UNIVERSIDADE CATÓLICA PORTUGUESA

A Comparison Between Statistical and Machine Learning Methods for Retail Market

Filipe Simões

Católica Porto Business School
2024

UNIVERSIDADE CATÓLICA PORTUGUESA

A Comparison Between Statistical and Machine Learning Methods for Retail Market Segmentation

Master Final Thesis
presented to Universidade Católica Portuguesa
to obtain a master's degree in management with a specialization in Business Analytics

by,

Filipe Simões

under the guidance of
Professor Aydin Teymourifar

Católica Porto Business School
March 24

Acknowledgments

Embarking on this academic journey has been a profound experience, and I owe a debt of gratitude to those whose support and guidance have been instrumental.

First, I state my deepest appreciation to my supervisor, Aydin Teymourifar. His unwavering support, expert guidance, and commitment to my academic growth have been the highlight of this investigation. I am grateful for the mentorship, encouragement, and the invaluable insights that shaped this thesis.

Family plays an indelible role in one's academic pursuits, and I express profound gratitude to my parents and brother. Their continuous support, understanding, and encouragement have been a constant source of strength and compromise during my both professional and academic moments.

To my circle of friends, thank you for your unwavering confidence and belief in my capabilities. Your camaraderie and encouragement have been vital in sustaining my spirit and motivation during this transformative period.

This acknowledgment is a testament to the collaborative efforts and encouragement I have received from these exceptional individuals. Each has played an integral role in shaping this academic milestone, and I am profoundly grateful for their contributions to this journey.

Agradecimentos

Embarcar nesta experiência académica tem sido uma experiência profunda, e devo um grande agradecimento àqueles cujo apoio e orientação foram fundamentais.

Em primeiro lugar, expresso o meu mais profundo apreço ao meu orientador, Aydin Teymourifar. O seu apoio inabalável, orientação especializada e compromisso com o meu crescimento académico têm sido o ponto alto desta investigação. Estou grato pelo encorajamento e pelos valiosos insights que moldaram esta tese.

A família desempenha um papel extraordinário nas aspirações académicas de alguém, e expresso profunda gratidão aos meus pais e irmão. O seu apoio contínuo, compreensão e encorajamento têm sido uma fonte constante de força e apoio durante os meus momentos profissionais e académicos.

Aos meus amigos, obrigado pela vossa confiança inabalável e crença nas minhas capacidades. A vossa camaradagem e encorajamento têm sido vitais para manter o meu espírito e motivação durante este período transformador.

Este reconhecimento é um testemunho dos esforços colaborativos e encorajamento que recebi destes indivíduos excecionais. Cada um desempenhou um papel integral na formação deste marco académico, e estou profundamente grato pelas suas contribuições para esta jornada.

Abstract

This study explores the dynamics of consumer behaviour in the retail sector to analyze the varied inclinations, patterns of purchase, and frameworks for making decisions among different groups of people. The study has identified six unique consumer clusters based on a thorough two different methodologies (k-means and Self-Organizing Maps ("SOMs")). Every cluster exhibits a unique profile that demonstrates the intricacy and subtlety of customer preferences. The results show notable differences in brand loyalty, price sensitivity, and the importance of sustainability, with ramifications for customized marketing strategies.

These findings have significant ramifications for retail enterprises. They emphasize how important it is for retailers to tailor their operational and marketing strategies to meet the diverse demands and wants of various consumer segments. It is recommended that retailers make use of these insights to augment customer interaction, refine inventory control, and formulate sophisticated sustainability programs that correspond with consumer preferences.

In addition, the study set a direction for more research. Researching how consumer preferences vary over time and evaluating the efficacy of various marketing strategies across various demographic groups, and would be impactful in the realm. A global viewpoint and an enlarged scope would also highlight the impact of culture on buying decisions, offering a thorough understanding of consumers in various economies.

Keywords: Self-Organizing Maps, Market Segmentation, Clustering, Retail, Decision-Making

Resumo

Este estudo explora a dinâmica do comportamento do consumidor no setor retalhista para analisar as diversas inclinações, padrões de compra e estruturas de tomada de decisão entre diferentes grupos de pessoas. O estudo identificou seis clusters de consumidores únicos com base em duas metodologias distintas (k-means e Mapas Auto-Organizados ("SOMs")). Cada cluster apresenta um perfil único que demonstra a complexidade e subtileza das preferências dos clientes. Os resultados mostram diferenças notáveis na fidelidade à marca, sensibilidade ao preço e importância da sustentabilidade, com repercussões para estratégias de marketing personalizadas.

Estas descobertas têm ramificações significativas para as empresas de retalho. Elas enfatizam quão importante é para os retalhistas adaptar as suas estratégias operacionais e de marketing para satisfazer as diversas exigências e desejos de diferentes segmentos de consumidores. Recomenda-se que os retalhistas aproveitem estas perceções para aumentar a interação com os clientes, refinar o controlo de inventário e formular programas de sustentabilidade sofisticados que correspondam às preferências dos consumidores.

Além disso, o estudo estabeleceu uma direção para mais investigação. Investigar como as preferências dos consumidores variam ao longo do tempo e avaliar a eficácia de várias estratégias de marketing em diferentes grupos demográficos seria impactante no campo. Uma perspetiva global e um escopo alargado também destacariam o impacto da cultura nas decisões de compra, oferecendo uma compreensão completa dos consumidores em diversas economias.

Palavras-chave: Mapas Auto-Organizados, Segmentação de Mercado, Clustering, Retalho, Tomada de Decisão

Index

Number of words: 9206

Figure Index

Table Index

Abbreviations

EFA:	Exploratory Factor Analysis
IoT:	Internet of Things
LFDA:	Local Fisher Discriminant Analysis
NN:	Neural Network
ML:	Machine Learning
PCA:	Principal Component Analysis
SOFM:	Self-Organizing Feature Map
SOMs:	Self-Organizing Maps
SVC:	Support Vector Clustering

Introduction

Mark segmentation is an essential technique in the business realm that allows companies the ability to comprehend their target markets more profoundly and tailor their offerings to meet customer needs efficiently. Over the years, extensive research has delved into various facets of market segmentation, encompassing its conceptualization, segmentation processes, methodological advancements, and the challenges inherent in implementation.

At the core of informed decision-making and strategic planning, particularly within the volatile landscape of the retail industry, lies successful market segmentation. This practice, which involves categorizing diverse markets into discrete groups based on shared criteria, proves indispensable for understanding the complexities inherent in such multifaceted environments.

The retail sector is characterized by its intricate ecosystem and complex structure. Traditional statistical methods have played a pivotal role in deciphering the complexities of retail markets, offering interpretability and a robust theoretical foundation. However, the dynamic nature of the industry calls for innovative approaches to identify and understand the subtle nuances that conventional statistical tools might overlook. In this evolving scenario, the field of machine learning emerges as a compelling domain for exploration. With its ability to recognize and predict complex patterns in consumer behaviour, machine learning presents a forward-thinking method to navigate the intricacies of the retail ecosystem.

This research aims to bridge the divide between traditional statistical methodologies and modern machine learning techniques, providing an

exhaustive evaluation of their respective strengths and limitations. It seeks to unearth the hidden structures within retail markets by employing a diverse range of analytical tools. This includes traditional methods like K-means and advanced machine learning techniques, notably neural networks such as SOMs, which are renowned for their ability to capture both the distribution and topology of data, placing them at the forefront of market segmentation research. Their grouping and visualization capabilities render them a promising tool for this purpose.

By examining the outcomes of these varied methodologies, this study endeavours to offer substantial insights into the comparative effectiveness of statistical and machine learning approaches in market segmentation. These insights aim to assist practitioners, researchers, and market stakeholders in making well-informed decisions regarding segmentation strategies, thus enhancing their understanding and application of these techniques in the dynamic world of retail.

1.1 General Framework

Market segmentation is important due to its allowance for better resource allocation and marketing activities. Businesses may better engage their target consumers by recognizing and understanding their distinct requirements and preferences. This tailored strategy increases customer happiness and loyalty, resulting in a more long-lasting relationship between consumers and brands. Market segmentation helps firms identify niche markets, allowing them to focus on certain sectors that match their strengths and skills. This specific focus enables more efficient resource allocation and better usage of marketing dollars, resulting in a higher return on investment.

Advanced analytical approaches, such as clustering algorithms and artificial intelligence, are commonly used in market segmentation. These methods allow

for more comprehensive analyses of massive datasets, revealing underlying patterns and linkages that typical segmentation algorithms may miss. This improves the precision and accuracy of market segmentation, enabling businesses to make better strategic decisions.

In conclusion, market segmentation is a dynamic and critical component of modern marketing strategy. Its advantages include better resource allocation, more consumer satisfaction, and the creation of tailored marketing campaigns. As organizations traverse an increasingly competitive marketplace, proper market segmentation remains critical for attaining long-term growth and preserving a competitive advantage in the market.

1.2 Objectives and Research Methodology

The key goal is to investigate the practical uses of various clustering approaches and evaluate the influence of AI alternatives on market segmentation in the retail industry. We aim to provide managerial insights that benefit both large and small firms.

Our goal is to identify the most exact and straightforward approach. The success and popularity of AI technologies prompted us to investigate the Neural Network (NN) approach.

The thesis consists of three steps. We begin with an exhaustive review of the literature. After that, we use the selected strategies in the database gathered. The final phase is to extract outputs and draw significant conclusions from the investigation.

1.3 Macrostructure

This dissertation is divided into six chapters that are all cohesive. Chapter 1 describes the research topic, aims, methodology, and overall structure. In Chapter 2, a thorough literature analysis explores existing scholarship, finding breakthroughs and knowledge gaps.

