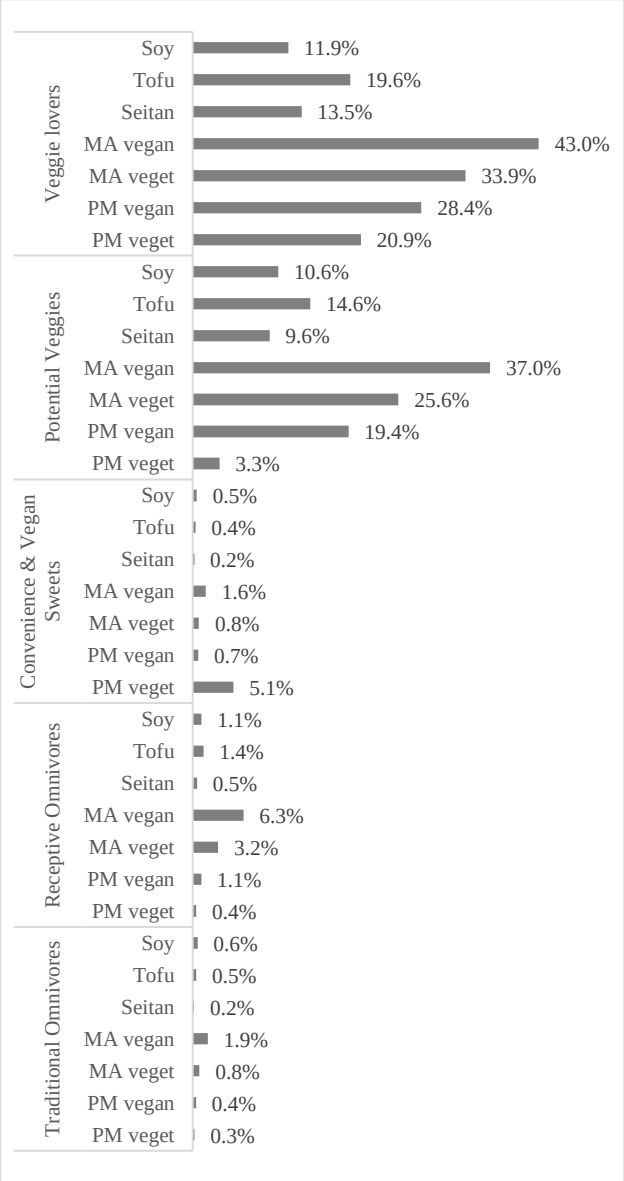

Figure I.1: Distribution of customers per cluster who purchased at least once, each alternative protein, considering the whole 6-month period

