UNIVERSIDADE CATÓLICA PORTUGUESA

Exploring Sustainability Patterns in Household Food Purchasing Habits

A Data Analytics Approach

Rui Filipe Peixoto Teixeira

Católica Porto Business School
April 2024

UNIVERSIDADE CATÓLICA PORTUGUESA

Exploring Sustainability Patterns in Household Food Purchasing Habits

A Data Analytics Approach

Masters Final Thesis
presented to Universidade Católica Portuguesa
to obtain a Master's Degree in Management with specialisation in Business Analytics

by

Rui Filipe Peixoto Teixeira

under the guidance of
Professor Vera Migueis

Católica Porto Business School, Universidade Católica Portuguesa
April 2024

Acknowledgements

As I close this significant chapter of my academic journey, I am filled with a deep sense of fulfilment and appreciation. The completion of this work brings a mix of relief and reflective nostalgia, symbolising the culmination of a six-year academic journey.

This journey would not have been possible without the invaluable support of important and exceptional people. Therefore, I would like to leave a special thanks:

To my parents for their unwavering love and support and whose belief in me has been a guiding light through this journey, providing strength and encouragement every step of the way. Your sacrifices have not gone unnoticed, and I am profoundly thankful for the foundation you've laid, allowing me to pursue my dreams.

To my younger brother, your support and our bond have been invaluable throughout this process. Your encouragement and the strength of our relationship have inspired me, offering solace and motivation during challenging moments. Your presence has been a constant reminder of the importance of perseverance and the power of family ties.

To my girlfriend, your motivation, and faith in my skills have been crucial throughout this adventure. Your constant support and trust in me have not only inspired me, but also filled my life with joy and meaning. I am eternally thankful for every moment of comfort, every word of support, and the never-ending inspiration you've offered. Thank you for being my partner and biggest supporter in this path that we call life.

To my dear friends, each moment of friendship we've shared has been invaluable, shaping this journey in ways words can hardly express. I cherish

every laugh and every challenge we've faced together. Here's to the memories we've made and the ones yet to come.

To my supervisor, Vera Miguéis for her constant support and expert guidance. Her insights and expertise have been invaluable to this work, providing both direction and inspiration.

To Joana Bôto, whose valuable suggestions greatly enhanced this thesis.

As I finish my thesis, I mark the conclusion of a major chapter in my life and look forward to the new experiences that lie ahead. To everyone who has joined me on this path, I extend my heartfelt gratitude for making these years truly precious.

Abstract

At a time when global environmental challenges are pressing and consumer options are more diversified than ever, there is an urgent need to integrate sustainable practices with the growing availability of food items that may not always be beneficial to healthy living. This work describes a critical step in using food retailers' data to turn complicated consumer interactions into meaningful insights. This initiative seeks to find opportunities to support more environmentally friendly and nutritionally healthy consumer choices by strategically analysing consumption habits.

This thesis used a complex approach to data cleansing, which included a variety of standardisation and normalisation procedures to match diverse data structures and formats. The outcome of these efforts was the merging of three data tables, which converged to form an integrated data table that not only contained a retailer's product offerings, but also included the complete nutritional information and calculated environmental implications of each product.

An important methodological highlight is the classification of clients into three unique segments based on macronutrient and micronutrient consumption, as well as environmental effect. This three-dimensional categorization is crucial, as it gives a full consumer profile and enables for a detailed knowledge of consumption behaviours across three important dimensions.

This thesis lays the framework for future research on the sustainability and nutritional effect of consumer behaviours. It uncovers critical information that might drive policy decisions targeted at directing household consumption towards more sustainable and healthier results by segmenting consumer data in great detail.

Keywords: Household Consumption Analysis, Nutritional Intake, Customer Segmentation, Environmental Sustainability Metrics, Clustering, Retail.

Resumo

Numa altura em que os desafios ambientais globais são urgentes e as opções dos consumidores são mais diversificadas do que nunca, existe a necessidade de integrar práticas sustentáveis com a crescente disponibilidade de produtos alimentares que podem nem sempre ser benéficos para uma vida saudável. Este trabalho descreve um passo fundamental na utilização dos dados dos retalhistas de produtos alimentares para transformar interações dos consumidores em informações úteis do ponto de vista analítico. Esta iniciativa procura encontrar oportunidades para apoiar escolhas de consumo com menos impacto ambiental e nutricionalmente saudáveis, analisando estrategicamente os hábitos de compra dos consumidores.

Este estudo utiliza uma abordagem de limpeza de dados composta por vários processos, que inclui uma variedade de procedimentos de normalização e padronização para fazer corresponder diversas estruturas e formatos de dados. Estes esforços culminam na fusão de três bases de dados, que convergem para formar uma base de dados integrada que não só contém toda a oferta de produtos de um retalhista, mas também inclui a informação nutricional completa e as implicações ambientais calculadas de cada produto.

Um aspeto importante da metodologia é a classificação dos clientes em três segmentos únicos com base no consumo de macronutrientes e micronutrientes, bem como no impacto ambiental. Esta categorização tridimensional é crucial, uma vez que fornece um perfil completo do consumidor e permite um conhecimento pormenorizado dos comportamentos de consumo em três dimensões importantes.

Esta tese estabelece as bases para a investigação futura sobre a sustentabilidade e o efeito nutricional do comportamento do consumidor,

utilizando dados de retalhistas, revelando informação que pode orientar decisões políticas destinadas a direcionar o consumo das famílias para que este seja mais sustentável e saudável.

Palavras-Chave: Análise do Consumo das Famílias, Consumo Nutricional, Segmentação de Clientes, Métricas de Sustentabilidade Ambiental, *Clustering*, Retalho.

Index

Figure Index

Table Index

Introduction

In the modern era, it is imperative to consider each human's sustainable practices, particularly those connected to our dietary habits. Our food choices significantly influence our carbon footprint, underscoring the importance of analysing household consumption patterns. Such analysis is not just about individual health but also about the collective environmental impact, highlighting the interplay between nutrition, sustainability, and well-being. This introduction lays the groundwork for a study dedicated to understanding the nuances of consumption behaviours and their broader sustainability implications.

The complex dynamics of consumer behaviour, particularly those related to food purchasing habits, have emerged as critical subjects of research in the expanding science of environmental sustainability. This research aims to explore the intricate link between household purchasing patterns and sustainability, with a focus on the retail sector.

The theoretical foundation of this work is based on analysing sustainability through the perspective of nutritional health and environmental effects of food consumption. Making use of quantitative methodologies, this thesis navigates through the intricate link between household consumption patterns and its health and environmental impacts. The quantitative study is driven by the R and SQL programming languages, which are well-known for their data manipulation and modelling capabilities and flexibility in assessing information from several data tables. The data used in this study is derived from a retailer store database, containing consumer transactions, customer information and product information, that is complemented by a secondary database that includes Environmental and Nutritional Indicators for each food subcategory.

The motivation behind this study is derived from the need to understand the sustainability of household consumption habits, particularly in terms of food purchases. Through detailed examination of transactional data, it aims to fill a substantial information gap about how consumers' food intake affects the environment and their health. By identifying different customer segments and understanding how their consumption interacts with sustainability factors, this study seeks to serve as a foundation for future research, eventually contributing to better informed policy and strategy formulation for encouraging sustainable consumption.

The study begins with a literature review that analyses existing research, followed by a methodology section that details data cleaning and modelling techniques. The following chapters dissect the findings, examining customer segmentation and delving into geographical and age distribution of customers, providing insights for targeted sustainability interventions.

1. Literature Review

1.1 Food Retail Impact on Sustainability

The growing necessity of solving sustainability challenges has sparked a surge in research in the retail and home consumption realms. This chapter aims to provide in-depth understanding of such challenges by examining the numerous approaches used, the various kinds of data evaluated, and the conclusions reached by different researchers, offering a comprehensive picture of the current research environment. Furthermore, it aims to reveal gaps in the current body of knowledge, emphasising areas that need further investigation.

Ruiz et al. (2019) provides an exhaustive exploration of the field, emphasising pivotal themes such as supply chain sustainability, food waste, efficient supermarket management, and carbon footprint, among others. This study establishes a foundation for future research by suggesting prospective pathways for additional investigation.

The effect of carbon footprint labelling in influencing consumer behaviour is a critical area of interest in this research field. According to research, labelling products with easily comprehensible symbols might help consumers better understand their carbon footprint. Rondoni and Grasso (2021) argue that the success of such carbon footprint labels is dependent on their clarity, comprehensibility, and perceived reliability, emphasising the importance of straightforward and trustworthy labelling in supporting informed consumer decisions. This emphasises the importance of straightforward and clear labelling to foster customer trust and informed decisions.

The relationship between socioeconomic status and food consumption sustainability is another significant area of exploration. Higher socioeconomic groups eat more energy-dense and animal-based goods in nations undergoing

fast nutrition transition, compromising the sustainability of their diets compared to lower socioeconomic groups (Eini-Zinab et al., 2021), underlining the possibility of tailored interventions to different socioeconomic groups, looking to promote healthier diets.

The potential role of consumption taxes in promoting more climate-friendly diets is another key area of investigation. High-income households contribute considerably to greenhouse gas emissions through their food choices, implying that changing these consumption patterns might result in considerable decreases in emissions (Edjabou & Smed, 2013). This point of view emphasises the environmental and public health benefits of dietary changes, as well as the possible role of consumption taxes in encouraging such changes.

Despite these findings, the research on the impact of retail data on home sustainability is scarce. While previous research has provided useful insights into consumer perceptions of carbon labelling and corporate social responsibility in the retail sector (Upham et al., 2011; Wagner et al., 2008), constant streams of data are required to understand the dynamics of household consumption patterns and their sustainability implications over time. This research aims to fill that void by giving a more thorough knowledge of the delicate interplay between family purchasing habits and sustainability.

1.2 Analytics in Retail

Navigating the complex world of retail involves the effective use of analytics, which are divided into descriptive, predictive, and prescriptive domains, each of which offers distinct insights and actionable paths. Descriptive analytics highlights past and current data by providing a retrospective study of sales, customer demographics, and inventories, resulting in a solid grasp of retail operations and consumer habits.

In contrast, predictive analytics takes a step further by using historical data to predict future events such as sales patterns or customer purchasing behaviours. This involves discovering patterns and correlations in data using statistical algorithms and machine learning approaches. In the research of Big Mart's sales data, Suryadevara (2020) conducted predictive analysis utilising machine learning algorithms to estimate sales patterns, emphasising the necessity of data leveraging for strategic inventory management and sales optimisation. This study analysed years of sales data using a variety of machine learning approaches such as regression models, decision trees, random forests, and gradient boosting. The goal was to discover significant factors impacting sales variations, such as product features, shop locations, and promotional activities, and the results showed that random forests provided the best correct forecasts of the models examined. The relevance of this study goes beyond Big Mart, as it lays the groundwork for other retail companies looking to use machine learning for better sales forecasting and inventory management. The research paper provides practical insights for retailers on how to execute data-driven strategies in a quickly changing retail sector by outlining the benefits and shortcomings of various prediction algorithms. This demonstrates predictive analytics' revolutionary impact in retail, enabling more accurate sales forecasting, effective inventory management, and the development of informed, data-driven plans to assure long-term profitability and consumer pleasure.

Prescriptive analytics goes beyond prediction by providing actionable suggestions that guide towards desired results, enhancing retail decision-making processes through data-driven strategies and actionable insights. For instance, a study by Van Dooren et al. (2014) used linear programming models to establish a balance between nutritional needs, environmental consequences, and cost-effectiveness. Existing data from 2007 to 2010 on the composition of the average Dutch diet was used in the study, which included various goods and their

respective weight (g/day), Kcal, greenhouse gas emissions (kg CO2 equivalent), and cost (€). The Linear programming model took into account a set of constraints such as limited greenhouse gas emissions, limited price, and a specific amount of nutrients. The researchers were able to analyse the effects of each constraint on price, greenhouse gas emissions, land usage, and energy consumption, resulting in the discovery of a diet with half the average cost of the and lower sustainable impact when compared to the Dutch diet, all while providing the required nutrients. These results highlight the important role of prescriptive analytics in identifying diets that are not only nutritional and cost-effective, but also environmentally sustainable, laying the foundations for shifts in retail policies and decisions that aim to influence customers towards a more sustainable consumption.

1.3 Exploring Sustainability Through Data Analytics

Data analytics is an essential component of every research project, translating raw data into relevant insights that serve as the foundation for findings and ideas. The choice of data analysis approach in the field of sustainability research, particularly in the retail sector, can have a considerable impact on the findings. The most appropriate analytical technique is determined by the type of the data, the research questions, and the overall aims of the study. Each technique provides distinct insights, whether it is a regression model to discover variable correlations, a linear programming which maximises or minimises results within restrictions, or graphical analysis to infer data patterns.

This chapter will delve into several studies, examining how multiple researchers have used data analysis to find trends, make conclusions, and provide solutions to critical sustainability concerns in the retail industry. The goal is to develop a better grasp of the power of data analysis in creating long-term transformation.

For instance, the study "Identifying Financially Sustainable Pricing Interventions to Promote Healthier Beverage Purchases in Small Neighbourhood Stores" analysed real sales data from two small San Francisco businesses over a two-year period. The use of real-world sales data lends a high degree of authenticity to the study, as it reflects actual consumer behaviour in a specific context. However, it restricts the findings' generalizability because they may not apply to diverse geographical regions or business kinds. Similarly, the research "Patterns of Legume Purchases and Consumption in the United States" relied on data from the National Health and Nutrition Examination Survey and retailer data from various sources, being limited by the fact that purchases from the food service sector, including restaurants, were not covered by the analysis. While this kind of data provides a general picture of consumer behaviour in the United States, it does not fully reflect the complexities of unique local contexts, highlighting the need for tailored strategies to approach local consumption issues.