Chapter 3 outlines the study problem and frames important questions, serving as a basic guide. Chapter 4 describes the approaches used, including inputs, outputs, and performance measurements, and provides insights into the chosen strategies.

The fifth chapter focuses on experimental outcomes, beginning with dataset characterization and progressing to descriptive analysis and findings discussion. Chapter 6 summarizes major findings and serves as a synopsis of the study.

The last section includes handpicked bibliographic sources that support the thesis's arguments, concepts, and theoretical foundations.

Literature Review

2.1 Market Segmentation

Mussa, Rosen (1978), and Moorthy (1984) investigated the optimal pricing strategies for independent products differentiated by quality, in markets defined by heterogeneous consumers whose quality perception may not be the same. Debo et al. (2005), on the other hand, extend this research by considering the simultaneous determination of product prices and manufacturing techniques in the context of a generic customer profile for manufacturable products. Their research attempts to supplement Ferrer's (2000) findings by investigating the conditions under which his cost savings analysis alone is sufficient to determine the viability of remanufacturing. This recurrent study presents the definition of market segmentation as the strategy of dividing a heterogeneous group into different segments based on their characteristics.

Linked to the retail market, Allred et al. (2006) investigate market segmentation in the context of e-commerce, revealing unique segments such as "e-shopping lovers" and "fearful conservatives" (Allred et al., 2006). This study demonstrates how recognizing and responding to the distinct requirements and preferences of different consumer categories may have a substantial impact on retail enterprise success.

Market segmentation is a fundamental idea in retail, playing a critical role in adapting marketing strategies, and customer-centric perspectives with the new era of data.

2.2 Statistical and Machine Learning Methods

2.2.1 Statistical Methods

Statistical approaches have a long history in financial analysis and market research, making them indispensable instruments for researching market segmentation. These methods make use of mathematical models and historical data to provide insight into market behaviour and investor preferences. Several statistical methods are especially pertinent to the research.

According to Christoph Ley et al., (2022), statistical methods are primarily a top-down technique: there is the assumption of the model from which the data was generated (i.e., techniques such as linear and logistic regression), and then estimate the model's unknown parameters from the data. In other words, it is expected that we understand how input variables relate to output, which simplifies the interpretation of results and makes the relationships between variables clear. The possible disadvantage is that the relationship between input and output is user-defined, which may result in a suboptimal (i.e., less accurate) prediction model if the chosen model does not effectively reflect the real input-output correlation.

This can happen if a user selects linear regression yet the relationship between input and output is non-linear, or when several input variables are involved.

2.2.1.1 Factor analysis.

Factor analysis is a statistical technique commonly used in social sciences to uncover underlying latent constructs, or factors, that explain patterns of correlations between a set of observed variables. It is extremely useful for lowering data dimensionality while retaining critical information, which aids in understanding complicated relationships within the data (Fabrigar et al., 1999).

This approach can contribute to the body of knowledge on retail market segmentation through a global review of relevant studies and the application of

factor analysis techniques, providing valuable insights into the factors that influence decisions.

Costello and Osborne (2005) published a paper outlining best practices in exploratory factor analysis (EFA) to ensure optimal results and analysis interpretation. The authors make four key recommendations for EFA researchers. First, they recommend assessing the suitability of data for factor analysis, including sample size and sampling adequacy. Second, they emphasize the importance of selecting an appropriate factor extraction method based on the characteristics of the data, such as principal components analysis or maximum likelihood estimation. Third, they advise examining the factor structure with various criteria such as eigenvalues, scree plots, and parallel analysis to determine the number of factors to keep. Finally, they emphasize the importance of interpreting and labeling the factors by substantive theory or practical considerations.

In conclusion, the main goal and difference lies in the objective of the study. Factor analysis seeks to identify factors that can explain the variable correlations found in the results.

2.2.1.2 Discriminant Analysis

Discriminant analysis, pioneered by Fisher in 1936, stands out as a widely utilized technique for linear supervised dimensionality reduction. It seeks an embedding transformation such that the between-class scatter is maximized, and the within-class scatter is minimized. FDA is a traditional but useful method for dimensionality reduction. However, it tends to give undesired results if samples in a class form several separate clusters (i.e., multimodal).

In addition to these ideas, multimodal problems urge in several contexts in the modern days which bring the idea of Local Fisher Discriminant Analysis (LFDA) by Sugiyama (2007) where the LFDA maximizes between-class separability and

preserves within-class local structure at the same time, being useful for multimodal problems.

As described by Mika et al. (1999), discriminant analysis is a conventional statistical technique, that has been limited to datasets that can be separated linearly. It is used to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.

2.2.1.3 Logistic regression

Logistic regression, as first proposed by Berkson A1944), is an important statistical tool used in several areas. This method displays its effectiveness in circumstances when the outcome of interest is binary, representing two categories. Joseph Berkson's work, lays the groundwork for logistic regression's use, demonstrating its effectiveness in evaluating binary sequences and showing intricate patterns within datasets.

David R. Cox (1958) refined and extended the logistic regression paradigm. Cox's contributions stress the statistical intricacies involved in assessing binary outcomes, providing a solid foundation for regression analysis on binary data.

Logistic regression is a versatile technique, especially for instances where the dependent variable is binary, such as success/failure, yes/no, or presence/absence. Its applications range from medical to economics and social sciences, all of which have binary consequences.

Engel (1988) presents the notion of “polytomous” logistic regression, which broadens the scope of logistic regression. Unlike its binary predecessor, multinomial logistic regression can manage cases with more than two discrete outcome categories. This study, develops into the complexities of managing several categories, providing a useful expansion to logistic regression for scenarios with more diverse outcomes.

The primary distinction is that multinomial logistic regression can handle dependent variables with different categories, making it an appropriate choice when the outcome variable has more than two options. Engel's work makes a substantial contribution to the understanding and use of multinomial logistic regression, expanding the analytical toolkit for scenarios requiring the modeling of categorical outcomes at several levels.

2.2.2 Machine learning methods

Machine learning technologies have gotten a lot of attention because of their capacity to extract important insights from data and automate complex operations. These methodologies provide a variety of tools for market segmentation in the context of retail markets, allowing for more complex categorization of consumers based on different criteria. The allure of machine learning resides in its capacity to rapidly examine large information, revealing complex patterns that traditional analytical methods may miss. In retail, this entails developing a thorough grasp of the qualities, and preferences of customers. This granular methodology allows for the construction of more accurate and focused segments that transcend broad classifications to capture varied consumer motives and purchasing habits.

Machine learning, in essence, enables dynamic segmentation by adjusting to the changing dynamics of retail markets over time. Because of this versatility, segmentation models remain relevant and sensitive to the ever-changing retail sector, delivering useful information for targeted marketing, service optimization, and revenue management strategies.

2.2.2.1 Cluster Analysis

Cluster analysis is a key technique in data mining, and it is an essential component of this study's methodological framework. Cluster analysis, as a dimension reduction approach, is critical for understanding complex data

structures and detecting meaningful patterns, both of which are required for this study.

Based on established factors, cluster analysis identifies groups or clusters of comparable data points. K-means clustering (Hartigan & Wong, 1979) and hierarchical clustering (Everitt et al., 2011) are the two main notable approaches. A case study on market segmentation using the R programming language and the K-means clustering algorithm is presented by Hung et al. (2019). Based on customer data, the study seeks to categorize and describe various market segments. The authors successfully divide the customer base into homogeneous groups using K-means clustering and offer details on the preferences and behaviours of each segment. The results highlight the usefulness of K-means clustering in understanding customer segmentation for targeted marketing strategies and show how well it performs in market segmentation.

2.2.2.2 SOMs

Unsupervised learning approaches have shown useful in market segmentation, especially when working with large and complicated datasets. These methods, which include clustering algorithms, seek to detect natural groupings or patterns in data without requiring explicitly labelled data. Unsupervised learning is an effective method for recognizing the intrinsic structures and relationships between data pieces.

Huang and Tzeng (2007) explore the application of support vector clustering (SVC) for marketing segmentation and compare it with other clustering methods such as k-means and self-organizing feature maps (SOFM). The study demonstrates the effectiveness of SVC through a case study of a drink company. SVC overcomes challenges like data type restrictions, sensitivity to outliers, high-dimensional data, and arbitrary cluster shapes. The paper concludes that SVC is a suitable tool for marketing segmentation and presents a hybrid segmentation

algorithm for hotel managers to identify customer segments and tailor marketing strategies accordingly.

SOMs are a subset of unsupervised learning that has shown to be an effective tool for market segmentation and data presentation. After being introduced in the 1980s, they have subsequently been used in a wide range of fields, including banking and market research (Kohonen, 1982).

SOMs work by arranging a grid of artificial neurons in a 2D lattice, with each neuron representing a prototype or cluster centre. During training, SOMs learn to map high-dimensional input data onto this grid while keeping data point associations. This topographic mapping allows for the discovery of underlying data structures and patterns, which is critical for market segmentation (Kohonen, 1990).

SOMs have been used successfully in financial and market research. In one work, an iterated greedy heuristic including SOMs was used to solve a market segmentation problem with various features.

2.2.2.3 Neural Networks

Supervised learning is a fundamental approach in machine learning. To learn the mapping function, this paradigm includes training a model on a labelled dataset that contains both inputs and outputs.

In the field of retail market segmentation, neural networks have emerged as a potent tool. These models, inspired by the neural architecture of the human brain, have the potential to learn complicated correlations among data (Kohonen, 1988). Understanding how neural networks learn and the possible challenges they provide is critical to our research's success.

Mitchell's key study (1997) presents an in-depth examination of guided learning. It highlights the value of data in training models that can make predictions or judgments without the need for explicit programming. According

to Mitchell, supervised learning is critical for tasks in which the algorithm must learn patterns and relationships from prior data to anticipate outcomes for future occurrences.