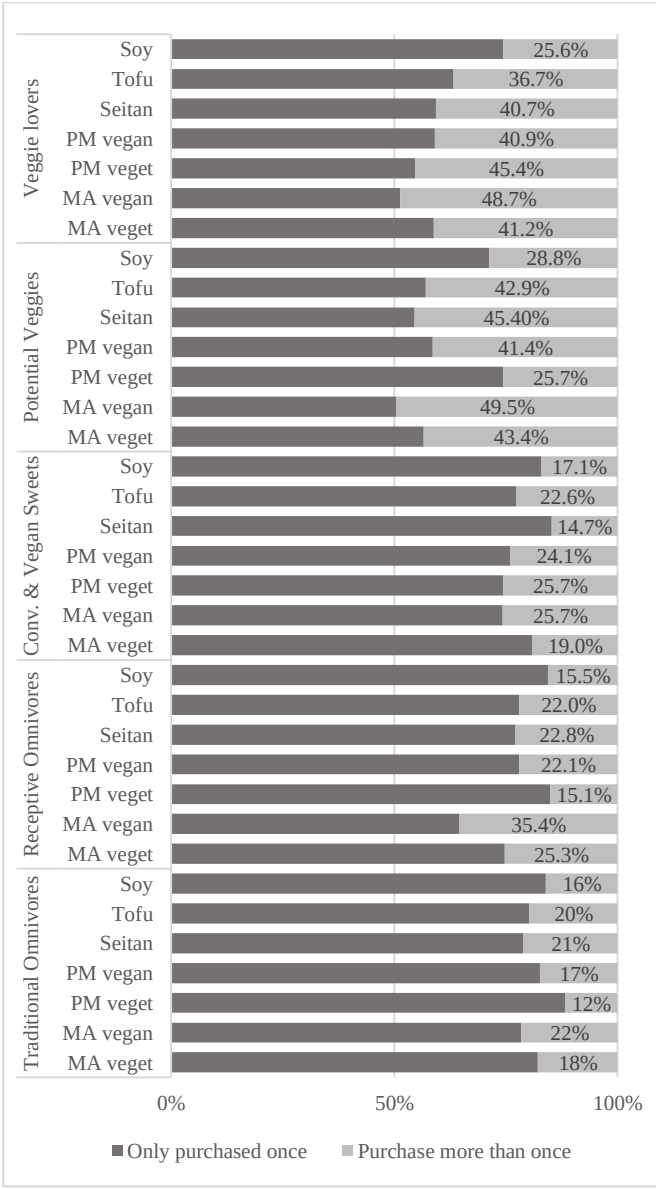

Figure I.2: Distribution of customers per cluster, by behavior and alternative protein, considering the whole 6-month period

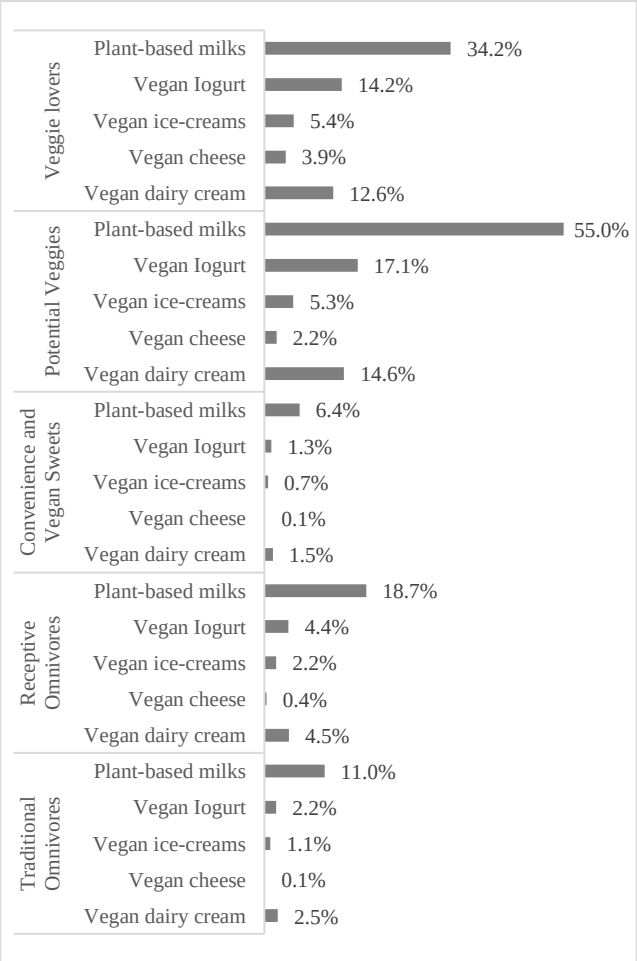

Figure I.3: Distribution of customers per cluster who purchased at least once, each dairy alternative, considering the whole 6-month period