The important role of analytics emerges as a thread that weaves through every decision-making process, generating strategies that are not only economically viable but also ecologically sustainable in the retail management realm. These studies highlight the significant influence of analytics in the retail sector, particularly by bridging the gap between operational efficiency and sustainability measures. The analysis of various analytical models, ranging from descriptive to prescriptive, can reveal pathways by which retailers can navigate the complex dynamics of market trends, consumer behaviour, and sustainability challenges, thereby crafting strategies that balance profitability and environmental responsibility.

However, a significant gap in existing research has been identified, particularly at the connection of retail analytics and sustainability trends, exposing a promising area for additional discovery and analysis. The use of

transactional retail data, which is not only more cost-effective but also more logistically practical than traditional survey techniques, provides a valuable source of information that can be analysed to discover patterns and trends in consumer behaviour and market dynamics, especially when combined with the possibility to evaluate this data more often and on a more regular basis, whether weekly or monthly, gives managers the ability to react to changes in the markets with agility and accuracy. Thus, this research not only adds to the current body of knowledge, but it also establishes a foundation for future research that aims to dive further into the connection of retail analytics and sustainability.

2. Methodology

Drawing on the insights into the CRISP-DM methodology outlined in Schröer et al.'s (2021), this thesis' methodology chapter will be structured to reflect the CRISP-DM's phases: business understanding, data understanding, data preparation, modelling, evaluation, and deployment. This structure not only follows the industry-standard process model for data mining projects, but it also provides a comprehensive framework for navigating the complicated process of data analysis, from the initial understanding of the business challenge to the implementation of data-driven solutions. By employing the CRISP-DM technique, the thesis not only adheres to best practices in data mining, but it also assures an organised and rigorous process that improves the clarity and trustworthiness of the results. This systematic technique allows for a consistent narrative throughout this thesis.

2.1 Data Sources

The data for this thesis comes from two different sources, totalling five tables. The first three tables are sourced from a comprehensive database provided by a well-known Portuguese retail shop. These tables are highly integrated, providing extensive information on the retailer's product offers, customer transactions, and consumer profiles, containing a wide range of factors, including product SKUs, brand descriptors, price, transaction information, and customer demographics, enabling for a multidimensional investigation of retail operations and consumer behaviour.

Bôto et al. (2014) conducted significant research to create the other two tables that are going to be used. The Nutritional data table, which delves into the details of food products, cataloguing critical nutritional information such as caloric

content and nutrient breakdowns, and the Environmental Impact dataset that quantifies the environmental impact of food products, including CO2 emissions, water consumption, and land usage per kilogramme of food. These data tables will be linked with the retailer's data tables to provide nutritional and environmental information to each product, allowing for a sustainability analysis based on household purchasing habits. The tables are represented in Figure 1, from which variable description is present in Table 9, Table 10 and Table 11 from the appendix, and Figure 2.

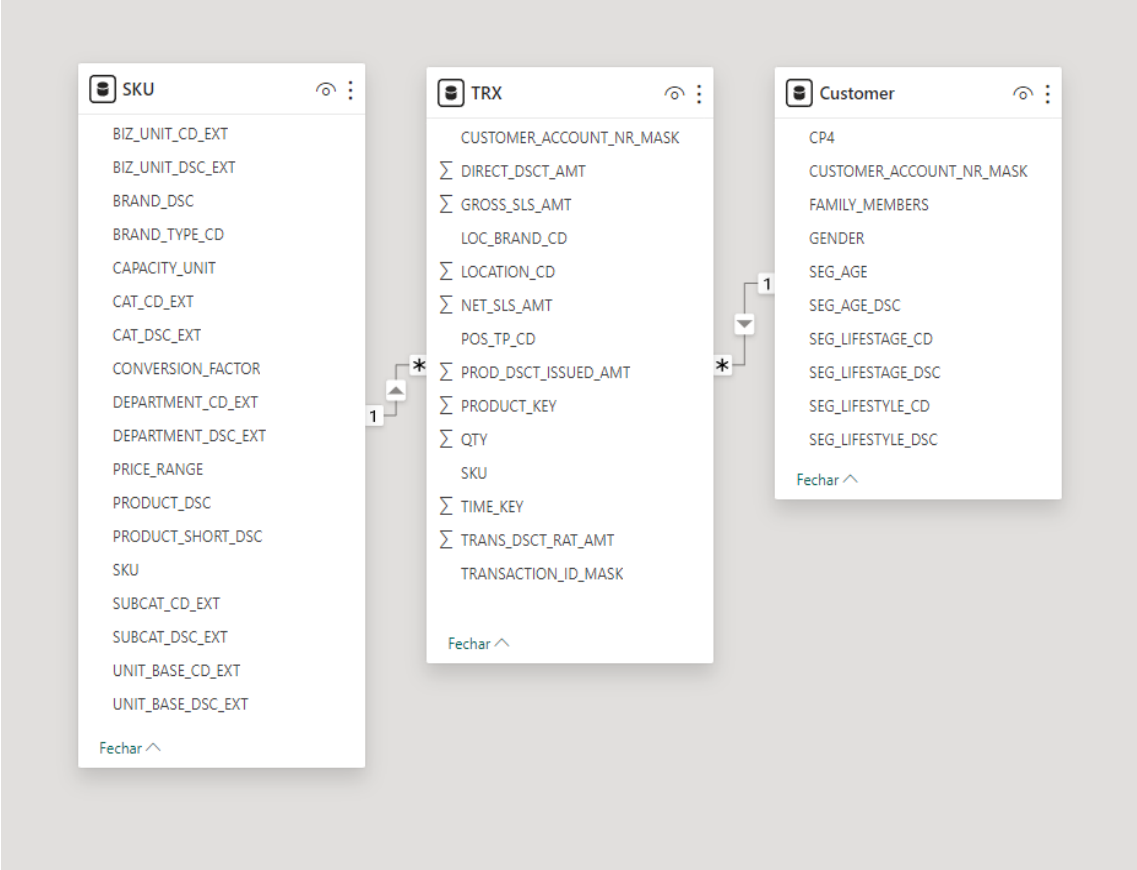

Figure 1: Product (SKU), Transaction (TRX) and Customer Data Tables

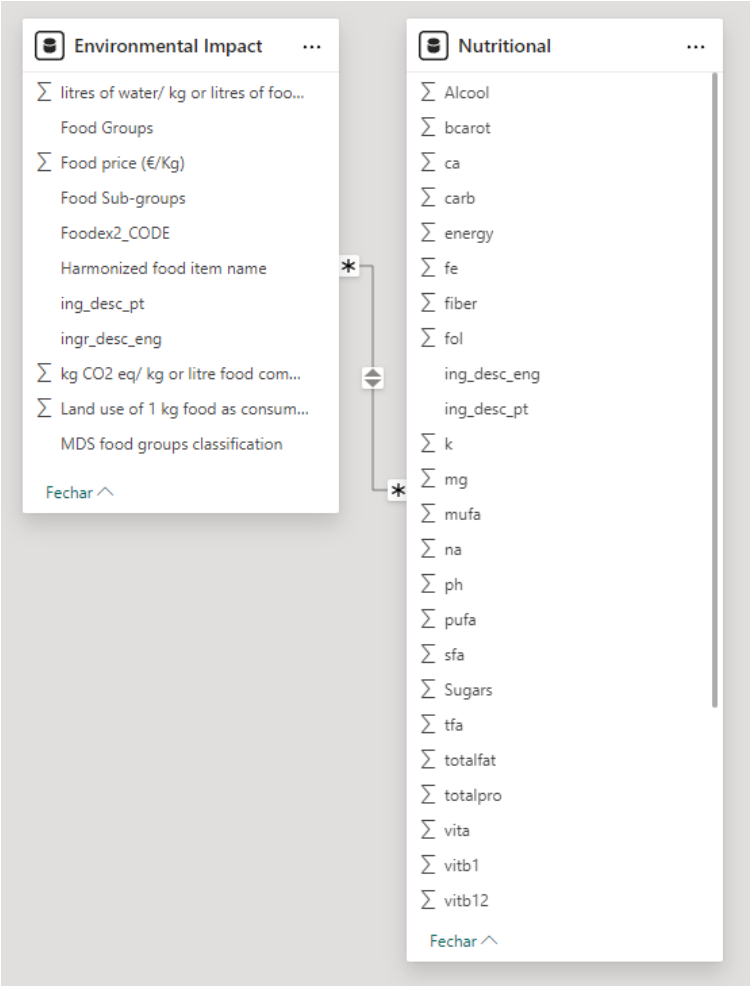

Figure 2: Nutritional and Environmental Impact Data Tables

2.2 Data Cleaning Process

The Data Cleaning Process section of this thesis is essential because it establishes the foundation for subsequent data analysis. It involves a series of steps to cleanse the Product, Nutritional and Environmental Impact tables, assuring its preparation for a successful merging process. This process is essential to synthesize separate tables into a singular, comprehensive table that embodies all product offerings from the retailer with their nutritional and environmental data.

2.2.1 Standardization of Product Descriptions

This analytical work begins with the critical step of importing the datasets, which involves transforming the data from its native storage format to a workable digital framework. Following the successful importation of the data, we immediately began a critical step aimed at improving the focus and relevancy of our datasets: the selective elimination of unnecessary variables. This purposeful process of elimination is driven by a focused approach to tables, in which only variables directly relevant to our research objectives are maintained.

An integral part of refining the Product table involved two key steps: replacing abbreviations in product descriptions and removing special characters. These steps were essential in transforming product descriptions into a more coherent and standardised format, which is essential to the subsequent data merging steps.

The procedure of replacing abbreviations was specifically designed to ensure uniformity across the dataset. While abbreviations are handy, they can generate ambiguity and inconsistency that might undermine data analysis. During our approach, we concentrated on a list of typical abbreviations seen in product descriptions and gradually replaced them with their full versions. The following abbreviations and alternatives were considered:

- "INTEG." was replaced with "Integral.";
- "INT." was replaced with "Integral.";
- "BOL." was replaced with "Bolacha.";
- "AZ." was replaced with "Azeite.";
- "BAT." was replaced with "Batata.";
- "MOID." was replaced with "Moida.";
- "MAS." was replaced with "Massa.";
- "ACUC" was replaced with "Acucar.";
- "QJ" was replaced with "Queijo.";

- "DE" was removed to enhance clarity and consistency.

This careful replacement of abbreviations was accompanied by the elimination of special characters from the product descriptions. Special characters can make text analysis difficult, frequently resulting in distorted or erroneous interpretations. By eliminating these features, our dataset was made more accessible and suitable to the subsequent analytical methods.

These procedures were performed together to standardise the product descriptions in the Product data table, having built the groundwork for our dataset's correct and insightful analysis by addressing the complex nature of textual data through systematic abbreviation replacement and the elimination of unusual characters.

2.2.2 Data Integration

The integration of the Product, Nutritional, and Environmental Impact data tables is the final stage of our data cleaning and preparation phase, encapsulating a complicated yet critical step in creating a comprehensive dataset for our research. This complicated merging process is based on detecting and using similar properties across the products names and the food names from the Nutritional and Environmental Impact datasets, which requires a comprehensive understanding of the data and a rigorous approach to assure the accuracy of the resulting dataset.

The initial phase in this integration process is to establish a common ground for linking, which is generally achieved through product descriptions. Given the intrinsic diversity of product nomenclature and the possibility of variance in descriptive terminologies, this phase uses text processing to harmonise the data. For example, normalising product descriptions in the Product data table required analysing the brand reference position, and concluding that the information following it was irrelevant. So, in order to normalise product descriptions, brand-

specific references and following information were deleted. This normalisation allows for more trustworthy comparisons with the Nutritional and Environmental Impact datasets, allowing the use of comparable text processing algorithms to guarantee that nutritional and environmental information are appropriately matched to the associated goods.

2.2.3 Transformation of Product Descriptions into Analytical Features

In order to allow the integration of data from several datasets, preliminary steps were necessary to guarantee precise product name matching. So, the words of each product description were separated into different columns. This method allowed for more detailed analysis, allowing to delve into the existence and frequency of matches within the product catalogue and the other datasets. By expanding the column of product descriptions into a broad format of numerous columns, each representing a unique word, the dataset was converted into a structured matrix that allowed for extensive textual analytics. This process was also applied to the Nutritional and Environmental Impact data tables, culminating in a different matrix for each dataset, facilitating the merging process through word matching.

The process of integrating and evaluating consisted of a mechanism that connects the Product dataset with the Nutritional and Environmental Impact datasets, individually. Through the dataset integration process, a new column entitled "Maximum Matched Words" was added to the Product dataset. This column indicates the maximum number of words that matched for any particular association. This was followed by the creation of separate columns for every unique match, where the number of matched words matched the value of the "Maximum Matched Words" column for each product item. The rows numbers from the related Nutritional and Environmental Impact datasets where a word

match was found were documented in these columns and were designated as "Nut_Match" and "Amb_Match," respectively. This strategy enabled us to populate each product row with several pairs of columns, each reflecting a unique match with food items from the other tables, resulting in a thorough mapping of textual overlaps between product descriptions and the properties of the other datasets.