Schmidhuber (2015) presents an in-depth look into deep learning in neural networks. Deep learning expands the ideas of supervised learning by incorporating deep neural networks, which have several layers. This architectural breakthrough enables the model to learn hierarchical data representations automatically, capturing complicated patterns and connections. Machine learning-based forecasting mechanisms have proved their significance in anticipating perioperative outcomes, hence improving decision-making for future courses of action, according to Rustam Furqan et. al (2020) investigations. These models, which are widely used in a variety of application areas for recognizing and prioritizing negative aspects, play an important role in threat assessment. This work demonstrates the effectiveness of ML models in projecting the amount of incoming COVID-19 cases, a current and significant hazard to public health. Furthermore, the results of the study canbe directly related to the importance of supervised models that can be useful to the main concept of this paper.

2.3 Application to Retail Markets

This modern approach to consumer segmentation is perfectly aligned with the integration of statistical and machine-learning methodologies for clustering in the retail sector. These advanced techniques enable retailers to extract meaningful patterns from massive and complex datasets, revealing previously unknown insights into consumer preferences.

D. Brown et al. (2018) demonstrated the effectiveness of behavioural segmentation in uncovering nuanced patterns in consumer preferences, particularly in understanding differences in purchasing behaviours across different customer segments. Machine learning algorithms that use historical

data improve this understanding even further, allowing for the development of personalized marketing strategies tailored to specific product categories or consumer segments.

Furthermore, the research by G. Davis et al. (2020) offers a comprehensive examination of the complex and dynamic nature of consumer expectations in the retail sector. The authors highlight the multidimensional and multifaceted aspects of consumer expectations that are influenced by a variety of factors. This acknowledgment is consistent with the study's central premise, which emphasizes the use of advanced statistical and machine learning methods to decode detailed patterns in consumer actions.

According to the findings of the literature review conducted by E. Garcia et al. (2022), statistical and machine learning methods present a promising avenue for improving the precision and utility of clustering in the retail industry. By incorporating these advanced techniques, retailers can gain a more sophisticated understanding of consumer behaviour, optimize pricing strategies, and ultimately improve the overall consumer experience.

In conclusion, the evolution of consumer segmentation, smoothly integrating with statistical and machine learning methodologies for clustering in the retail sector, represents a significant advancement in understanding and responding to consumer interactions.

2.4 Innovative Methods of Data Collection in Retail

There are various trends in data collection innovations for retailing, with Internet of Things (IoT) serving as a prominent example. The convergence of the IoT and the age of data represents a transformative force in retail market segmentation. As elucidated by Kakhi et al. (2022), the integration of IoT technologies and artificial intelligence is driving a paradigm shift, presenting both challenges and opportunities. The seamless connectivity of devices and the prolific generation of data in the retail environment offer an unprecedented depth of insights into consumer behaviours.

In the landscape of retail, the study by Allred et al. (2006) highlights the significance of market segmentation for understanding diverse consumer preferences. The IoT contributes to this understanding by enabling the continuous collection of real-time data from various touchpoints. This influx of data allows retailers to move beyond traditional segmentation based on demographic or geographic variables, as discussed by Allred et al. (2006), and delve into dynamic and personalized segmentation strategies.

The IoT's impact lies not only in its ability to capture consumer preferences but also in the potential for real-time adaptation. By harnessing data from connected devices, retailers can respond promptly to shifting consumer behaviours, optimizing segmentation strategies on the fly. This responsiveness aligns with the findings of Allred et al. (2006), emphasizing the need for retailers to adapt their strategies to meet evolving consumer expectations.

In summary, the IoT and the era of data, as explored by Kakhi et al. (2022), have the potential to revolutionize market segmentation in retail. This transformative impact stems from the continuous flow of real-time data, providing retailers with the tools to create more nuanced, dynamic, and personalized segmentation strategies that align closely with the ever-changing landscape of consumer preferences.

Problem Definition and Research Questions

3.1 Problem Definition

Developing specialized marketing strategies is made more difficult by the peculiarities of the retail industry, which are defined by a wide range of consumer behaviours and a constantly changing environment. To successfully personalize retail tactics, this heterogeneous market must be efficiently classified. The challenge is unfortunately made more complex by the lack of understanding regarding the relative merits of statistical and machine learning approaches in retail market segmentation.

The underlying issue stems from an unclear understanding of which segmentation strategies—based on statistical or machine learning frameworks—produce the best results for detecting consumer attributes. One major challenge is the lack of clarity on the trade-offs between interpretability, scalability, and accuracy in these systems. Furthermore, little research has been done on the real-world effects of segmentation applications on managerial choice-making in the setting of retail marketplaces.

By performing a thorough evaluation of statistical and machine learning techniques in the field of retail market segmentation, this study seeks to address this important topic. The lack of clear standards for this kind of comparison and a poor comprehension of the managerial fallout from various segmentation strategies draw attention to important gaps in the state of the art. Closing these gaps will be essential to improving retail decision-making processes and optimizing tactics in the constantly changing retail offerings landscape.

3.2 Research Questions

The research questions that are addressed in this dissertation are presented in the following table.

Research Question	Objective
In retail market segmentation, how do statistical and machine learning approaches compare?	Conduct a thorough review of statistical and machine learning methodologies in the context of retail market segmentation to determine their respective strengths, shortcomings, and practical consequences.
What are the details of consumer habits that statistical approaches may overlook?	Determine and examine minute details in consumer behaviour that conventional statistical tools could miss, highlighting the necessity for creative solutions in the changing retail environment.
How do machine learning techniques contribute to dynamic retail market segmentation?	Examine how machine learning can adapt to shifts in consumer habits to help merchants find the best segmentation tactics.
What is the impact of statistical and machine learning methods on marketing and operational strategies in retail?	Analyze how these innovative methods affect pricing tactics to improve retail pricing techniques for increased customer interaction.

Table 1- Research questions and objectives

Research Methods

4.1 Methods Definition

The selection of a methodological paradigm is crucial for steering the research process, determining the tools we employ, the data we gather, and how we interpret the outcomes. In the field of retail market segmentation, using a methodological paradigm can help to understand and improve service delivery resulting in better results for clients.

The current study is based on a methodological paradigm that employs a quantitative approach, including a survey. This technique tries to provide objective measurements that can be used to better understand and enhance retail market segmentation quality.

To assess whether there is a trend in customers' preferences for low-cost/luxury services and their connection to wealth status, we asked where the customers make their purchases to understand what type of service they address. Similar questions were asked to determine how comfortable patients are with digital markets and with which types of products can extrapolate that kind of purchase.

To better understand what drives customers' decisions and, as a result, what can encourage their loyalty, we asked respondents to rate several characteristics including quality, waiting time, price, and distance in terms of their influence on their choice of market.

By designing a new questionnaire, our study enhances the depth and comprehensiveness of understanding the factors influencing patient quality, loyalty, and satisfaction.

4.1.1 Data collection and measurement instruments

This research focuses on the diverse retail market environment. To collect as much information as possible, there was made available online survey through a variety of digital media, including Instagram, Facebook, and LinkedIn. The survey was available on these platforms between December 2023 and January 2024.

One of the most significant considerations when conducting a survey or study is sample size, as it affects the validity and precision of your findings. The number of replicates or observations in our statistical sample is referred to as the sample size.

Furthermore, the task of establishing the ideal number of clusters is handled with strategic thinking. The Elbow Method, a well-known technique, is used to identify the point at which adding more clusters no longer significantly improves the segmentation model.

In parallel, a trial-and-error method is taken, testing with various cluster configurations. This iterative procedure includes repeating segmentation studies with changing cluster counts and evaluating their effectiveness on the data set. The goal is to find a compromise, ensuring that the segmentation model accurately captured the subtle patterns present in the diversified retail market environment.

.

Experimental Results

5.1 Dataset characterization

Our database is compiled using a questionnaire available on numerous web platforms. There are 23 total questions. A total of 101 responses were gathered, and their relationships were investigated and assessed. The questions were designed to understand each financial, social, and behavioural context. To improve the reliability of the results, the questionnaire is primarily limited to multiple-choice items.

5.2 Descriptive analysis

In this subchapter, we will look at the overall properties of the dataset and undertake a descriptive analysis to get a better understanding of the data's original state.

When we look at the dataset, we see replies from people of all ages, with a noticeable preponderance of young adults. The distribution is the outcome of the questionnaire's sharing strategies, such as social feeds and company networking. It's important to remember that most responders can make economic decisions.

Looking at gender distribution, we see an equitable representation that avoids bias in any output. Notably, responses are from a variety of cities, which adds great value to the research. Porto stands out as the hub for the major network with almost 60% of the answers. As a result, the household income distribution is naturally broad, reflecting differences in average incomes among countries.

In line with the global trend of higher educational attainment, around 80% of respondents have at least a bachelor's degree as their highest level of education.

Regarding the decision-making questions, we focused on two key factors: fresh and non-fresh products.

When it comes to fresh products, a large majority of people say they are "somewhat uncomfortable" making purchases online. Quality, price, and proximity to work/home, respectively, emerge as the most important factors in their purchasing decisions.

On the other hand, when it comes to non-fresh products, 50% of respondents say they feel "very comfortable" shopping online. Similarly, quality, price, and proximity to home/work are the primary criteria that influence their decision-making. Figures 1–10 summarize the descriptive analysis.