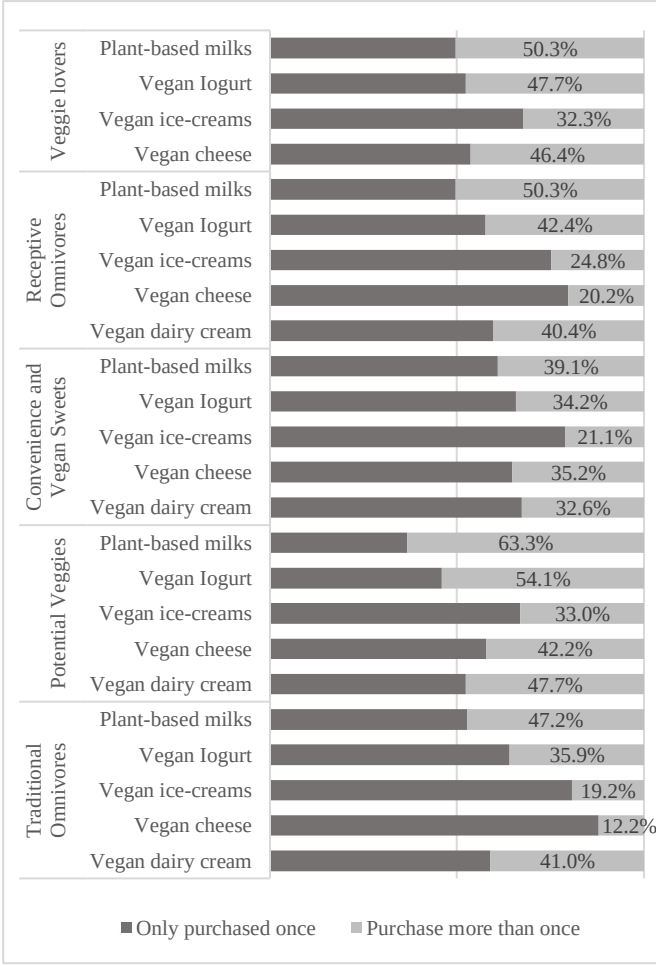

Figure I.4: Distribution of customers per cluster, by behavior and dairy alternative, considering the whole 6-month period

Appendix J- Selected classification model

Model Estimation

- **Algorithm:** CART
- **Number of customers considered:** 4 thousand
- **Partition:** 70% train set, 30% test set
- **Stop criterium rules:**
  - Maximum depth: 20
  - Minimum records in parent node: 200
  - Minimum records in child node: 100
- **Results**:
  - Tree depth: 13
  - Number of Rules (final nodes): 24

	Metrics	*Traditional Omnivores*	*Receptive Omnivores*	Convenience & Vegan Sweets	*Potential Veggies*	*Veggie Lovers*
Train set (70%)	**Sensitivity**	97.2%	91.5%	95.7%	91.3%	69.6%
	Specificity	99.5%	97.3%	97.7%	94.2%	97.4%
	Precision	98.0%	89.4%	91.4%	79.7%	87.3%
	Overall accuracy	88.9%				
Test set (30%)	**Sensitivity**	97.1%	90.8%	95.7%	93.6%	70.7%
	Specificity	99.4%	97.6%	97.7%	94.5%	97.9%
	Precision	97.6%	90.1%	91.5%	81.4%	89.2%
	Overall accuracy	89.7%				

Application of the model

	Metrics	*Traditional Omnivores*	*Receptive Omnivores*	Convenience & Vegan Sweets	*Potential Veggies*	*Veggie Lovers*
Original dataset	**Sensitivity**	97.3%	91.1%	96.0%	92.1%	69.9%
	Specificity	99.1%	99.6%	99.4%	98.1%	99.8%
	Precision	99.9%	87.6%	75.0%	48.5%	49.0%
	Overall accuracy	96.9%				

Rules for classifying *Traditional Omnivores* - contains 1 rule

Rule 1 for Traditional Omnivores

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index <= 29,985263
and PM_vegetarian_Share_Index <= 102,773144
and PM_vegan_Share_Index <= 72,147896
then **Traditional Omnivores**

Rules for classifying *Potential Veggies* - contains 5 rules

Rule 1 for Potential Veggies

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index <= 29,985263
and PM_vegetarian_Share_Index <= 102,773144
and PM_vegan_Share_Index > 72,147896
and PM_vegan_Share_Index <= 445,500647
then **Potential Veggies**

Rule 2 for Potential Veggies

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index > 29,985263
and MA_vegetarian_Share_Index <= 320,273631
then **Potential Veggies**

Rule 3 for Potential Veggies

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index > 11,404924
and MA_Tofu_Soja_Seitan_Share_Index <= 330,186468
then **Potential Veggies**

Rule 4 for Potential Veggies

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index > 7,724186
and MA_vegan_Share_Index <= 149,616422
and PM_vegetarian_Share_Index <= 239,804002
and MA_vegetarian_Share_Index <= 291,063848
then **Potential Veggies**

Rule 5 for Potential Veggies

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index > 13,972030
then **Potential Veggies**

Rules for classifying *Convenience and Vegan Sweets* - contains 4 rules

Rule 1 for Convenience and Vegan Sweets

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index <= 29,985263
and PM_vegetarian_Share_Index > 102,773144
and PM_vegetarian_Share_Index <= 449,380609
then **Convenience and Vegan Sweets**