Subsequently, a meticulous analysis was conducted on the columns containing the counts of matched words. This analysis consisted of comparing the number of words matched for a certain product in the Product table to the total number of words in the corresponding food item of the Nutritional or Environmental Impact tables. For example, if a product description had two matching terms with a specific food item in the Nutritional dataset, which comprised a total of three words, the matching efficiency would be 66% (2 out of 3 words matched). This percentage was a key factor in determining the best match for each product, emphasising food items from the Nutritional and Environmental datasets with the highest matching efficiency.

In cases where many food items had the same match % for a given product, additional factors were used to determine the best match. When compared to the Nutritional dataset, the food item with the largest calorie intake was preferred. When compared to the Environmental Impact dataset, the food item with the largest CO2 emissions per kg of food was selected, which was consistent with the research's environmental considerations. This selection criterion specifically intended to emphasise the matches with the greatest potential negative impact, whether in terms of increased calorie consumption or environmental load.

The result of these efforts is a uniform dataset that not only connects products with their nutritional information, but also with their environmental effect evaluations, by having a specific column that matches with a specific food item for both nutritional and environmental datasets. This result is more than just

technical, as it marks a big step toward developing a complete analytical framework capable of accommodating complex research questions on product sustainability and customer behaviour.

2.2.4 Data Filtering

The Transactions table needed to be filtered before the customer information could be analysed. This thesis focused on a specific temporal scope by only evaluating transactions from the year 2022. The goal is to distil the data so that it accurately reflects customer behaviour over a specific time period. Further enhancing is obtained by concentrating on recurrent client patterns, categorized as those with at least 12 separate purchase instances, since non-recurrent customers would have a sizeable impact on the analysis, since there are numerous, as seen in Figure 3, which only reflects customers with less than 50 purchase instances, in order to properly analyse the non-recurrent customers.

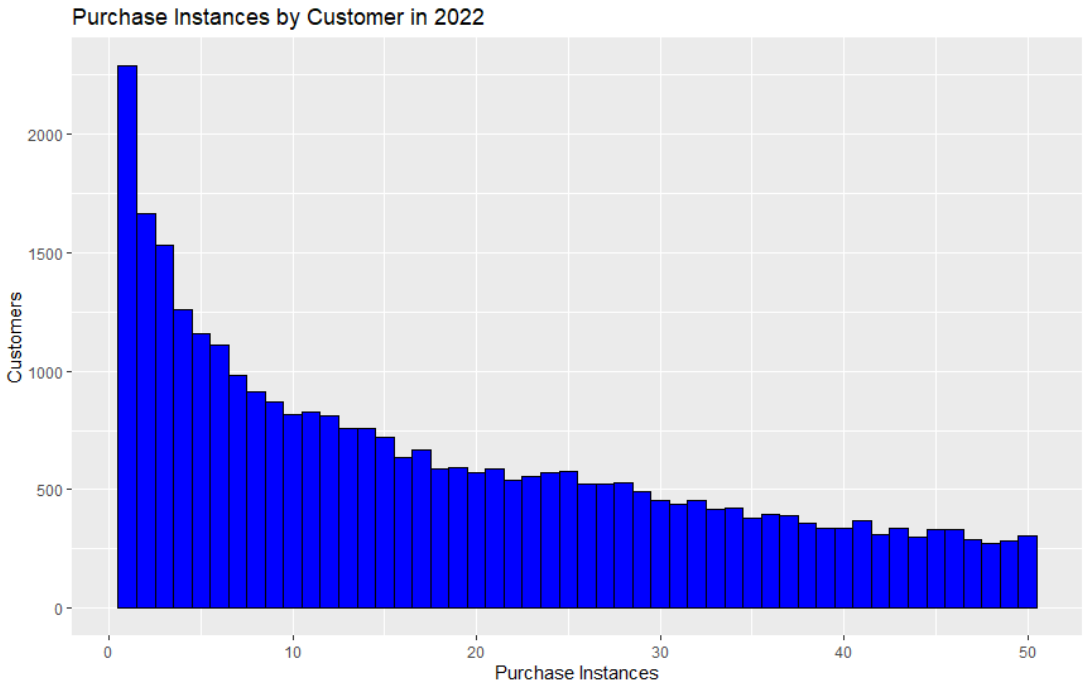

Figure 3: Purchase Instances by Customer

Furthermore, to allow for an accurate per person assessment of nutritional and environmental implications, records of customers who did not provide family size information were eliminated. The accuracy of this research depends on an accurate picture of household consumption, which demands this data filtering step.

2.2.5 Data Aggregation

After preparing the table for analysis, one essential step remained: combining nutritional and environmental data at the consumer level. This step was methodically carried out using a series of SQL queries built with accuracy to incorporate nutritional and environmental characteristics into the product table. Initially, this procedure entailed attaching nutritional and environmental data to each product and altering these statistics based on the conversion rates (grammes or kilogrammes) appropriate to that item. Following that, for each client transaction, these corrected values were multiplied by the quantity of each product purchased to ensure that the generated metrics accurately reflected actual consumption. This procedure culminated in the aggregation of these enhanced statistics at the customer level, resulting in a comprehensive overview of each customer's nutritional consumption and environmental effect. To ensure that the study accounted for household size, these aggregated statistics were then normalised on a per capita basis by dividing the total consumption by the number of family members, thereby standardising the data for meaningful analysis.

2.3 Data Modelling

2.3.1 Data Clustering

The k-means clustering approach was chosen as the method for categorising clients into separate groups based on their consumption habits across many factors. This technology, known for its ability to discover homogeneous subgroups within datasets, enabled a systematic approach to categorising individuals based on their nutritional consumption and environmental effect. The employment of this technique aims to identify important patterns and correlations in consumer data, giving information on customers' food habits and environmental footprints.

The z-score approach was used to normalise the dataset before clustering, removing inequalities in different scales by standardising data to a similar scale, increasing the efficiency of the k-means clustering algorithm in classifying clients based on nutritional and environmental characteristics.

The k-means clustering approach was used to examine three sets of variables: macronutrients (energy, lipids, carbs, and protein), micronutrients (salt, monounsaturated acids, vitamin B12, and fibre), and environmental effect (CO2 emissions, water consumption, and land use per kilogramme of food). This clustering process was conducted across the following segments individually:

- **Macro nutritional Segment Clustering:** The first segment focuses on client purchases connected to macro-nutritional characteristics. This includes their preferences for energy (kcal), proteins, carbohydrates and lipids. Each customer was effectively classified into a different cluster based on their macro nutritional shopping patterns within this segment.
- **Micro Nutritional Segment Clustering:** The second segment focuses on micro-nutritional factors, including vitamin B12, fibre, salt and

monosaturated acids. Using the k-means algorithm enabled the clustering of clients based on their particular micro nutritional preferences, with each customer having a profile that is consistent with others in their cluster, indicating comparable micro nutritional purchase habits.

- **Environmental Impact Segment Clustering:** This research investigates how customers' purchases affect the environment, taking into account aspects such as carbon footprint, water use, and land use associated with the things purchased by customers. This segmentation connected each consumer to a group with comparable environmental consequences, allowing us to better understand and categorise their environmental consumption habits.

The Elbow method, a methodology employed for determining the ideal cluster count in k-means clustering, was used to find the quantity of clusters to be selected for each cluster analysis. In this analysis, there were selected 3 clusters for each of the segments, as it's perceived by the information present in Figure 4, Figure 5 and Figure 6, respectively. This method, which focuses on determining the point at which increasing the number of clusters no longer results in considerably lower within-cluster variances, serves as a heuristic guide rather than providing precise accuracy.