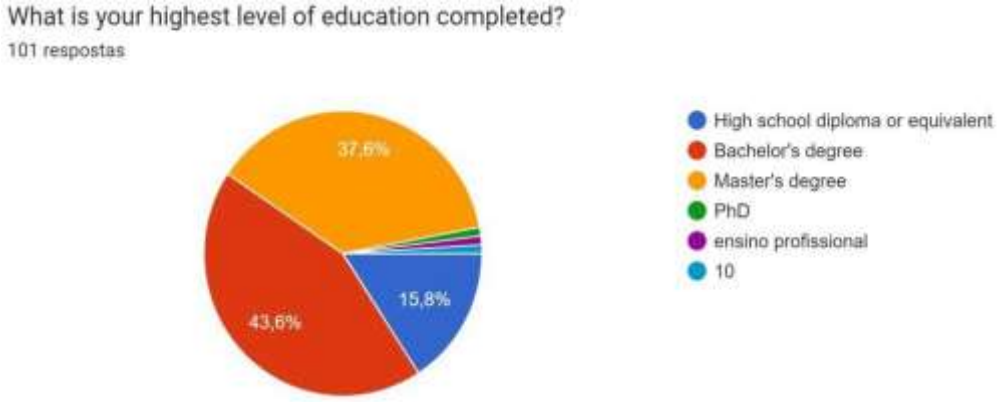

Figure 1-Highest Level of education completed.

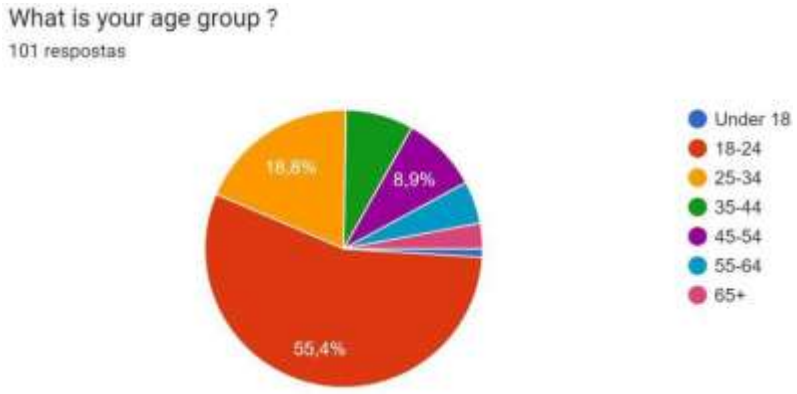

Figure 2-Age group.

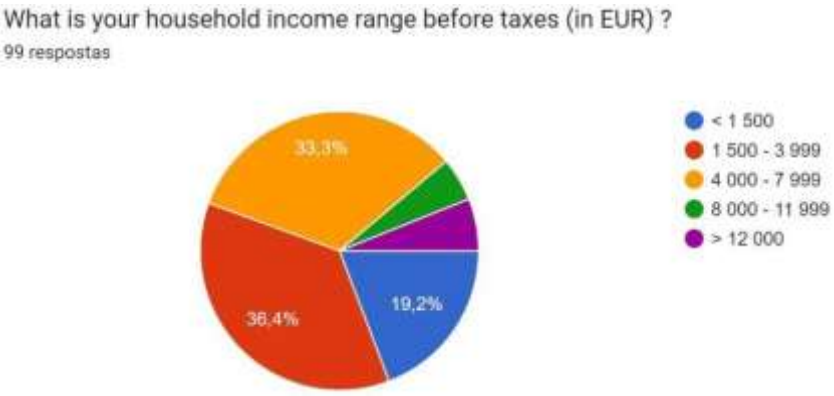

Figure 3-Household income before taxes.

.

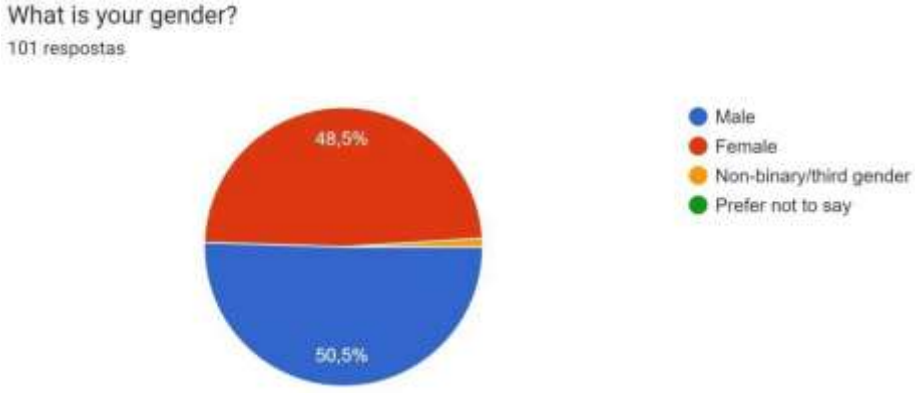

Figure 4-Gender.

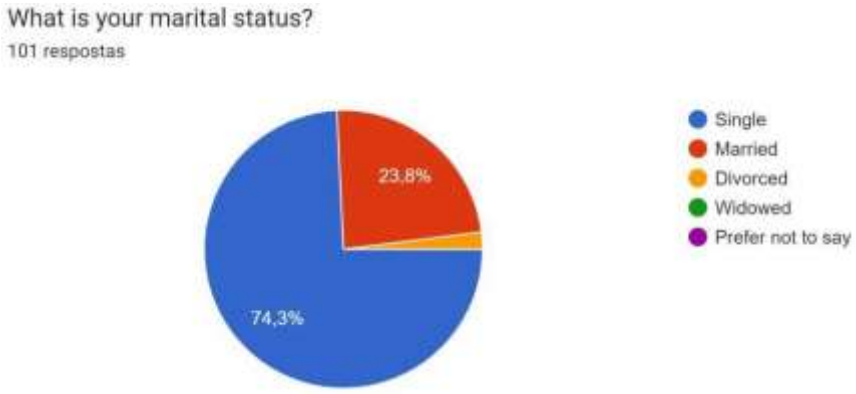

Figure 5- Marital Status.

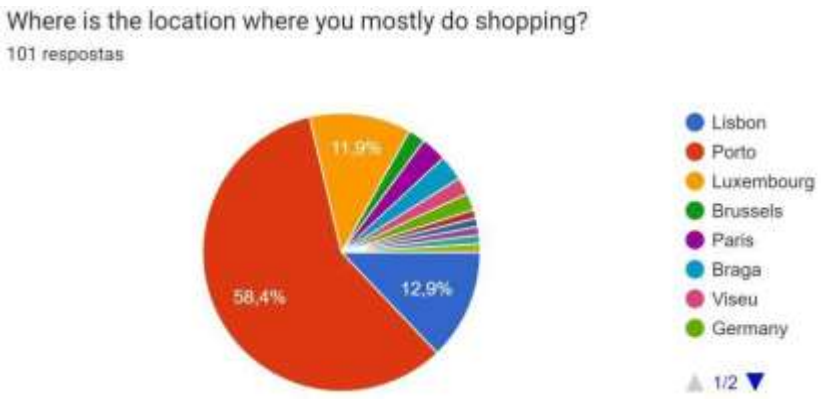

Figure 6- Location of shopping.

.

Figure 7- Online purchasing of fresh products.

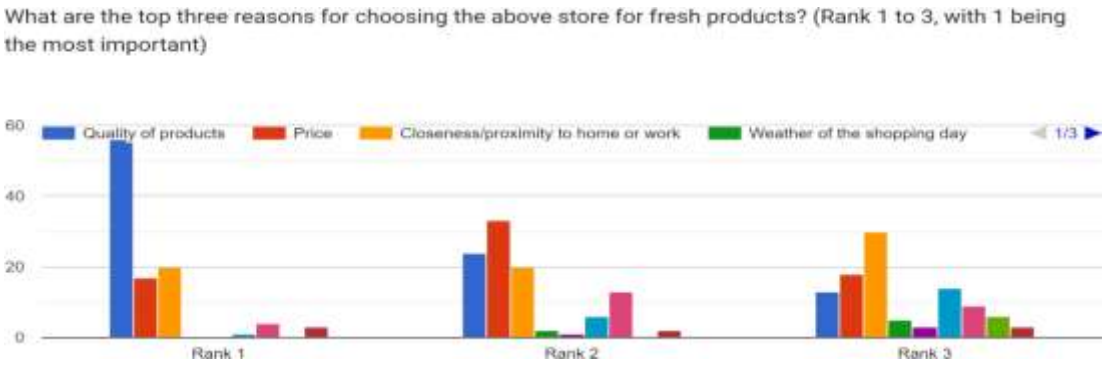

Figure 8- Top reasons to buy fresh products.

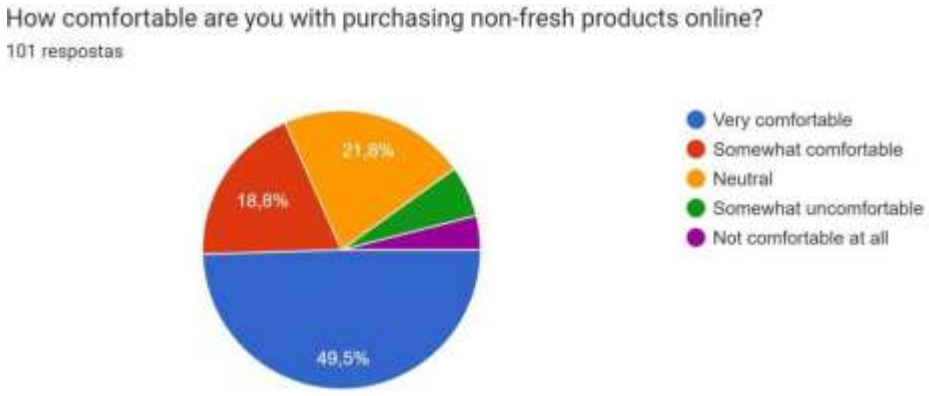

Figure 9- Acceptance of online purchase of non-fresh products.