Rule 2 for Convenience and Vegan Sweets

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index > 17,905692
then **Convenience and Vegan Sweets**

Rule 3 for Convenience and Vegan Sweets

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index > 14,934366
then **Convenience and Vegan Sweets**

Rule 4 for Convenience and Vegan Sweets

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index > 9,898330
then **Convenience and Vegan Sweets**

Rules for classifying *Receptive Omnivores* - contains 3 rules

Rule 1 for Receptive Omnivores

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index > 10,225290
then **Receptive Omnivores**

Rule 2 for Receptive Omnivores

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index > 12,480733
then **Receptive Omnivores**

Rule 3 for Receptive Omnivores

if ReadyFoods_vegan_Share_Index > 13,629744
and MA_vegan_Share_Index <= 23,595801
and ReadyFoods_vegan_Share_Index <= 220,801852
and MA_vegetarian_Share_Index <= 72,895209
and MA_Tofu_Soja_Seitan_Share_Index <= 23,088018
then **Receptive Omnivores**

Rules for classifying *Veggie Lovers* - contains 11 rules

Rule 1 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index <= 29,985263
and PM_vegetarian_Share_Index <= 102,773144
and PM_vegan_Share_Index > 72,147896
and PM_vegan_Share_Index > 445,500647
then **Veggie Lovers**

Rule 2 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index > 29,985263
and MA_vegetarian_Share_Index > 320,273631
then **Veggie Lovers**

Rule 3 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index <= 11,404924
and MA_vegetarian_Share_Index > 29,985263
and MA_vegetarian_Share_Index > 320,273631
then **Veggie Lovers**

Rule 4 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index <= 7,724186
and Biscuit_Cakes _vegan_Share_Index <= 12,480733
and Snacks_vegan_Share_Index <= 9,898330
and ReadyFoods_veget_Share_Index <= 14,934366
and Sweet_Dessert_vegan_Share_Index <= 17,905692
and Sauces_Creams_vegan_Share_Index <= 10,225290
and MA_Tofu_Soja_Seitan_Share_Index > 11,404924
and MA_Tofu_Soja_Seitan_Share_Index > 330,186468
then **Veggie Lovers**

Rule 5 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index > 7,724186
and MA_vegan_Share_Index <= 149,616422
and PM_vegetarian_Share_Index <= 239,804002
and MA_vegetarian_Share_Index > 291,063848
then **Veggie Lovers**

Rule 6 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index > 7,724186
and MA_vegan_Share_Index <= 149,616422
and PM_vegetarian_Share_Index > 239,804002
then **Veggie Lovers**

Rule 7 for Veggie Lovers

if ReadyFoods_vegan_Share_Index <= 13,629744
and Dairy_vegan_Share_Index <= 13,972030
and MA_vegan_Share_Index > 7,724186
and MA_vegan_Share_Index > 149,616422
then **Veggie Lovers**

Rule 8 for Veggie Lovers

if ReadyFoods_vegan_Share_Index > 13,629744
and MA_vegan_Share_Index <= 23,595801
and ReadyFoods_vegan_Share_Index <= 220,801852
and MA_vegetarian_Share_Index <= 72,895209
and MA_Tofu_Soja_Seitan_Share_Index > 23,088018
then **Veggie Lovers**

Rule 9 for Veggie Lovers

if ReadyFoods_vegan_Share_Index > 13,629744
and MA_vegan_Share_Index <= 23,595801
and ReadyFoods_vegan_Share_Index <= 220,801852
and MA_vegetarian_Share_Index > 72,895209
then **Veggie Lovers**

Rule 10 for Veggie Lovers

if ReadyFoods_vegan_Share_Index > 13,629744
and MA_vegan_Share_Index <= 23,595801
and ReadyFoods_vegan_Share_Index > 220,801852
then **Veggie Lovers**

Rule 11 for Veggie Lovers

if ReadyFoods_vegan_Share_Index > 13,629744
and MA_vegan_Share_Index > 23,595801
then **Veggie Lovers**