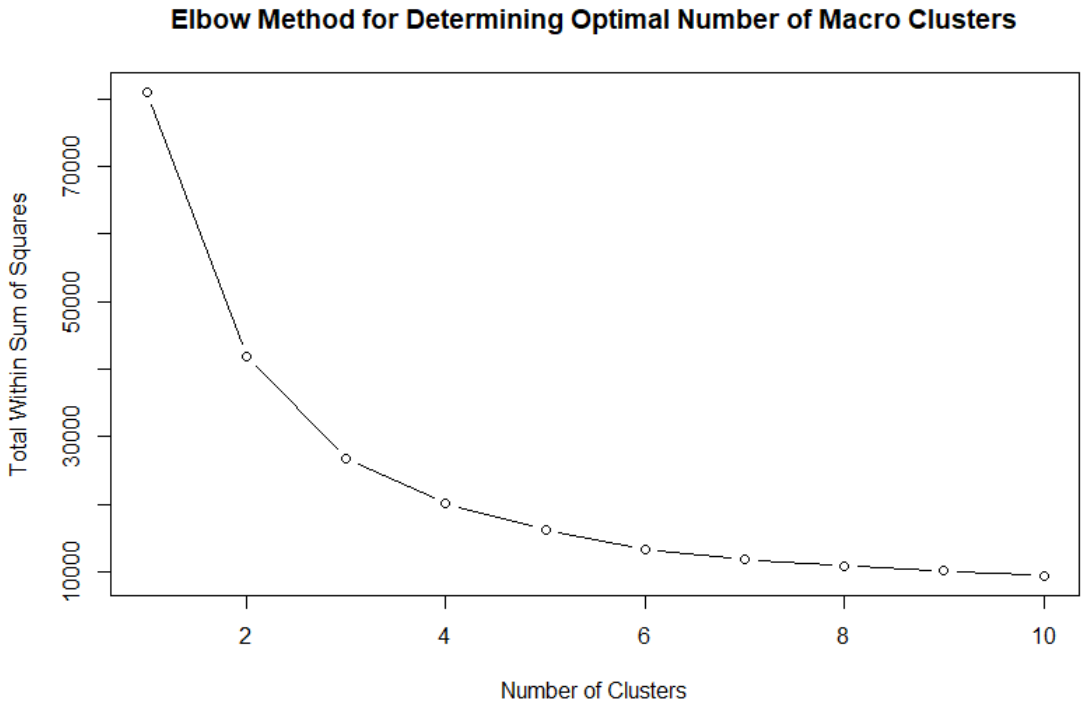

Figure 4: Elbow Method of Macronutrient Clustering

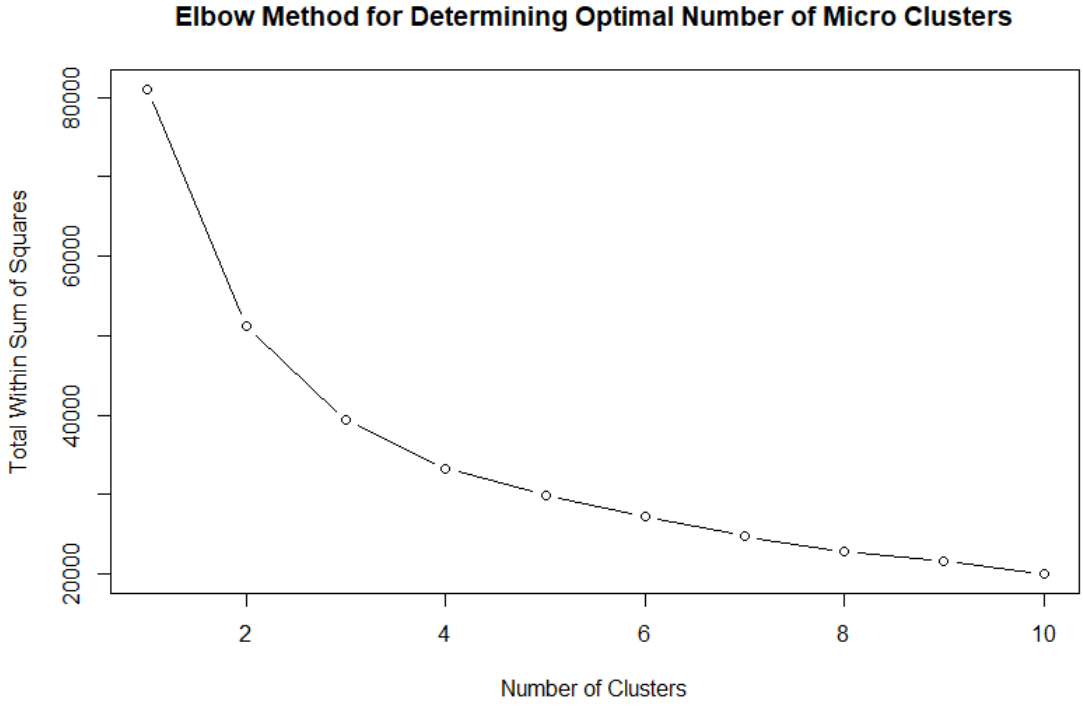

Figure 5: Elbow Method of Micronutrient Clustering

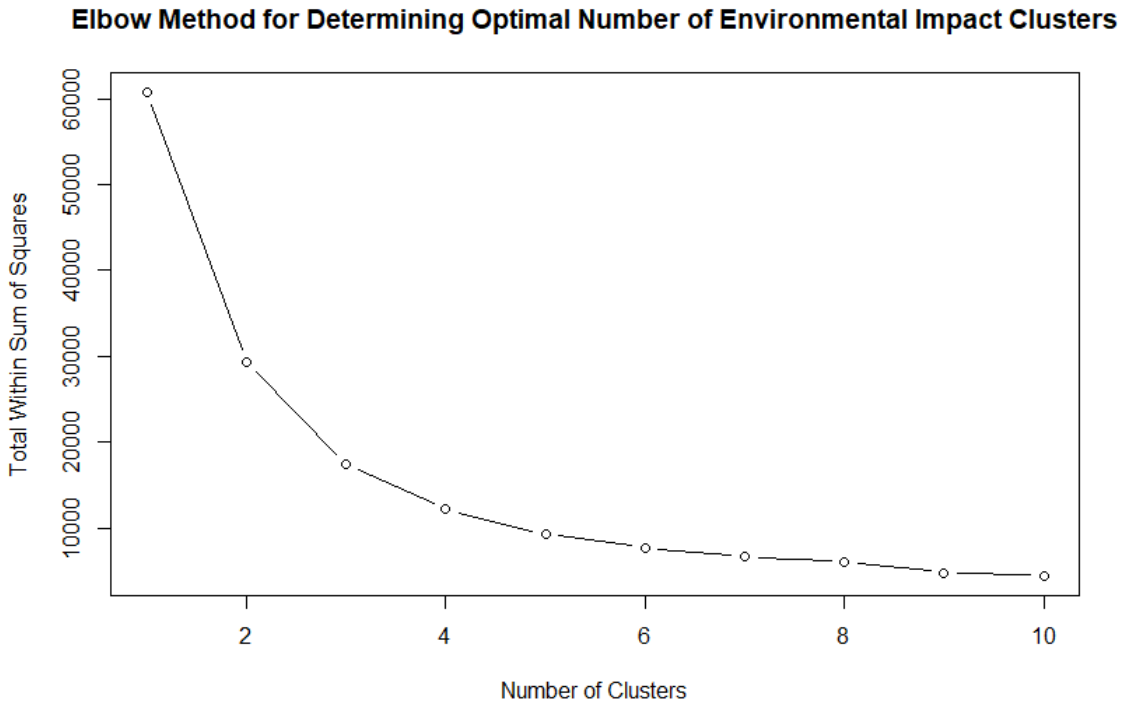

Figure 6: Elbow Method of Environmental Impact Clustering

3. Findings and Discussion

This section of the research is dedicated to an in-depth analysis of the data obtained from consumer segmentation based on purchase behaviour, with a focus on nutritional and environmental factors. The k-means clustering technique was meticulously applied to identify different customer groups or clusters, exhibiting patterns and trends that give important insights into consumer preferences and behaviours. The following analysis will dissect these segments, providing a thorough examination of the underlying traits that distinguish each group.

To lay the groundwork for a thorough assessment of customer segments, particularly from a nutritional standpoint, the European Food Safety Authority's (EFSA) rules on the daily recommended consumption for key nutrients for adults of both genders were used. Such data was averaged to create an annualised baseline of recommended intake for each nutritional indicator examined in this study. The estimated averages serve as crucial criteria for evaluating the eating habits observed in our customer categories. These benchmarks, which are summarised into Table 1, give a measurable basis for comparing the actual consumption patterns revealed by our consumer segmentation to the EFSA's ideal dietary requirements.

	Energy (Kcal)	Protein (g)	Lipids (g)	Carbohydrates (g)	Fibre (g)	Vitamin B12 (µg)	Salt (g)	Monosaturated Acids (g)
Daily	2383	89	66	357	25	4	ALAP [1]	ALAP[1]
Yearly	869804	32618	24161	130471	9125	1460	ALAP[1]	ALAP[1]

Table 1: EFSA's Average Yearly Recommendation for Nutrient Consumption

3.1 Macronutrients Clustering Analysis

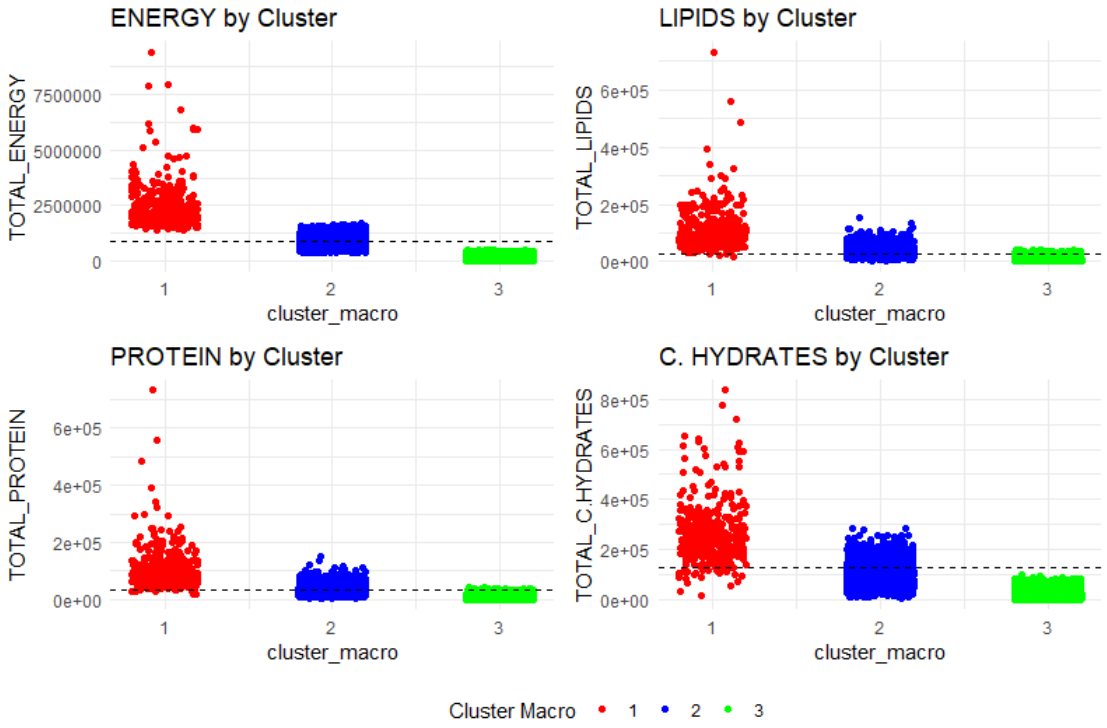

Figure 7: Macronutrients Clustering

The Macro Cluster analysis, as shown in Figure 7, demonstrates distinct consumption patterns across the three groups. In this analysis, energy is quantified in kilocalories, while proteins, lipids, and carbohydrates are measured in grams. Customers in Cluster 3 had nutritional intakes that are somewhat lower

[1] ALAP – As Low as Possible

than ideal, which might reflect a conservative approach to consumption or perhaps more health-conscious decisions. In contrast, Cluster 1's consumption generally exceed the recommended intake, indicating a tendency for increased nutrient consumption. Cluster 2 represents a balanced consumption, with core values that oscillate about the reference benchmarks with some being slightly higher and other being slightly lower, but all roughly fitting with the recommended intake amounts.

A deeper look at the customer dispersion within each cluster for each specific nutrient reveals a more detailed picture. The study breaks down the association between nutrient consumption and purchase occurrences, as shown in Figure 12, Figure 13, Figure 14 and Figure 15 from the appendix. This study demonstrates that the number of purchase occurrences does not always correspond with nutrient consumption. This discovery reinforces the assumption that frequency of shopping does not immediately correlate with the volume or nutritional worth of food purchased, since a consumer who shops less regularly may acquire a bigger or more nutrient-dense haul than a regular shopper.

Furthermore, the financial expenditure analysis for nutrient intake reveals a consistent trend: higher nutritious consumption correlates with increasing expenditures. Customers in Macronutrient Cluster 1, who have greater levels of macronutrient consumption, tend to spend more, illustrating consumer behaviour in which the amount or quality of food choices influences total expenditure.

3.2 Micronutrients Clustering Analysis

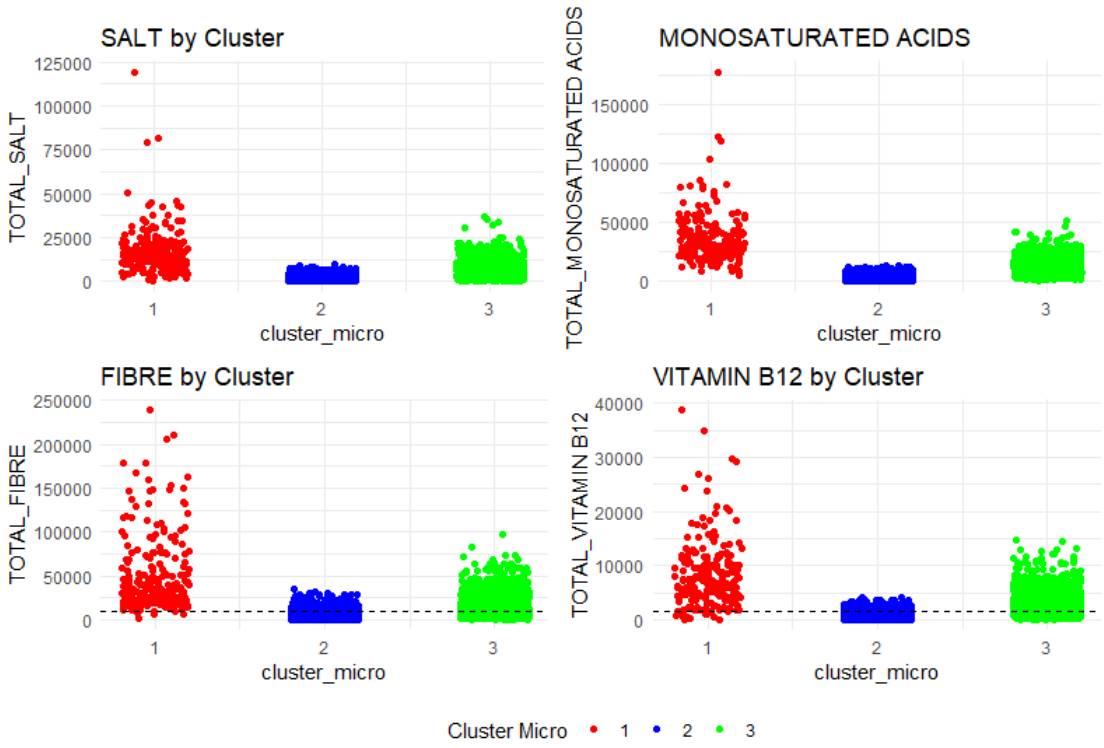

Figure 8: Micronutrients Clustering

Figure 8 illustrates a micro-nutrient cluster analysis, which provides a detailed look at consumer dietary habits, with an emphasis on salt, monounsaturated acids, fibre, and vitamin B12. According to the European Food Safety Authority (EFSA), there is no 'ideal' daily consumption for salt or monounsaturated acids; instead, it is recommended that the consumption of such nutrients should be kept to a minimum. Micronutrient Cluster 2 appears as the section closest to the required consumption amounts for vitamin B12 and fibre, indicating a dietary profile that is consistent with the EFSA's recommendations. This cluster is especially notable for its decreased intake of salt and monounsaturated acids, showing a preference for better nutritional options.

Micronutrient Cluster 3 includes clients with intermediate consumption levels, frequently exceeding the optimal nutritional benchmarks, indicating a tendency of overconsumption relative to dietary guidelines. Micronutrient

Cluster 1, on the other hand, although having fewer observations and a sparser distribution in the graph, shows nutrient consumption that is significantly higher than the optimal range.

Subsequent analyses, shown in Figure 16, Figure 17, Figure 18 and Figure 19 from the appendix, explores the relationship between nutrient consumption, purchase frequency, and spending. The findings confirm that the number of purchase instances has no direct association with nutrient consumption, validating the assumption that the frequency of purchases does not imply quantity of nutrient intake. On the other side, there is a positive relationship between spending and nutrient consumption, with more expenditure frequently accompanied increased nutrient intake.

3.3 Environmental Impact Clustering Analysis

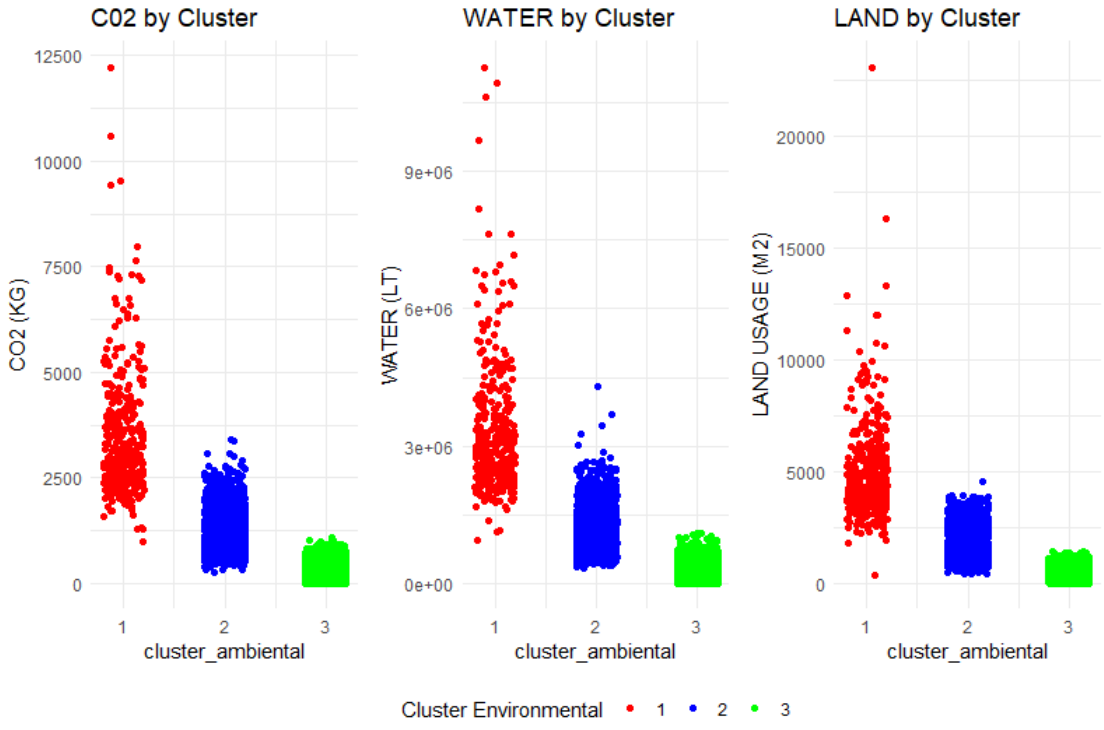

Figure 9: Environmental Impact Clustering

When analysing the environmental effect of food intake, lower values are associated with better outcomes because there are no set 'optimal' criteria for CO2 emissions, water, and land utilisation. From all the clusters, Environmental Impact Cluster 3 is the one composed by customers with the lowest average environmental effect across all three variables, indicating a consumption pattern linked with more environmentally responsible choices, as shown in Figure 9.