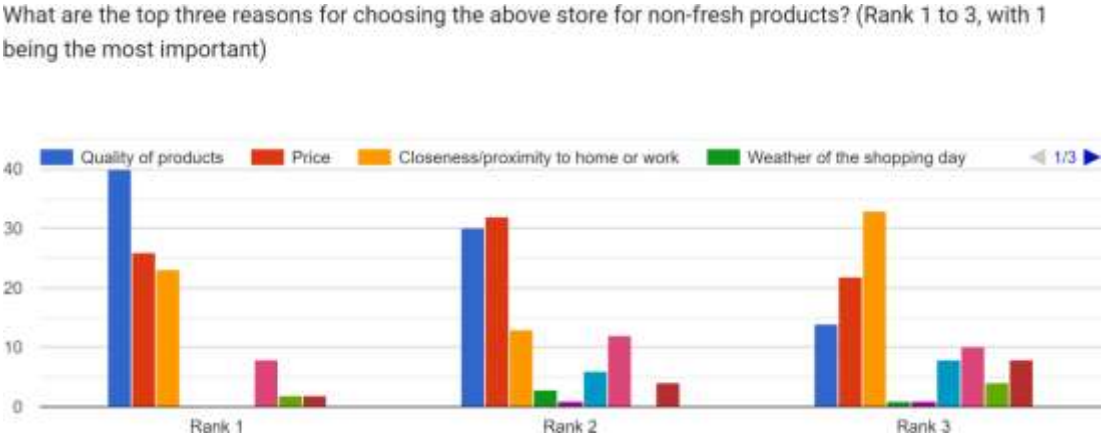

Figure 10- Top decision-making reasons to buy non-fresh products

5.3 Findings

In this subchapter, we aim to describe the results obtained by the methods used for a later discussion.

As seen in Figure 11, employing the methodology section where the Elbow method is explained, we determined that six clusters are optimal. This selection is utilized in our experiments.

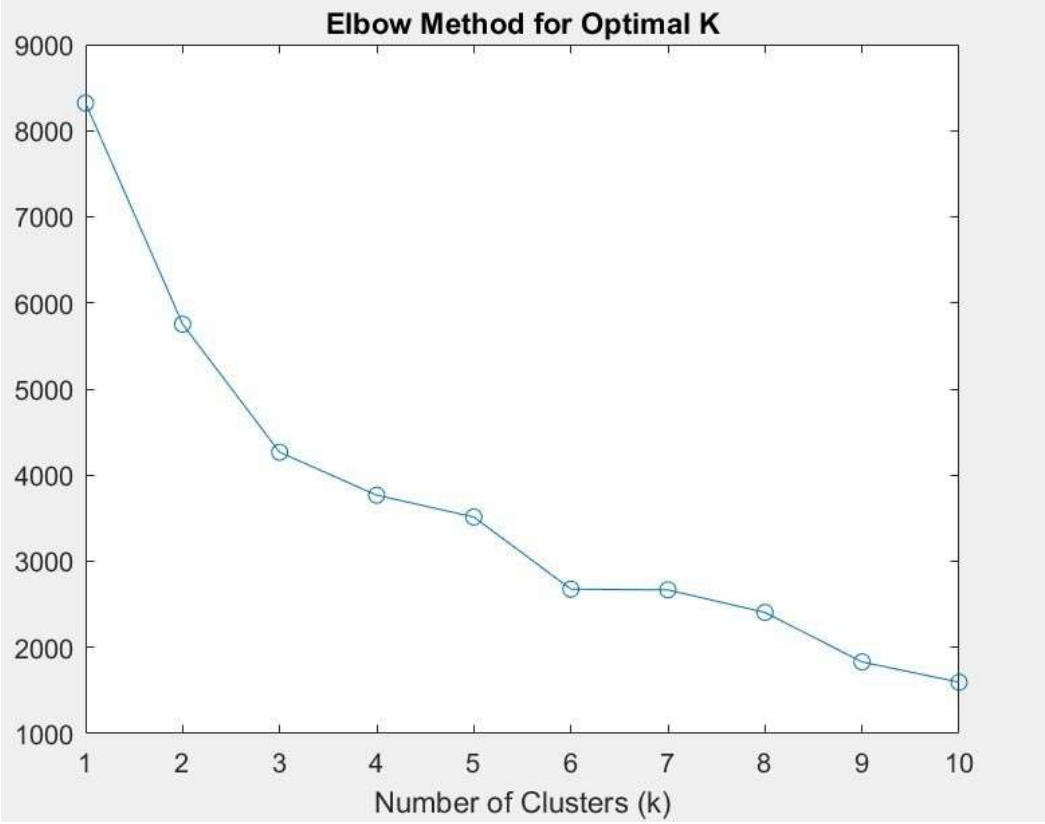

Figure 11- Identify the number of clusters using the Elbow method

5.3.1 Findings using SOMs

Cluster 1 is a group of 20 consumers who share a combination of behavioural and demographic characteristics. Most people in this cluster are married, the majority are female, and their average age is 36. This cluster demonstrates people who have completed a bachelor's degree as their most advanced degree, the household's income before taxes is approximately €2,625 per year. Luxembourg is the main shopping location, and Carrefour is the store of choice for both non-perishable and perishable goods.

These consumers typically buy fresh goods including meat, dairy, fruits, and vegetables once a week. When it comes to fresh food, price is the most important aspect in their decision-making. Weather and proximity to home or work are respectively the next crucial factors. They indicate a "neutral" degree of comfort with buying fresh food online.

Regarding non-perishable food, they do this shopping once a month. When deciding where to buy non-fresh food, the location of the home or job, weather conditions on the day of buying, and weight of purchase are ranked as the most important considerations. When it comes to buying non-fresh goods online, they exhibit a "somewhat comfortable" attitude.

They place "somewhat significant" value on brand names, and they are prepared to pay more for goods that come from sustainable or ecologically friendly sources.

Cluster 2 comprises a group of 14 selective consumers. Most people in this cluster are single men, with an average age of 25. The educational level normally reaches a bachelor's degree, and the average household income before taxes is €2,250. Luxembourg emerges as the predominant shopping location, with *El Corte Inglês* taking the lead as the preferred retailer for perishable items, as opposed to non-perishable goods purchased at *Lidl*.

Fresh products are purchased every week, and pricing is the primary influence in their store selection. Proximity to work influences the top reasons for choosing a store. In-store purchasing is preferred, with online fresh food purchases rated as "neutral."

Moving to non-fresh products, this cluster displays customers who shop every week. The underlying theme in decision-making is pricing. In-store shopping is the preferred mode of purchase for fresh products, and customers feel "somewhat comfortable" making non-fresh purchases online.

Brand names are considered "somewhat important" in their purchasing decisions, and there is a noticeable readiness to pay a premium for environmentally friendly or sustainably sourced products.

Cluster 3 is made up of a separate group of young consumers, with an average age of 23, who are primarily single females with a bachelor's degree. Their household income before taxes is roughly €2,805.

Porto stands out as the main retail destination. Local markets are the primary suppliers of fresh items, while for non-fresh items, customers prefer *Pingo Doce*.

When buying fresh food, the most important elements to consider are price, weather conditions, and store hygiene (Ranked from 1 to 3).

In-store shopping is the preferable option for purchasing fresh foods, as internet fresh food transactions are rated as "somewhat uncomfortable." Turning to non-fresh (non-perishable) products, the frequency of shopping is as well weekly. Similarly, price, weather conditions, and overall store ambiance are paramount in their decision-making.

When it comes to non-fresh (non-perishable) things, shopping occurs every week. Similarly, they prioritize price, weather conditions, and overall store environment when making purchasing decisions.

This cluster is distinguished by a "neutral" comfort level with online purchasing of non-fresh products. Brand names are rated "somewhat important" in purchase decisions, indicating a nuanced concern for quality and brand reputation. Interestingly, the cluster indicates a readiness to pay a premium for products that promote environmentally friendly or sustainable sources.

Cluster 4 represents a customer group that is primarily composed of single females with bachelor's degrees and an average age of 22. Before taxes, their average household income is roughly €1,770.

They decided to purchase fresh goods once a week, and their decision is influenced by the variety, quality, and weather of the goods, in that order. When purchasing these kinds of things, these shoppers prefer the in-store experience.

After switching to non-fresh (non-perishable) goods, the preferred retailer is *Continente*, and monthly shopping is the frequency of shopping.

Price is the primary deciding factor for this kind of product, with weather and product diversity coming in second and third.

When purchasing non-fresh goods online, "somewhat comfortable" is the mark. They rate the significance of brand names as "somewhat important," indicating a careful balancing act between product quality and brand identification. It's interesting to note that the cluster reflects a pragmatic attitude by showing a reluctance to spend more for things that are environmentally friendly or sustainably generated.

Cluster 5 encapsulates the essence of the young, unmarried male generation in Porto who have a strong educational background and are primarily bachelor's degree holders. The average age of this group is 22, and their household income before taxes is an average of €1,710.

Every week, they purchase fresh goods, preferring to shop in person. When making the decision, factors like price and ease of access to work or home are top-ranked.

As stated above, this one also shows a weekly shopping frequency at *Pingo Doce* for non-fresh items, highlighting the significance of reasonable prices and ease of access to home and work.

These customers reveal a "somewhat comfortable" attitude regarding purchasing non-fresh products online.

In their decision-making process, brand names have a "somewhat important" influence, implying a complex fusion of cost and brand awareness. Notably, this cluster demonstrates a deliberate effort towards sustainability by demonstrating a willingness to pay more for goods that are sourced responsibly or ecologically friendly.