Appendix K- Deployment end-to-end

1. Join the table of Sales with the table of Customers
2. Aggregate data by client and group of items, at the level of detail that is pretend, and calculate the total amount spend during the 6 months period (Xi, where i=item group i)
3. Use a Pivot function which will generate a new dataset where each row corresponds to customer and each column to an item group
4. In the new dataset, change negative values and null/blank values for 0.
5. For each item group, create a new variable that correspond to the average monthly expenditure over the 6 months ($Mi = \frac{Xi}{6}$)
6. Calculate a new variable which corresponds to the total amount spend per month ($T = \sum Mi$)
7. Create new variables which are the share of average monthly expenditure on each item group ($Si = \frac{Mi}{T}$)
8. To have the previous variables indexed, divide the share of average monthly expenditure on each item by the average share from the same item (Ii = $\frac{Si}{\bar{S}}$)
9. From the new indexed variables (Ii), select the ones that are related to the item groups considered as input in this project
10. Apply the decision tree model
11. Identify the customer's classification
12. Follow a maintenance plan (updating needs) and a continuous improvement plan (a process of identifying segment migration)

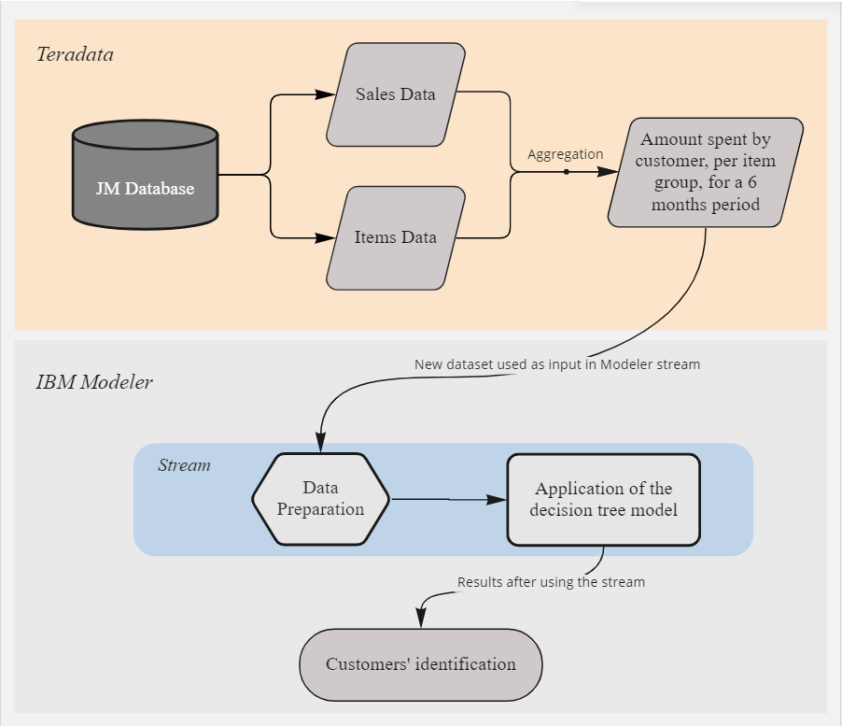

Figure K.1: Workflow from deployment

Appendix L- Outputs of Decisiom Tree Models to Recommendations

Veggie Lovers (higher frequency vs. lower frequency than the mean)

Model	Rule Precision	Rule Odds	Rule Lift	Rule Support	Rules
CHAID	86.35	6.33	0.467	7.77%	If Condiments_Share_Index > 0 and Condiments_Share_Index <= 0.868930966932591 and Takeaway_Share_Index > 0 then **Higher_Frequency_VeggieLovers**
CHAID	92.53	12.38	0.784	4.95%	If Condiments_Share_Index > 0 and Condiments_Share_Index <= 1.55166244095105 and Takeaway_Share_Index > 0 and Bazar_Share_Index > 0 and Nuts_Share_Index > 0 then **Higher_Frequency_VeggieLovers**

Receptive Omnivores (higher amount spend vs. lower amount spend than the mean)

Model	Rule Precision	Rule Odds	Rule Lift	Rule Support	Rules
CHAID	69.4%	2.27	0.013	2.03%	If Fish_Share_Index > 0.60847563930715 and Fish_Share_Index <= 1.6768456571604 and (Bazar_Share_Index > 0 and Bazar_Share_Index <= 0.882499403098957) and Pulses_Share_Index > 0 and Personal_Hygiene_Share_Index > 0.539267259446728 and Meat_Share_Index > 0.860099035782217 Then **Higher_AmountSpend_ReceptiveOmnivores**