Environmental Impact Cluster 1, on the contrary, represents the opposite end of the spectrum, being composed by customers with the greatest impact on the environment. This might be suggestive of consumption choices that, either to the nature of the items or the amount consumed, have a larger ecological impact. Environmental Impact Cluster 2 has an intermediate position, characterised by its customer's environmental effect sitting between the two extremes, implying a moderate amount of environmental impact from food intake.

Figure 20, Figure 21 and Figure 22 from the appendix, provide a disaggregated examination of the relationship between each environmental element, the frequency of purchases and monetary spending. The statistics show no clear association between the number of purchase instances and a customer's environmental effect, refuting the notion that frequent shopping always results in increased ecological strain. However, there is a positive association between expenditure levels and environmental effect, indicating that consumers who spend more have a higher environmental footprint.

3.4 Customers within Clusters

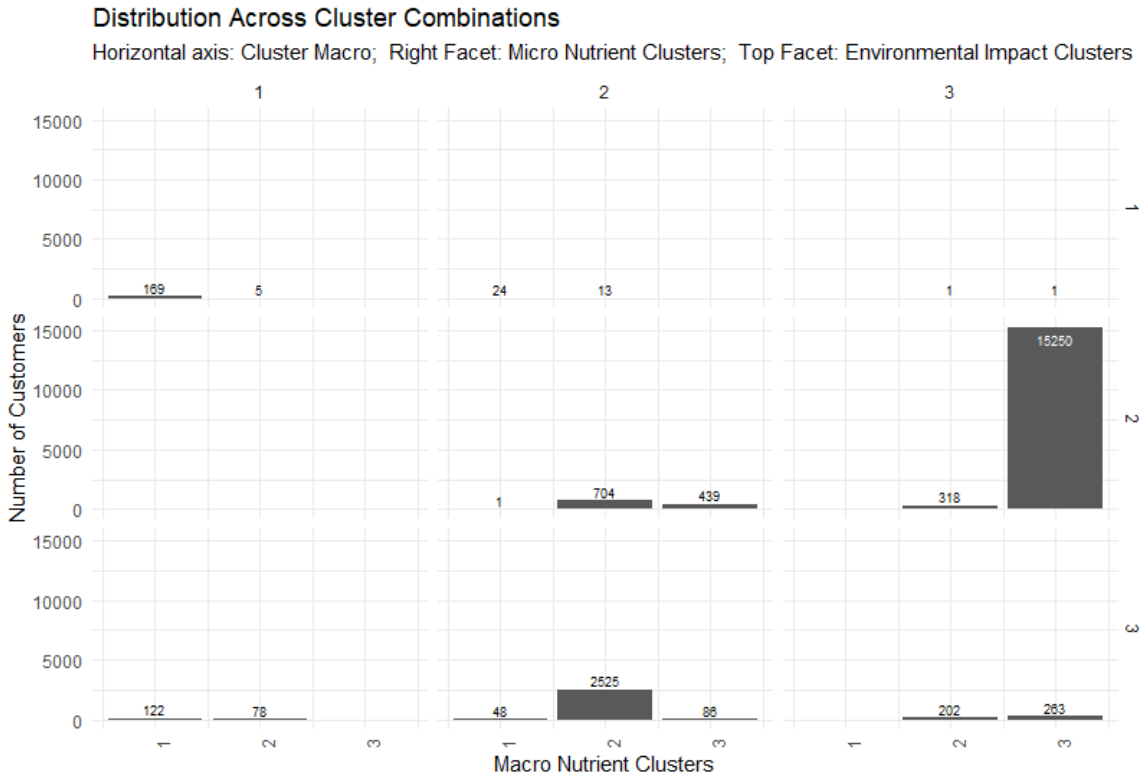

Figure 10: Customer Distribution by Combination of Clusters

The Figure 10 depicts the distribution of customers across different cluster combinations. The horizontal axis shows the Macronutrient Clusters. The Micronutrient Clusters are presented on the right facet, while the Environmental Impact Clusters are presented at the top. The vertical axis represents the number of customers in each cluster combination. The visualisation compares the number of customers across these 3 dimensions, offering a multidimensional picture of the combination of segments.

The analysis of customer distribution across clusters gives significant insights regarding their macronutrient, micronutrient, and environmental consumption patterns. As seen in Figure 10, there is a large concentration of customers within particular cluster combinations, indicating common eating and shopping habits.

The graph indicates that the highest concentration of customers is found in Cluster 3 for both macronutrient and environmental clusters, as well as Cluster 2 for micronutrient cluster, indicating a segment of customers with lower than ideal macronutrient intake, nearly ideal micronutrient consumption, and a low environmental footprint when compared to customers in other clusters. This clustering pattern indicates that dietary choices that lead to decreased macro and micronutrient consumption are associated with a low environmental effect.

Furthermore, the graph shows a significant concentration of customers within Macronutrient Cluster 2, Micronutrient Cluster 3, and Environmental Impact Cluster 2. This confluence denotes a customer group that follows the recommended macronutrient intake levels while slightly exceeding the optimal micronutrient intake. Concurrently, these persons have an intermediate environmental impact. This distribution validates the hypothesised association, which states that an increase in macro and micronutrient consumption is related with a higher environmental footprint for consumers. Such findings contribute to a better understanding of eating patterns and their larger ramifications, indicating that maintaining a nutritional balance may be critical in minimising environmental consequences.

3.4.1 Customer Age Distribution

The analysis of client demographics within each cluster combination also included a review of the age distribution. This stage was conducted with the deliberate aim of determining the prevalent age ranges at each cluster combination. The most populated cluster combination, consisting of Macronutrient Cluster 3, Micronutrient Cluster 2, and Environmental Impact Cluster 3, had a wide age representation (15,250 observations). Notably, 23% of those in this group were between the ages of 45 and 55, with 20% in the 35 to 45

age bracket and 18% in the 55 to 65 age bracket. The distribution across other age segments was less prominent, showing a generally equal age distribution, with a considerable share falling in the 45 to 55-year-old group, accounting for nearly a quarter of the total observations.

Conversely, the cluster combination of 2,525 observations, constituted by customers that are within Macronutrient Cluster 2, Micronutrient Cluster 3, and Environmental Impact Cluster 2, showed a concentration of adults aged 45 to 55, accounting for 24% of the observations. The succeeding age group of 35 to 45 years accounted for 21% of the sample, with lower amounts scattered throughout other age groups. This data shows that, while this sector has a wide age range, there is a noticeable preponderance of middle-aged customers.

3.4.2 Customer Geographic Distribution

The combination of clusters provided data for a geographical analysis of customers, which aimed to determine the major districts in Portugal represented within particular cluster combinations. This geographical investigation used postal code data from the customer database, first simplifying postal codes to their first two digits, and then translating the various postal codes into their respective districts. This simplification approach efficiently delineates the geographical area of Portugal, allowing for the grouping of these codes by district.

Focusing on the two combinations with the most observations provides valuable information regarding client dispersion across geographies. For the combination of Macronutrient Cluster 3, Micronutrient Cluster 2, and Environmental Impact Cluster 3, referred as combination "323", which includes 15,250 customers, the distribution per district is as follows:

Postal Code	District	% of Cluster
40	Lisbon	20%
10	Porto	19%
75	Setúbal	8%
47	Braga	7%
24	Leiria	6%

Table 2: Distribution of Customers within Cluster Combination "323" by District

The analysis illustrates the geographical composition of this cluster combination, showing Lisbon and Porto as the districts with the highest customer presence. A similar analysis was performed on the combination of Macronutrient Cluster 2, Micronutrient Cluster 3, and Environmental Impact Cluster 2, referred as combination "232", which included 2,525 observations. The regional distribution revealed similar patterns:

Postal Code	District	% of Cluster
40	Lisbon	20%
10	Porto	19%
75	Setúbal	8%
47	Braga	6%
24	Leiria	6%

Table 3: Distribution of Customers within Cluster Combination "232" by District

Cluster combinations indicate that Lisbon and Porto have the greatest number of customers, with Setúbal, Braga, and Leiria all making substantial contributions.

3.4.3 Customer Distribution by District

3.4.3.1 Customers within Unfavourable Segments

This section delves into a detailed examination of client distribution throughout the top five districts, concentrating on how they correspond with optimal consumption levels and environmental effect clusters. The goal is to identify the groups of customers in each district that do not follow specified consumption patterns, for example those who satisfy one of the optimum guidelines but fall short of the other targets. This granular method enables a more detailed understanding of how geographical location affects food habits and environmental impact across different consumer segments.

Further assessment of the consumer distribution throughout the top five districts reveals an interesting trend, particularly when the incidence of unfavourable clustering is considered. Table 4 shows the fraction of customers in each district that are in unfavourable clusters, defined as those with high nutritional consumption or high environmental impact. This distinction is crucial for understanding the complexities of consumer behaviour across geographic regions.

District	**% of customers within at least 1 unfavourable cluster**
Lisbon	3%
Porto	2%
Setúbal	2%
Braga	1%
Leiria	3%

Table 4: Customers within at least 1 unfavourable cluster

The data reveals that just a small percentage of customers in all assessed districts fall into unfavourable clusters, indicating healthy and sustainable consumption patterns. Notably, Lisbon and Leiria have the highest percentages,

with 3% of their customer base assigned to clusters linked with either excessive nutrient intake or a substantial environmental imprint. Porto and Setúbal have a little better picture, with only 2% of consumers falling within these less favourable groups. Braga distinguishes out for its lower incidence of 1%, implying a better alignment with ideal consumption and environmental norms among its residents.

3.4.3.2 Customers with Ideal Nutrient Consumption and Environmental Impact

In this section, we delve into the intersection of balanced nutrition and environmental sustainability in the top five districts, focusing on the percentage of customers who are able to align their consumption with optimal levels of both micro and macronutrients while also having a low environmental impact.

District	**% of customers with ideal Nutritional Consumption and low Environmental Impact**
Lisbon	1%
Porto	1%
Setúbal	2%
Braga	1%
Leiria	2%

Table 5: Customers with Ideal Nutritional Consumption and Low Environmental Impact Clusters

This uniform low percentage underscores a critical insight: despite growing awareness and initiatives aimed at encouraging sustainable living and healthy eating habits, Portugal faces substantial challenges in achieving these aims on a larger scale. The data suggests that while there is a conscientious portion of the community trying towards sustainability and nutritional adequacy, the majority

remains outside the desired thresholds for creating a significant positive impact on both health and environment.

3.4.3.3 Customers with Ideal Macronutrient Consumption

This section of the analysis focuses on a specific consumer behaviour pattern across the previously mentioned districts, emphasising the prevalence of people who consume ideal macronutrient levels but do not meet ideal standards for micronutrient intake or environmental impact. Table 6 provides a thorough look at the percentage of clients in each district who fit into this consumption niche.

District	% of customers within Ideal Macronutrient Consumption Segment
Lisbon	17%
Porto	17%
Setúbal	17%
Braga	16%
Leiria	16%

Table 6: Customers within Ideal Macronutrient Consumption Cluster

This data highlights an important insight: while a sizable fraction of the population follows recommended macronutrient intake levels, there is still a gap in obtaining full nutritional balance while minimising environmental effect.

3.4.3.4 Micronutrient Consumption and Environmental Impact Analysis by District

This section focuses on analysing the customers within each district that achieve an optimal mix of micronutrient consumption and minimal environmental effect. The data is especially striking when compared to the previously analysed category of customers who manage macro and micronutrient balance as well as environmental concerns. Table 7 offers a

detailed breakdown of the percentage of clients in each district that fall into this consumption niche.

District	% of customers within ideal Micronutrient Consumption and Environmental Impact segments
Lisbon	77%
Porto	77%
Setúbal	77%
Braga	81%
Leiria	77%

Table 7: Customers Within Ideal Micronutrient and Environmental Impact Clusters

The data shows that 77% of customers in Lisbon, Porto, Setúbal, and Leiria, and 81% in Braga, fall within our concept of ideal micronutrient consumption and environmental sustainability. This suggests that a sizable proportion of the population follows eating patterns that are both healthy and sustainable in terms of micronutrients and the environment.

However, when this data is analysed in the context of the entire nutritional and environmental sustainability research, a more complex story emerges. Despite the high percentage of consumers in these areas complying to optimum micronutrient consumption and environmental impact requirements, there is a noticeable divergence in macronutrient intake.

This indicates that, while the percentage of consumers in optimum groups in terms of micronutrient consumption and environmental effect is quite high, there is still room for these customers to enhance their macronutrient intake.