Cluster 6 depicts the Conscious Shopper in Brussels, a group of youthful, unmarried women who, on average, are 24 years old and show consideration for

their buying choices. Most of their educational background consists of bachelor's degrees, and their average household income before taxes is €1,920.

This cluster shops once a month when it comes to buying fresh produce. When choosing retailers for fresh products, customers put cleanliness, quality, and proximity first, preferring in-store experiences.

Monthly shopping frequency at *Pingo Doce* is consistent with non-fresh (non-perishable) preferences, which prioritize variety, affordability, and meteorological conditions (ranks 1 through 3, respectively).

They disclose a "somewhat comfortable" stance on non-fresh product purchases made online.

They give brand names "somewhat important" weight when making decisions, reflecting a combination of cost and brand awareness. This cluster, however, illustrates a nuanced approach to sustainability by indicating a hesitant attitude toward paying more for goods that are sustainably sourced or ecologically friendly.

5.3.2 Findings using K-means Clustering

Cluster 1 is a group consisting of 7 individuals characterized by specific behavioural and demographic attributes. This cluster predominantly consists of young females, with an average age of 26 years. Most individuals in this group are single, and they have achieved an undergraduate degree as their highest level of educational attainment. The average household income before taxes stands at approximately €3,285.71 annually. These consumers predominantly conduct their shopping in Brussels, with Carrefour being the preferred retailer for both perishable and non-perishable goods.

Members of this cluster engage in the weekly purchase of fresh goods such as fruits, vegetables, dairy, and meat. Quality, store proximity, and weather are the primary factors influencing their decision-making when selecting a retailer for fresh produce. They demonstrate a 'somewhat comfortable' level of ease with the notion of purchasing fresh goods online.

For non-perishable items, such as canned goods, pasta, and toiletries, shopping is typically done monthly at *Mercadona*. Cleanliness, store cleanliness, and the day's weather conditions are the most significant factors they consider when choosing a store for non-fresh goods. Their comfort level with buying non-perishable items online is also reported as 'somewhat comfortable'.

Brand name holds 'somewhat significant' importance in their purchasing decisions. Additionally, there is a readiness to pay a premium for products that are environmentally friendly or sustainably sourced, although this willingness has its limits.

Cluster 2 represents a distinct group of 24 consumers, predominantly young, single males with an average age of 24 years. The majority have attained a bachelor's degree and report an average household income before taxes of €3,062.50. These consumers typically perform their shopping activities in Luxembourg, with Carrefour being their retailer of choice for both perishable and non-perishable goods.

The members of this cluster exhibit a weekly shopping pattern for fresh produce such as fruits, vegetables, dairy, and meat. Price, store proximity, and weather are the key determinants for their choice of Carrefour for fresh product purchases. They show a 'somewhat comfortable' level of ease with the concept of buying fresh goods online but prefer to make their purchases in-store.

When it comes to non-perishable items, this cluster also shops every week, favouring *Lidl* for such purchases. In choosing a store for non-fresh goods, price is the predominant factor, followed again by price sensitivity and weather conditions. Like their habits with fresh goods, they prefer in-store shopping for non-perishable items and express a 'somewhat comfortable' attitude towards online purchasing in this category.

The brand name holds 'somewhat important' significance in their decision-making process. Furthermore, there is a conditional willingness among this cluster to pay a premium for environmentally friendly or sustainably sourced products, indicating a selective approach to ethical consumption.

This cluster's behaviour highlights a price-conscious and convenience-oriented segment within the retail market, with a partial inclination towards environmental considerations.

Cluster 3 encapsulates a dynamic group of 15 young, single males with an average age of 19 years. These individuals have generally completed their bachelor's degrees and have an average household income of €2,333.33 before taxes. Their shopping activities are frequently conducted in Porto, with the Continent being their preferred choice for purchasing fresh products like fruits, vegetables, dairy, and meat every week.

When selecting *Continente* for their fresh goods, price emerges as the most critical factor, followed by weather conditions and, again, price sensitivity, reflecting a strong inclination towards cost-effective shopping choices. These consumers exhibit a neutral comfort level with purchasing fresh goods online, but they predominantly opt for in-store shopping.

For non-perishable goods such as canned goods, pasta, and toiletries, *Pingo Doce* is the go-to retailer, with weekly visits being the norm. Price and proximity to their location are the primary factors influencing their store choice for non-fresh items, indicating a preference for convenience and value for money. Notably, they report a "very comfortable" attitude towards purchasing non-fresh products online, suggesting a greater openness to e-commerce in this category compared to their purchasing habits for fresh products.

Brand name carries a somewhat important weight in their purchasing decisions, and there is an apparent willingness to invest more in products that are environmentally friendly or sustainably sourced, although this willingness has financial limitations.

This cluster's characteristics suggest a consumer segment that is highly price-conscious, values convenience, and has a measured interest in sustainability, balanced by budget considerations.

Cluster 4 is characterized by a group of 12 young, single females who have completed a bachelor's degree and report an average household income before taxes of €2,125.00. These individuals frequently shop in Porto, with *Pingo Doce* being their selected retailer for purchasing fresh products such as fruits, vegetables, dairy, and meat every week.

The key factors influencing their choice of *Pingo Doce* for fresh goods are the quality of the products, the weather conditions, and the variety offered. They have a neutral attitude toward purchasing fresh items online and show a clear preference for in-store shopping, indicating that the tactile and immediate nature of shopping in physical stores is important to them.

For non-perishable goods, these consumers turn to the Continent, and they do so monthly. The proximity of the store to their home or workplace, the weather at the time of shopping, and the variety of products available are the main considerations for their store choice, suggesting a preference for convenient and diverse shopping experiences. When it comes to purchasing non-fresh goods

online, they report feeling somewhat comfortable, which may indicate a willingness to engage with e-commerce more readily for these types of items.

The importance of brand names in their purchasing decisions is rated as somewhat important, reflecting a discerning approach to product selection that factors in brand reputation. Additionally, their willingness to pay a higher price for environmentally friendly or sustainably sourced products is conditional, indicating a consideration for ethical products that are balanced with financial practicality.

Overall, Cluster 4 depicts a segment that values quality and variety in their shopping, is conscious of convenience, and has a tentative interest in sustainability that is influenced by cost considerations.

Cluster 5 comprises a total of 11 mature demographics of 47-year-old married males, most holding a bachelor's degree and reporting an average household income before taxes of €4,363.64. This group conducts their shopping primarily in Luxembourg, with a strong preference for local markets when it comes to sourcing fresh products like fruits, vegetables, dairy, and meat every week.

When it comes to their reasons for choosing local markets for fresh produce, price is the most decisive factor, followed by the weight of the products, which could indicate a preference for bulk purchasing or an aversion to carrying heavy items for long distances. Store cleanliness is also a significant consideration, reflecting a concern for shopping in a hygienic environment. They have a neutral stance on purchasing fresh products online but show a clear preference for the tactile experience of in-store shopping.

For non-perishable goods, these consumers frequent Cactus monthly. The key factors influencing their choice for non-fresh product purchases are quality, cleanliness, and the variety of options available, suggesting a discerning approach to these types of items. They are somewhat comfortable with purchasing non-fresh products online, indicating a willingness to utilize e-commerce platforms for convenience.

The brand name holds somewhat important significance in their purchasing decisions, and there is a conditional willingness to pay a premium for environmentally friendly or sustainably sourced products, indicating a conscientious approach to shopping that is balanced with budgetary considerations.

In essence, Cluster 5 represents a demographic that values quality and ethical considerations in their shopping, prioritizes convenience, and maintains a level of price sensitivity in their purchasing decisions.

Cluster 6 profiles a 32-observation group of 35-year-old single males, predominantly holding bachelor's degrees and earning an average household income before taxes of €1,812.50. Located in Luxembourg, these individuals demonstrate a weekly routine of purchasing fresh products like fruits, vegetables, dairy, and meat from *Pingo Doce.*

The proximity of the store to their home or workplace is the paramount factor influencing their choice of *Pingo Doce* for fresh products, followed by the price and again the store's closeness, indicating a strong preference for convenience and affordability. Despite the availability of online shopping, their comfort with purchasing fresh products online is neutral, and they prefer in-store experiences.

For non-perishable items such as canned goods, pasta, and toiletries, their retailer of choice is *Mercadona*, with shopping conducted monthly. Similar to their fresh product shopping preferences, closeness and price are significant factors for selecting *Mercadona*, alongside the convenience of the store's location. They are somewhat comfortable with the idea of purchasing non-fresh products online, suggesting a moderate openness to e-commerce for these typesof goods.

The brand name holds a neutral place in their decision-making process, suggesting that while they recognize brands, it is not a primary driver of their purchase decisions. Their willingness to invest more in environmentally friendly or sustainably sourced products is tentative, indicating a consideration for ecological factors that may be dependent on the cost implications.

Cluster 6 thus represents a consumer segment that prioritizes location convenience and price in their shopping habits, with a balanced approach to brand recognition and a conditional interest in sustainability.

5.4 Contribution to the Literature

The comparative analysis of SOMs and K-means clustering techniques reveals distinct differences in their abilities to delineate cluster shape and structure, flexibility in handling data, interpretability, and suitability for high-dimensional data representation.

SOMs excel in identifying and illustrating complex structures and relationships inherent within the dataset. Unlike K-means, which typically enforce a spherical shape on the clusters due to their reliance on centroid-based clustering, SOMs allow for the emergence of a more nuanced cluster geometry that can capture the intrinsic topology of the data. This characteristic of SOMs is particularly advantageous when the dataset contains non-linear relationships that a simple circular boundary cannot adequately represent.