3.4.3.5 Cluster Combinations by District

As we delve deeper into customer segmentation statistics, it is possible to look at how cluster combinations are distributed throughout Portugal's five districts with the most observations. Understanding the prevalence of these combinations within each district allows us to get a sense of how customers' nutritional and environmental footprints vary by area. Cluster combinations have been decomposed individually in

Table 8, serving as foundation to interpret the pie charts in Figure 11, which cluster combination is ordered starting from Macronutrient Cluster, followed by the Micronutrient Cluster and Environmental Impact Cluster.

Table 8: Clusters Description

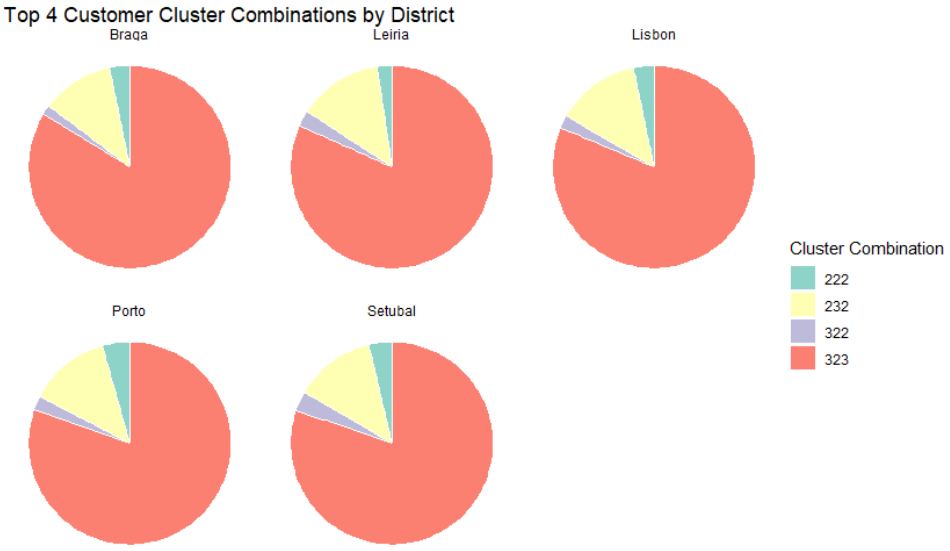

Figure 11: Distribution of Cluster Combinations by District

From the visual data representation of the biggest four cluster combinations distribution by district, we may determine that cluster combination “323” is the most common across all districts, indicating consumers who have a lower-than-ideal macronutrient consumption but an optimal micronutrient intake and

environmental impact. This implies a cautious yet ecologically mindful approach to consuming.

Conclusion

This thesis took a thorough approach, beginning with an in-depth investigation of the dataset at our disposal. The major purpose was to navigate through the intricacies of retail data in order to discover useful insights for sustainability and health. Our goal was not only to analyse the data, but also to convert it into a format that would allow for extensive study of the nutritional content and environmental effect of the retailer's product offers. This method involved thorough data cleaning and merging to ensure the integrity and trustworthiness of the information employed in our subsequent research.

Preparing the datasets for analysis required an extensive set of steps to clean and integrate the data. This approach was critical in obtaining correct nutritional and environmental information for each product. The tools and procedures used were diverse and complex, ranging from basic data cleaning methods to more advanced data transformation techniques. This phase's meticulousness laid the groundwork for the subsequent analytical work, ensuring that the insights gained were both correct and relevant to this thesis' objectives.

With a clean and extensive dataset in hand, this study moved on to the modelling step, which turned the focus to consumer segmentation. The choice to implement three unique clustering activities was intentional, with the goal of dissecting the customer base from several perspectives, with them being the macronutrient intake, the micronutrient consumption, and the environmental effect. This strategy was critical for acquiring a deeper understanding of Portuguese clients' food patterns and the consequences for environmental sustainability. Each clustering attempt contributed to a layered analysis,

allowing us to group clients not just based on their food choices, but also on the larger environmental implications of those decisions.

Portugal's transition to sustainable consumption has shown good indications, most notably in the significant grouping of consumers with optimal micronutrient intake and low environmental impact. This demonstrates a growing knowledge about health and sustainability in dietary choices. However, the small proportion of the population that fits both optimum nutritional and environmental parameters deserves to be thought about, as this section is far less frequently observed than expected, indicating that there's a lot of work to be done regarding shifting consumer habits into a more healthy and sustainable consumption. Regarding this matter, addressing the macronutrient intake gap, that seems to be shared by most of the Portuguese population represented in this analysis, requires policies targeted at sensitising and educating these specific customer segments. It is vital to emphasise the tremendous effects that macronutrient consumption has on an individual's health, and at the same time dispel the misperception that consuming optimal nutritional thresholds requires more money or consumption. Instead, it is possible to achieve this equilibrium by fine-tuning existing consumption patterns. Such a complex approach can promote alignment with health guidelines while maintaining sustainability goals.

While just a small percentage of the entire customer base has higher-than-optimal nutritional intakes, the health consequences are considerable. The ideal approach would consist of countrywide awareness programmes that would increase the current knowledge of customers on the negative health effects of excessive nutrient consumption. Consumers would be guided towards more balanced dietary choices that promote long-term health and well-being by

being provided focused information that highlights the hazards associated with high salt consumption and over-nutrition.

Addressing the segment of customers that have a larger environmental effect presents a unique challenge. A potential start would be the addition of visible, understandable environmental effect labels on products. Such labels would allow customers to make more informed selections, perhaps leading to a shift towards items with a reduced environmental impact. If these informational techniques fail to effect change, stronger governmental actions should be considered. For example, altering VAT rates to favour environmentally friendly items might act as a financial incentive for customers to shift towards sustainability. This might create a market that naturally gravitates towards lower-impact items, lowering the overall environmental load.

The prescriptive insights presented in this study are only a few of the many potential solutions targeted at balancing human health and the environment. The recommendations presented here are meant to encourage additional debate and creativity in policymaking and consumer education, encouraging lower environmental impact and improved quality of life. The goal of this study is to provide the groundwork for future investigations on household purchasing habits. The route to sustainable and health-conscious living is varied and difficult, but by rigorous research and innovative, educated policymaking, we can direct society towards a more responsible and nutritious future.

Limitations

In the pursuit of deriving meaningful insights into the nutritional and environmental behaviours of Portuguese consumers, it's crucial to recognize the inherent limitations of our dataset and analytical approach. These limits, although not decreasing the significance of our findings, do define the context in which they should be interpreted.

One important drawback is the dataset's focus on consumer behaviour, which is mostly supplied from a single retailer. This dataset does not account for the likelihood that customers would also buy things from other retailers, which could introduce bias into our study by not completely capturing all of consumers' shopping behaviours.

Furthermore, the quality of demographic information, particularly household size and age groups, adds another degree of complication. If this information was not consistently updated, there is a danger of distorting per capita consumption patterns. Households that have recently undergone changes, such as the addition of new members or the ageing of existing ones, may have consumption patterns that are inaccurately portrayed due to obsolete data. This difference might result in an overestimation or underestimating of per capita consumption.

Another factor to consider is the approach used to calculate optimal consumption numbers. By averaging these values for adults, the study may neglect the complex nutritional requirements of different family compositions, particularly those with children. Households with younger members may have different nutritional needs and environmental footprints than those with only adults, implying that a one-size-fits-all approach to optimal eating may not be entirely accurate.

These limitations, inherent to the nature of the data and the analytical framework, underscore the importance of cautious interpretation of the findings. While the findings of this thesis provide useful insights into consumer behaviour in terms of nutrition and sustainability, they should be seen as indicative rather than final. The potential impact of unaccounted purchases from other retailers, the variability in household composition, and the generalisation of ideal consumption values all point to the importance of understanding the underlying complexities when drawing conclusions and recommendations.

Statement

During the preparation of this work the author used:

CHATGPT in order to retrieve suggestions of optimization of R programming functions;

QUILLBOT in order to improve the quality of writing;

After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

Bibliography

Bôto, J.M., Neto, B., Miguéis, V., & Rocha, A. (2014) "Development of the Dietary Pattern Sustainability Index (DIPASI): a novel multidimensional approach for assessing the sustainability of an individual's diet", *Sustainable Production and Consumption* (under revision)

Edjabou, L. D., & Smed, S. (2013). The effect of using consumption taxes on foods to promote climate friendly diets – The case of Denmark. *Food Policy*, *39*, 84–96. https://doi.org/10.1016/j.foodpol.2012.12.004

Eini-Zinab, H., Shoaibinobarian, N., Ranjbar, G., Norouzian Ostad, A., & Sobhani, S. R. (2021). Association between the socio-economic status of households and a more sustainable diet. *Public Health Nutrition*, *24*(18), 6566–6574. Scopus. https://doi.org/10.1017/S136898002100402X

Nau, C., Kumanyika, S., Gittelsohn, J., Adam, A., Wong, M. S., Mui, Y., & Lee, B. Y. (2018). Identifying Financially Sustainable Pricing Interventions to Promote Healthier Beverage Purchases in Small Neighborhood Stores. *Preventing Chronic Disease*, *15*, 160611. https://doi.org/10.5888/pcd15.160611

Rondoni, A., & Grasso, S. (2021). Consumers behaviour towards carbon footprint labels on food: A review of the literature and discussion of industry implications. *Journal of Cleaner Production*, *301*, 127031. https://doi.org/10.1016/j.jclepro.2021.127031

Ruiz-Real, J. L., Uribe-Toril, J., Gázquez-Abad, J. C., & De Pablo Valenciano, J. (2019). Sustainability and Retail: Analysis of Global Research. *Sustainability*, *11*(1), Artigo 1. https://doi.org/10.3390/su11010014

Semba, R. D., Rahman, N., Du, S., Ramsing, R., Sullivan, V., Nussbaumer, E., Love, D., & Bloem, M. W. (2021). Patterns of Legume Purchases and Consumption in the United States. *Frontiers in Nutrition*, *8*. https://www.frontiersin.org/articles/10.3389/fnut.2021.732237

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A Systematic Literature Review on Applying CRISP-DM Process Model. *Procedia Computer Science*, *181*, 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Suryadevara, C. K. (2020). *Predictive Analysis for Big MartSales using Machine Learning Algorithms* (SSRN Scholarly Paper 4591993). https://papers.ssrn.com/abstract=4591993

Upham, P., Dendler, L., & Bleda, M. (2011). Carbon labelling of grocery products: public perceptions and potential emissions reductions. *Journal of Cleaner Production, 19(4)*, 348-355.

Van Dooren, C., Marinussen, M., Blonk, H., Aiking, H., & Vellinga, P. (2014). Exploring dietary guidelines based on ecological and nutritional values: A comparison of six dietary patterns. *Food Policy*, *44*, 36–46. Scopus. https://doi.org/10.1016/j.foodpol.2013.11.002

Wagner, T., Lutz, R. J., & Weitz, B. A. (2009). Corporate hypocrisy: Overcoming the threat of inconsistent corporate social responsibility perceptions. *Journal of Marketing, 73(6)*, 77-91.

Appendixes

Variable	Description
CUSTOMER_ACCOUNT_NR_MASK	Customer Masked Identifier
DIRECT_DSCNT_AMT	Discount Value Associated to a Product
GROSS_SLS_AMT	Sales Value
LOC_BRAND_CD	Brand Identifier
LOCATION_CD	Store Identifier
NET_SLS_AMT	Net Sales Value
POS_TP_CD	Identifier of Phisical or Online Transaction
PROD_DSCNT_ISSUED_AMT	Product Discount Value Associated to a Client Card
QTY	Quantity of Product
SKU	Product Identifier
TIME_KEY	Date of the Transaction
TRANS_DSCNT_RATT_AMT	Transaction Discount Value Associated to a Client Card
TRANSACTION_ID_MASK	Transaction Masked Identifier

Table 9 - A1: Variable Description of Transaction (TRX) Table

Variable	Description
SKU	Product Identifier
PRODUCT_DSC	Product Description
UNIT_BASE_CD_EXT	Product Base Code
UNIT_BASE_DSC_EXT	Product Base Description
SUBCAT_CD_EXT	Product Subcategory Code
SUBCAT_DSC_EXT	Product Subcategory Description
CAT_CD_EXT	Product Category Code
CAT_DSC_EXT	Product Category Description
BIZ_UNIT_CD_EXT	Business Unit Code
BIZ_UNIT_DSC_EXT	Business Unit Description
DEPARTMENT_CD_EXT	Department Code
DEPARTMENT_DSC_EXT	Department Description
PRODUCT_SHORT_DSC	Product Short Description
BRAND_DSC	Brand Description
BRAND_TYPE_CD	Brand Type
PRICE_RANGE	Price Segment
CONVERSION_FACTOR	Measurement
CAPACITY_UNIT	Unit of Measurement

Table 10 - A2: Variable Description of Product (SKU) Table

Variable	Description
CUSTOMER_ACCOUNT_NR_MASK	Customer Masked Identifier
GENDER	Gender
FAMILY_MEMBERS	Number of Family Members
CP4	Postal Code
SEG_LIFESTYLE_CD	Lifestyle Segmend Code
SEG_LIFESTYLE_DSC	Lifestyle Segment Description
SEG_AGE	Age Segment Code
SEG_AGE_DSC	Age Segment Description
SEG_LIFESTAGE_CD	Lifestage Segment Code
SEG_LIFESTAGE_DSC	Lifestage Segment Description