When considering flexibility and the nature of data representation, SOMs demonstrate superior adaptability. They are proficient at managing a diverse range of data structures and distributions, making no prior assumptions about the homogeneity of cluster sizes or densities. K-means, by contrast, operates under the assumption that clusters are of equivalent size and density, which can be a significant limitation in datasets where this is not the case.

The interpretability and usability of the clustering results also vary between the two methods. K-means offers a straightforward approach, where each data point is explicitly assigned to one cluster, making the interpretation of results relatively easy. However, this clarity comes at the expense of detailed insight into the data's structure. SOMs, on the other hand, provide a richer and more intricate representation of data relationships through their 2D grid layout, albeit with

increased complexity in interpretation. The SOM's visualization allows for the observation of clusters and their relationships in a way that is not possible with K-means, highlighting gradients and transitions between different clusters.

In the context of dimensionality and visualization, SOMs hold a significant advantage, particularly in higher-dimensional spaces. They are designed to reduce dimensions while preserving topological and metric relationships between the data points, offering a 2D visual representation of multi-dimensional data. K-means, in isolation, struggle with high-dimensional data due to the curse of dimensionality. It often requires prior dimensionality reduction techniques, such as Principal Component Analysis (PCA), to become effective for visualizing and interpreting clusters in higher-dimensional spaces.

In conclusion, both SOMs and K-means have their unique strengths and weaknesses. The choice between them should be guided by the specific requirements of the data structure, the desired insights from the clustering results, and the complexity of relationships within the dataset. This comparison underscores the importance of selecting the appropriate clustering technique to uncover the true structure of the data in pursuit of meaningful and actionable insights.

5.5 Managerial implications

Understanding the distinctive features and behaviours of consumer clusters is crucial for retailers to design effective marketing and operational strategies. This analysis gives useful information on the diverse interests and habits of six identified customer clusters.

Given the different consumer clusters discovered, there is potential to create unique marketing and operational tactics that appeal to specific populations. We propose to look at practical strategies drawn from the discovered groupings.

Beginning with Cluster 1, which is primarily made up of married people with higher incomes, a focused strategy could include promotions geared toward

family-oriented products or exclusive loyalty programs for repeat high-value customers. Initiatives like family bundle deals or the establishment of a loyalty card programthat offers discounts on bulk purchases could effectively appeal to this group.

Moving on to Cluster 2, the main strategy should highlight Weekly flash sales on new non-fresh items, which could increase the frequency of the buying of these products. Looking at the average age and income, using the mobile app or online website should be a straightforward way to engage this type of client. As we can see in the previous descriptive analysis, the geographic location of the retail shop should be carefully decided to be around universities or business areas filled with young people that can be associated with cluster 2.

Cluster 3, which focuses mostly on local marketplaces in Porto, provides an opportunity to position the brand as a supporter of local businesses. Marketing activities could include collaboration with local farmers for one-of-a-kind in-store events, as well as the use of social media to depict the journey of locally sourced products, fostering a sense of community. Large surfaces, on the other hand, can incorporate typical fresh-product zones to appeal to cluster 3 customers.

For Cluster 4, which demonstrates a practical approach to purchases, marketing strategies should emphasize cost savings and value for money. Initiatives such as a "Smart Shopper" campaign, featuring weekly discounts and cost-effective product bundles, along with a rewards system for consistent monthly shoppers, could be highly effective. To attract this kind of client, the main goal of firms should be to improve their operational backgrounds to achieve economies of scale and deliver the best prices.

Marketing initiatives aimed at Cluster 5, which has an eco-consciousmindset, should include sustainability messaging. This could include launching a sustainability campaign with eco-friendly product lines, as well as implementing practical steps like biodegradable packaging.

Finally, Cluster 6, which is characterized by conscientious and caring buyers, shows an emphasis on quality and cleanliness. Retailers might pitch the brand as a source of high-quality, carefully picked products. We recognize that these individuals expect the shop to keep the hygiene of the items and infrastructure in exemplary circumstances. Adopting a minimalist decorating design should promote engagement among young adults, as suggested by this cluster.

Examining the characteristics that contribute to success in all clusters reveals that most people prefer to buy non-fresh products online and in-store for fresh products. Given this knowledge, businesses can employ focused techniques, such as online-exclusive promotions or fresh products, to encourage and build a popular trend in consumer behaviour. This not only facilitates certain types of purchases but also builds client trust.

The first step in this direction is to develop an operating structure that maximizes value for clients. Businesses that prioritize customer involvement and happiness can establish the groundwork for developing trust in the retail relationship. Furthermore, the geographical environment becomes increasingly important, influencing not only the effectiveness of the distribution system but also the expansion of the client base, given the importance of closeness to home or work to the majority, strategic placement and outreach become critical components in success. Aligning these tactics allows organizations to efficiently react to their broad consumer base's changing preferences, laying the framework for long-term market success.

Conclusion and future work

The present investigation explored the complexities of customer behaviour in the retail sector, revealing a heterogeneous array of preferences, purchasing patterns, and criteria for making decisions among six different clusters. These results offer insightful information for both theoretical and applied research. Age, gender, marital status, and educational background are just a few of the demographic factors that combine to create distinct customer profiles and highlight how complicated consumer choices may be. The study has important ramifications for retailers looking to improve their marketing and operational strategies in addition to providing insights on how customers make choices.

Opportunities for future research directions exist, providing the potential to increase our understanding of the industry. A thorough investigation into changing customer preferences over time is necessary for merchants hoping to maintain their competitiveness in a changing market.

Moreover, a thorough analysis of how various marketing strategies affect customer groups might enable businesses to better customize their approaches. Additionally, examining customers from a global viewpoint might reveal cultural traits that affect buying decisions. This is particularly important for companies that operate in the increasingly interconnected global economy.

As a result, this study not only wraps up a thorough investigation into contemporary consumers but also establishes the foundation for further study that will be essential for companies navigating the always-shifting retail market.

Bibliography

A Machine-Learnt Approach to Market Segmentation and Purchase Prediction Using Point-Of-Sale (POS) Data. (n.d.). Springerprofessional.De. Retrieved 7 February 2024, from https://www.springerprofessional.de/en/a-machine-learnt-approach-to-market-segmentation-and-purchase-pr/23545800

Allred, C. R., Smith, S. M., & Swinyard, W. R. (2006). E-shopping lovers and fearful conservatives: A market segmentation analysis. International Journal of Retail and Distribution Management, 34(4–5), 308–333. Scopus. https://doi.org/10.1108/09590550610660251

Barros, A. J. D., & Hirakata, V. N. (2003). Alternatives for logistic regression in cross-sectional studies: An empirical comparison of models that directly estimate the prevalence ratio. BMC Medical Research Methodology, 3, 1–13. Scopus. https://doi.org/10.1186/1471-2288-3-21

Bauer, H. U., & Schöllhorn, W. (1997). Self-Organizing Maps for the Analysis of Complex Movement Patterns. Neural Processing Letters, 5(3), 193–199. https://doi.org/10.1023/A:1009646811510

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. Journal of Machine Learning Research, 7, 2399–2434.

Berkson, J. (1944). Application of the Logistic Function to Bio-Assay. Journal of the American Statistical Association, 39(227), 357–365. https://doi.org/10.2307/2280041

Burrell, J. (2016). How the machine 'thinks': Understanding opacity in machine learning algorithms. Big Data and Society, 3(1). Scopus. https://doi.org/10.1177/2053951715622512

Chen, D., Sain, S. L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. Journal of Database Marketing and Customer Strategy Management, 19(3), 197–208. Scopus. https://doi.org/10.1057/dbm.2012.17

Chen, J. (2024). The pricing and quality effects of network structure choice: Evidence from American Airlines' international route cancellations. Journal of Air Transport Management, 114. Scopus. https://doi.org/10.1016/j.jairtraman.2023.102489

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. Machine Learning, 20(3), 273–297. Scopus. https://doi.org/10.1023/A:1022627411411

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Practical Assessment, Research, and Evaluation, 10(7). Scopus.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2), 215–232. https://doi.org/10.1111/j.2517-6161.1958.tb00292.x

Debo, L. G., Toktay, L. B., & Van Wassenhove, L. N. (2005). Market Segmentation and Product Technology Selection for Remanufacturable Products. Management Science, 51(8), 1193–1205. https://doi.org/10.1287/mnsc.1050.0369

Dow, J. K., & Endersby, J. W. (2004). Multinomial probit and multinomial logit: A comparison of choice models for voting research. Electoral Studies, 23(1), 107–122. https://doi.org/10.1016/S0261-3794(03)00040-4

Draper, N. R., & Smith, H. (2014). Applied regression analysis (p. 716). Scopus. https://doi.org/10.1002/9781118625590

Dumakude, A., & Ezugwu, A. E. (2023). Automated COVID-19 detection with convolutional neural networks. Scientific Reports, 13(1). Scopus. https://doi.org/10.1038/s41598-023-37743-4

Engel, J. (1988). Polytomous logistic regression. Statistica Neerlandica, 42(4), 233–252. https://doi.org/10.1111/j.1467-9574.1988.tb01238.x

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). Cluster Analysis. John Wiley & Sons.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological Methods, 4(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. Science, 315(5814), 972–976. Scopus. https://doi.org/10.1126/science.1136800

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Journal of the Royal Statistical Society. Series C (Applied Statistics), 28(1), 100–108. https://doi.org/10.2307/2346830

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Huerta-Muñoz, D. L., Ríos-Mercado, R. Z., & Ruiz, R. (2017). An iterated greedy heuristic for a market segmentation problem with multiple attributes. European Journal of Operational Research, 261(1), 75–87. Scopus. https://doi.org/10.1016/j.ejor.2017.02.013

Hung, P. D., Ngoc, N. D., & Hanh, T. D. (2019). K-means clustering using R A case study of market segmentation. 100–104. Scopus. https://doi.org/10.1145/3317614.3317626