Table 11 - A3: Variable Description of Customer Table

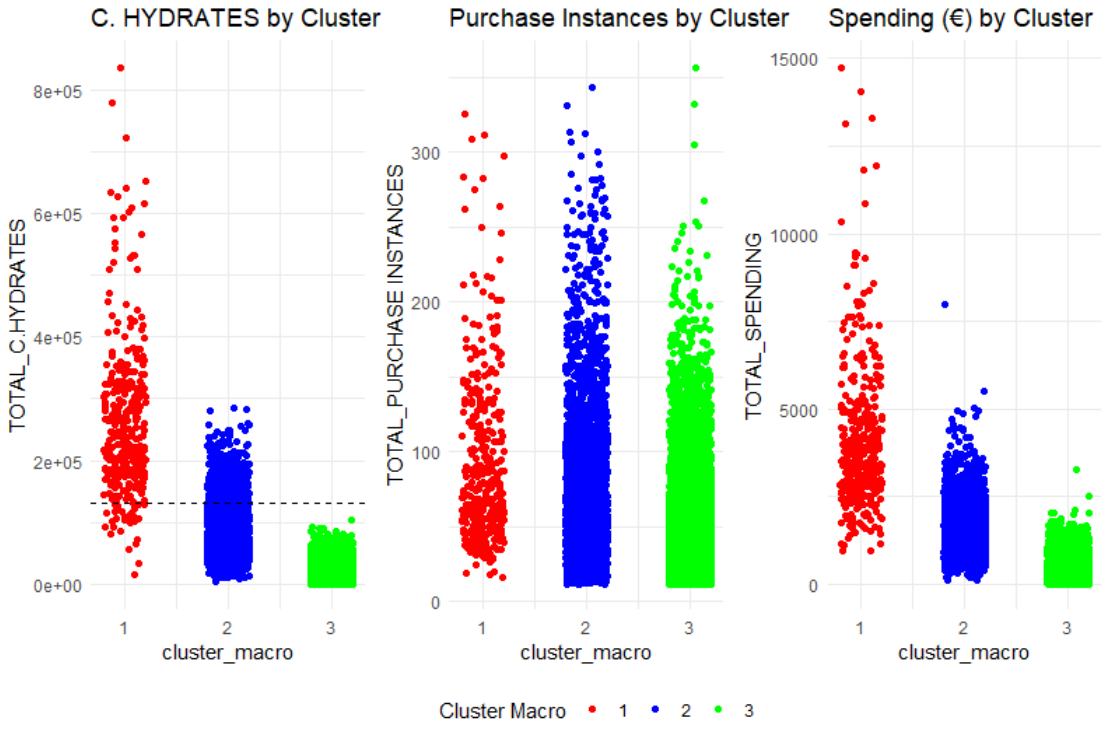

Figure 12 - A4: Macronutrient Clustering Analysis by Carbohydrates Intake

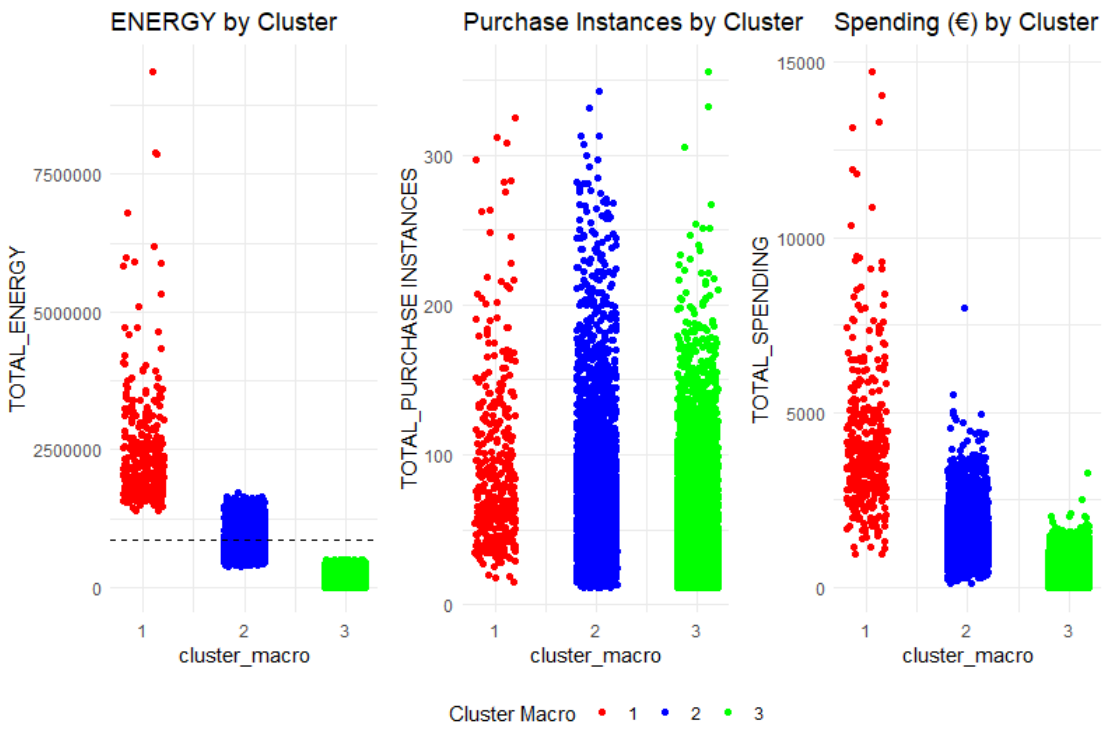

Figure 13 - A5: Macronutrient Clustering Analysis by Energy Intake

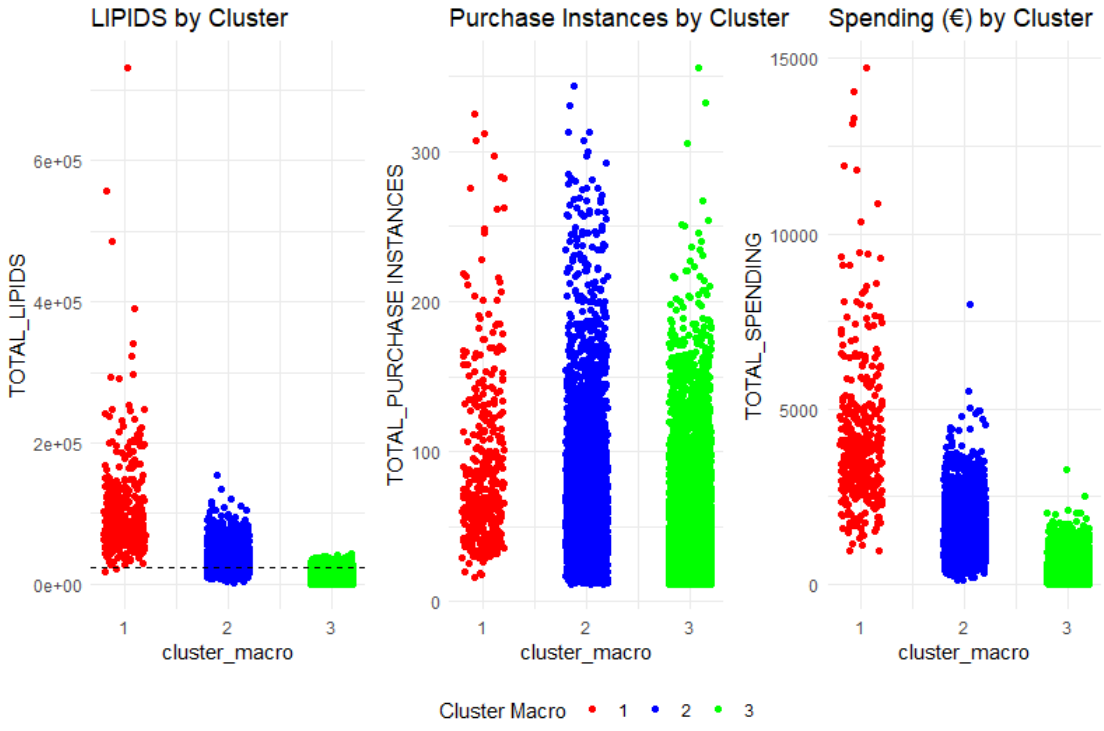

Figure 14 - A6: Macronutrient Clustering Analysis by Lipids Intake

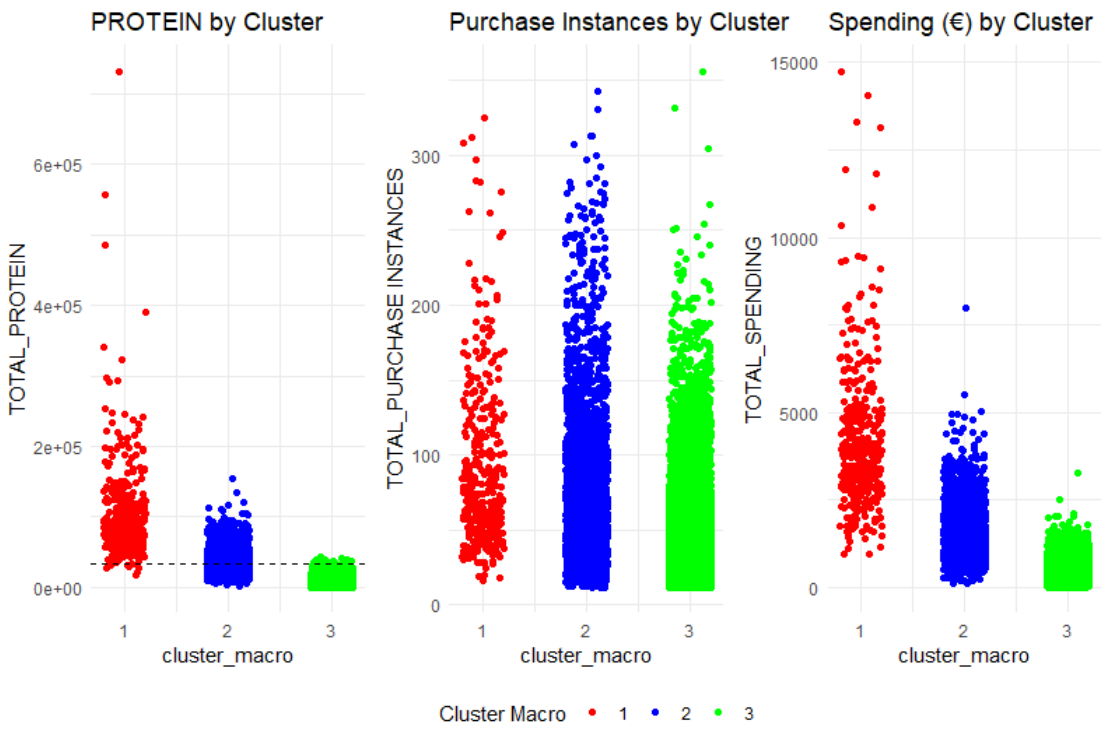

Figure 15 - A7: Macronutrient Clustering Analysis by Protein Intake

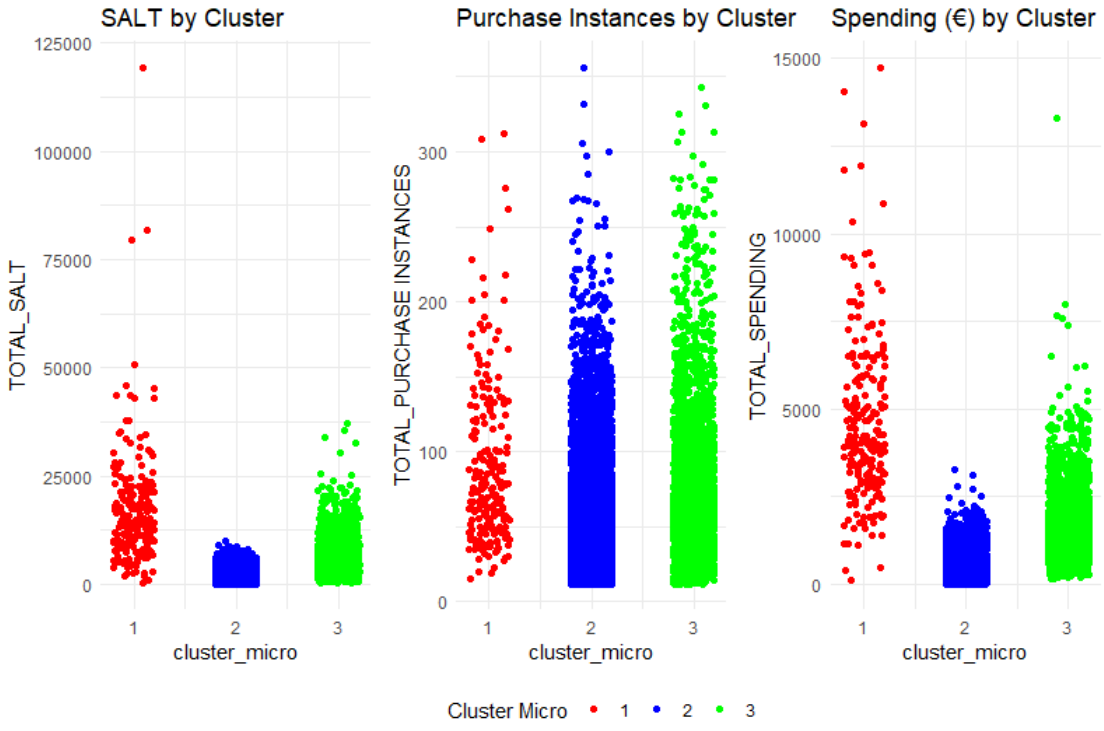

Figure 16 - A8: Micronutrient Clustering Analysis by Salt Intake

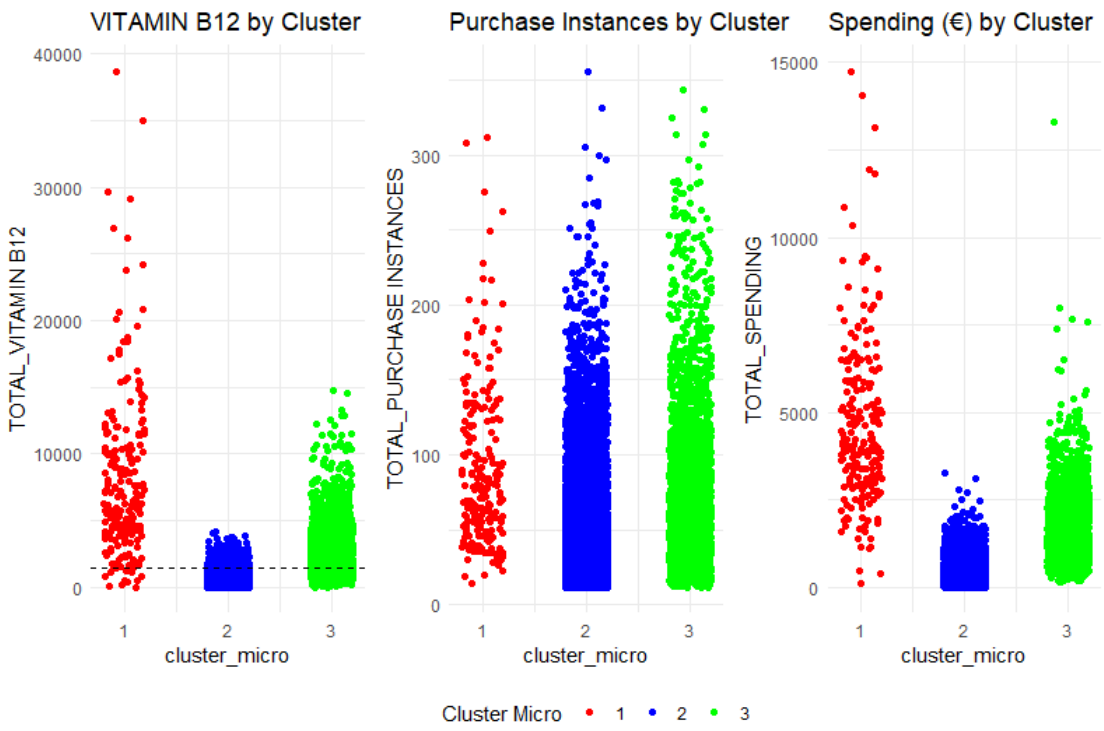

Figure 17 - A9: Micronutrient Clustering Analysis by Vitamin B12 Intake

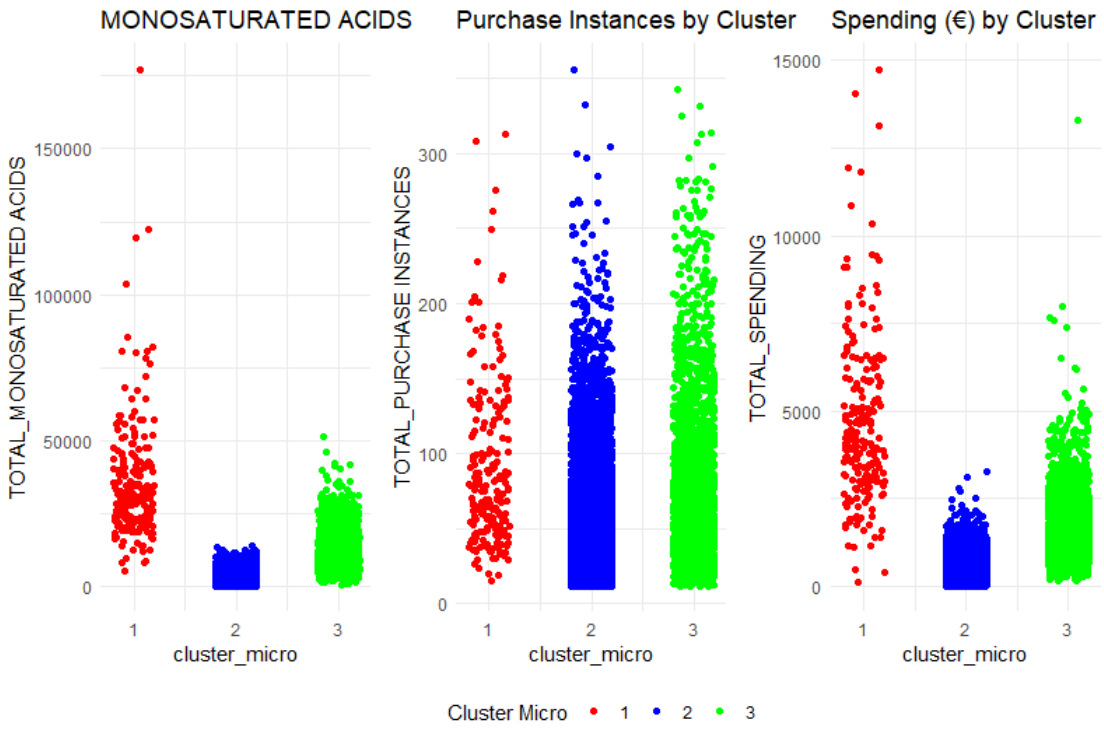

Figure 18 - A10: Micronutrient Clustering Analysis by Monosaturated Acids Intake

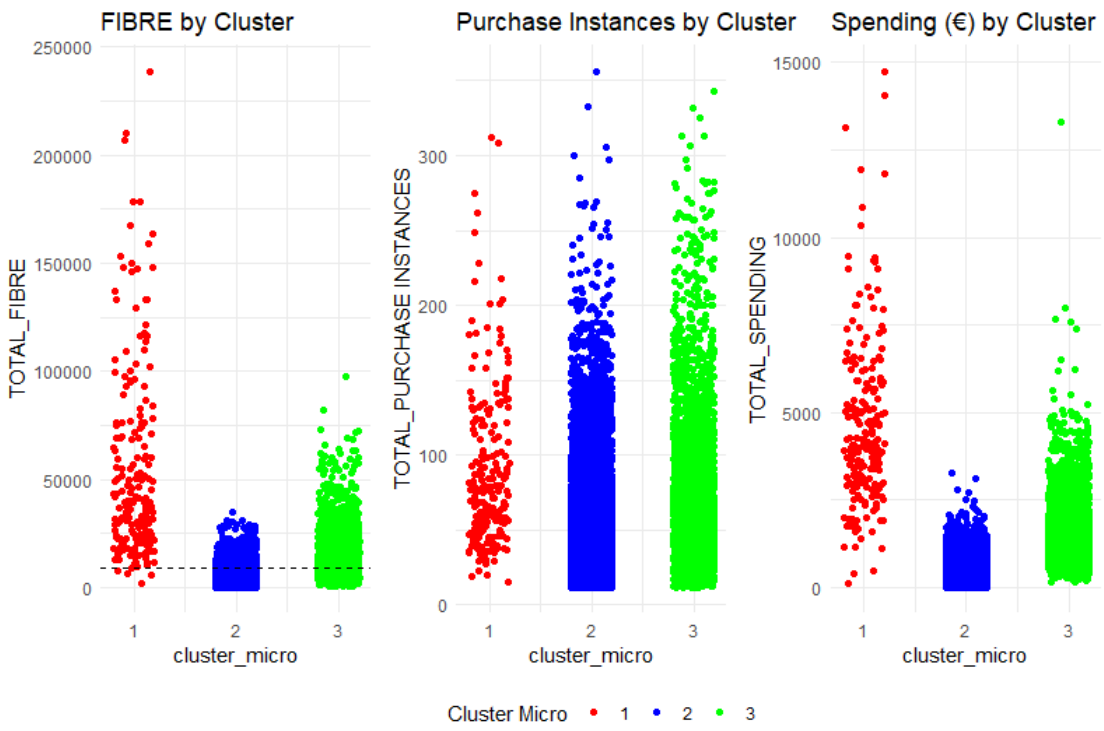

Figure 19 - A11: Micronutrient Clustering Analysis by Fibre Intake

Figure 20 - A12: Environmental Impact Clustering Analysis by Water

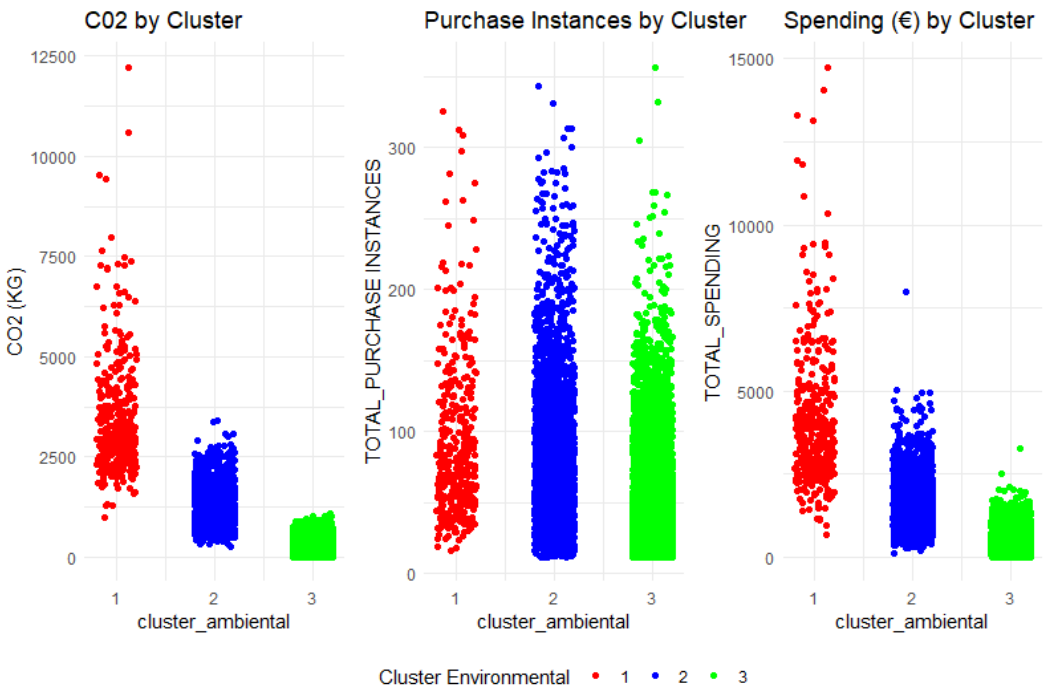

Figure 21 - A13: Environmental Impact Clustering Analysis by CO2

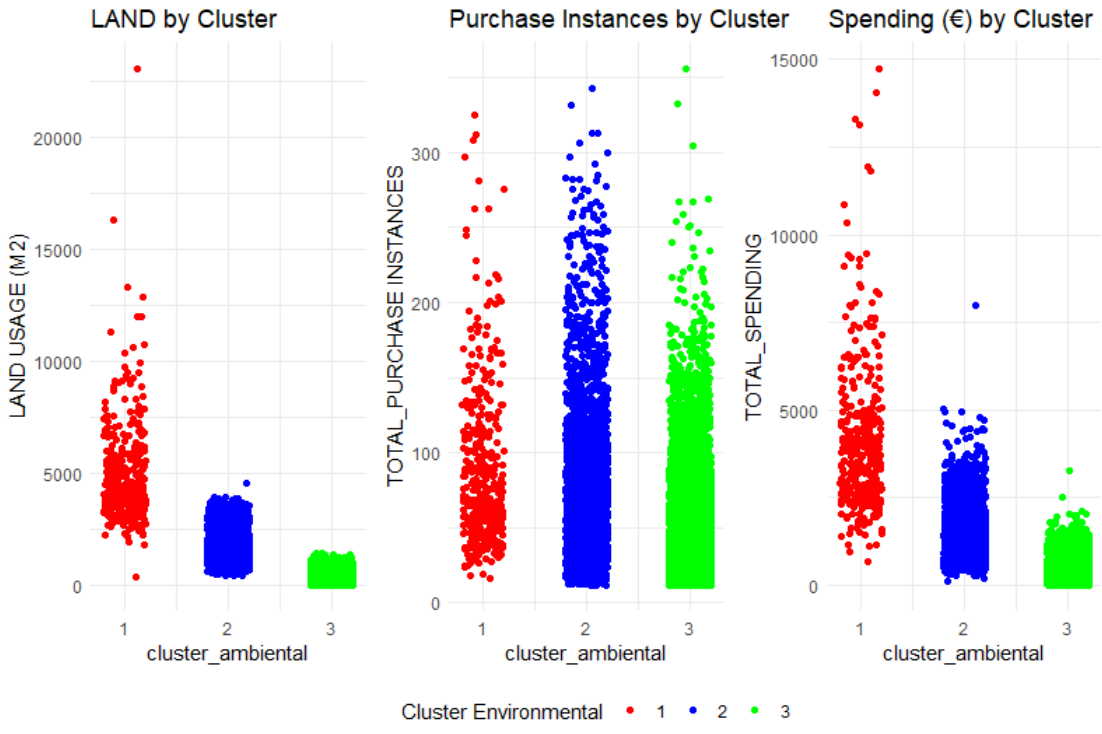

Figure 22 - A14: Environmental Impact Clustering Analysis by Land Usage