Kakhi, K., Alizadehsani, R., Kabir, H. M. D., Khosravi, A., Nahavandi, S., & Acharya, U. R. (2022). The internet of medical things and artificial intelligence: Trends, challenges, and opportunities. Biocybernetics and Biomedical Engineering, 42(3), 749–771. Scopus. https://doi.org/10.1016/j.bbe.2022.05.008

King, G., & Zeng, L. (2003). Logistic regression in rare events data. Journal of Statistical Software, 8, 137–163. Scopus. https://doi.org/10.18637/jss.v008.i02

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43(1), 59–69. https://doi.org/10.1007/BF00337288

Kohonen, T. (1988). An introduction to neural computing. Neural Networks, 1(1), 3–16. Scopus. https://doi.org/10.1016/0893-6080(88)90020-2

Kohonen, T. (1990). The self-organizing map. Proceedings of the IEEE, 78(9), 1464–1480. https://doi.org/10.1109/5.58325

Kohonen, T. (1999). Analysis of processes and large data sets by a self-organizing method. 1, 27–36. Scopus. https://doi.org/10.1109/IPMM.1999.792450

Kohonen, T. (2013). Essentials of the self-organizing map. Neural Networks, 37, 52–65. Scopus. https://doi.org/10.1016/j.neunet.2012.09.018

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., & Tischer, T. (2022). Machine learning and conventional statistics: Making sense of the differences. Knee Surgery, Sports Traumatology, Arthroscopy, 30(3), 753–757. https://doi.org/10.1007/s00167-022-06896-6

Li, J., & Lewis, H. W. (2016). Fuzzy Clustering Algorithms—Review of the Applications. 282–288. Scopus. https://doi.org/10.1109/SmartCloud.2016.14

Mika, S., Ratsch, G., Weston, J., Scholkopf, B., & Mullers, K. R. (1999). Fisher discriminant analysis with kernels. Neural Networks for Signal Processing IX: Proceedings of the 1999 IEEE Signal Processing Society Workshop (Cat. No.98TH8468), 41–48. https://doi.org/10.1109/NNSP.1999.788121

Mitchell, T. M. (1997). Machine Learning. McGraw-Hill.

Mohamed, A. A. (2020). An effective dimension reduction algorithm for clustering Arabic text. Egyptian Informatics Journal, 21(1), 1–5. https://doi.org/10.1016/j.eij.2019.05.002

Murat, S. (n.d.). A brief review of feed-forward neural networks. Retrieved 15 July 2023, from https://dergipark.org.tr/en/pub/aupse/article/890416

Patankar, N., Dixit, S., Bhamare, A., Darpel, A., & Raina, R. (2021). Customer Segmentation Using Machine Learning. https://doi.org/10.3233/APC210200

Regmi, S. R., Meena, J., Kanojia, U., & Kant, V. (2022). Customer Market Segmentation using Machine Learning Algorithm. 1348–1354. Scopus. https://doi.org/10.1109/ICOEI53556.2022.9777146

Roy, S., Menapace, W., Oei, S., Luijten, B., Fini, E., Saltori, C., Huijben, I., Chennakeshava, N., Mento, F., Sentelli, A., Peschiera, E., Trevisan, R., Maschietto, G., Torri, E., Inchingolo, R., Smargiassi, A., Soldati, G., Rota, P., Passerini, A., … Demi, L. (2020). Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound. IEEE Transactions on Medical Imaging, 39(8), 2676–2687. Scopus. https://doi.org/10.1109/TMI.2020.2994459

Schmidhuber, J. (2015). Deep Learning in neural networks: An overview. Neural Networks, 61, 85–117. Scopus. https://doi.org/10.1016/j.neunet.2014.09.003

Sugiyama, M., & Jp, C. T. A. (n.d.). Dimensionality Reduction of Multimodal Labeled Data by Local Fisher Discriminant Analysis.

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition | SpringerLink. (n.d.). Retrieved 2 October 2023, from https://link.springer.com/book/10.1007/978-0-387-84858-7

Thomas, P. (2009). Semi-Supervised Learning by Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien (Review). IEEE Transactions on Neural Networks, 20, 542.

Ustebay, S., Yelmen, I., & Zontul, M. (n.d.). Customer Segmentation Based on Self-Organizing Maps: A Case Study on Airline Passengers.

Van Hulle, M. M. (2012). Self-organizing Maps. In G. Rozenberg, T. Bäck, & J. N. Kok (Eds.), Handbook of Natural Computing (pp. 585–622). Springer. https://doi.org/10.1007/978-3-540-92910-9_19

Vapnik, V. N. (1999). An overview of statistical learning theory. IEEE Transactions on Neural Networks, 10(5), 988–999. Scopus. https://doi.org/10.1109/72.788640

Verdu, S. V., Garcia, M. O., Senabre, C., Marin, A. G., & Franco, F. J. G. (2006). Classification, Filtering, and Identification of Electrical Customer Load Patterns Through the Use of Self-Organizing Maps. IEEE Transactions on Power Systems, 21(4), 1672–1682. https://doi.org/10.1109/TPWRS.2006.881133

Appendix

Appendix 1 – Survey Questionnaire

A comparison between statistical and machine learning methods for retail market segmentation.

This thesis undertakes a thorough investigation into market segmentation in the retail sector, specifically focusing on the comparison between statistical and machine learning methods. It aims to unravel the intricate patterns that define consumer preferences and behaviors within the realm of retail. By employing both statistical and machine learning techniques, the study seeks to identify the most effective strategies for categorizing diverse consumer segments. The research delves into the nuances of the retail market, meticulously examining the performance and practical implications of each segmentation approach. Through this effort, we aim to not only enhance our understanding of market dynamics but also provide valuable insights for stakeholders navigating the complex landscape of retailofferings.

Furthermore, I thank you for allowing me to collect the necessary data for the research.

For Survey Circle users (www.surveycircle.com): The Survey Code is: HAH6-GFNQ- NAPS-6DPY

What is your age group?*

- Under 18
- 18-24
- 25-34
- 35-44
- 45-54
- 55-64
- 65+

What is your gender?

- Male Female
- Non-binary/third gender Prefer not to say
-

What is your highest level of education completed? *

- High school diploma or equivalent
- Bachelor's degree
- Master's degree
- PhD
- Other: _____

What is your household income range before taxes (in EUR)?

- [ ] < 1 500
- [ ] 1 500 - 3 999
- [ ] 4 000 - 7 999
- [ ] 8 000 - 11 999
- [ ] > 12 000

Where is the location where you mostly do shopping? *

- [ ] Lisbon
- [ ] Porto
- [ ] Luxembourg
- [ ] Brussels
- [ ] Paris
- [ ] Other: _____

What is your occupation? *

_____

When purchasing fresh (perishable) products like fruits, vegetables, dairy, and meat, which store do you usuallyprefer?

- [ ] Continente
- [ ] Pingo Doce
- [ ] Mercadona
- [ ] Aldi
- [ ] El corte inglés
- [ ] Lidl
- [ ] Auchan
- [ ] Cactus
- [ ] Carrefour
- [ ] Local Markets
- [ ] Other: _____

How frequently do you purchase fresh products like fruits, vegetables, dairy, and meat, from this store?

- [ ] Daily
- [ ] Weekly
- [ ] Monthly
- [ ] Rarely

What are thetopthreereasonsforchoosingthe above storeforfreshproducts? (Rank 1 to 3, with 1 being the most important)

	Quality of products	Price	Closeness/proximity to home or work	Weather of the shopping day	Weigh of the order	Store cleanliness and hygiene
Rank 1						
Rank 2						
Rank 3						

How do you prefer to shop for fresh products? *

- In-store shopping
- Online ordering with home delivery
- Online ordering with click-and-collect service
- Other: _____

Does your preferred store offer your preferred method of shopping for fresh * products?

- Yes
- No
- Not sure

How comfortable are you with purchasing fresh products online? *

- ◯ Very comfortable
- ◯ Somewhat comfortable
- ◯ Neutral
- ◯ Somewhat uncomfortable
- ◯ Not comfortable at all

When purchasing non-fresh (non-perishable) products like canned goods, pasta, and toiletries, which store do you usually prefer?

- ◯ Continente
- ◯ Pingo Doce
- ◯ Mercadona
- ◯ Aldi
- ◯ El corte inglés
- ◯ Lidl
- ◯ Auchan
- ◯ Cactus
- ◯ Carrefour
- ◯ Local Markets
- ◯ Other: _____

How frequently do you purchase non-fresh products like canned goods, pasta, and toiletries, from this store? *

- Daily
- Weekly
- Monthly
- Rarely

Whatarethetopthreereasonsforchoosingtheabovestorefornon-freshproducts? (Rank 1 to 3, with 1 being the most important)

	Quality of products	Price	Closeness/proximity to home or work	Weather of the shopping day	Weigh of the order	Store cleanliness and hygiene
Rank 1						
Rank 2						
Rank 3						

How do you prefer to shop for non-fresh products?

- In-store shopping
- Online ordering with home delivery
- Online ordering with click-and-collect service
- Other: _____

How comfortable are you with purchasing non-fresh products online? *

- Very comfortable
- Somewhat comfortable
- Neutral
- Somewhat uncomfortable
- Not comfortable at all

How important is a brand name to you when making a purchase decision?

- Very important
- Somewhat important
- Neutral
- Not very important
- Not at all important

Are you willing to pay a higher price for products thatare environmentally friendly or sustainably sourced? *

- Yes, definitely.
- Yes, but only up to a certain point
- Maybe, it depends on theproduct
- Not really.
- No, not at all.

If you could change one thing about the way you shop, what would it be?

Please share any additional comments or suggestions you have about your shopping experience.