NOVA
IMS
Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação
Master Program in Statistics and Information Management

Pricing Network Analysis of a supermarket chain in Portugal

Ana Catarina Lopes da Silva

Project Work presented as partial requirement for the Master's Degree of Statistics and Information Management

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

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

PRICING NETWORK ANALYSIS OF A SUPERMARKET CHAIN IN PORTUGAL

BY

Ana Catarina Lopes da Silva

Project Work presented as partial requirement for the Master's Degree of Statistics and Information Management, with a specialization in Information Analysis and Management

Advisor: Flávio L. Pinheiro, PhD

November 2021

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to Prof. Flávio Pinheiro, from Nova Information and Management School (IMS), for his patient guidance, enthusiasm, and assistance during my research and writing progress. His willingness to give his time to provide his knowledge and give me constructive recommendations on this project. With his valuable knowledge and guidance, it was possible to overcome some of the challenges presented by the data and complete this work.

A special thank you to my family and closest friends for their constant encouragement, support, and understanding of all my absences for the past two years, that I invested and committed to finishing my master's degree.

Last but not least, I would like to thank my company, Deloitte Portugal, which provided me access to the data and to the architecture implemented to export a sample that was used to perform this research.

ABSTRACT

With the growth of digital transformation, the business is now conducted in the digital age, opening new markets and new business opportunities with a simple click, sitting at home. Along with it, as the internet grows, more data is exposed, which leads us to competitive advantages, especially during the pandemic and in the post-pandemic period.

This work explores daily price data from the online website of one of the largest supermarket chains in Portugal, covering the period between February 2020 to March of 2021. To observe the market and the competition behavior, we applied the science of networks, a data mining technique that provides deeper insight into the structure of the retailer market.

We used the Maximum Spanning Tree and added the most relevant links from the correlation matrix to represent the retail market. As a starting point, we were able to build five networks that represent the daily price changes for products, categories, and brands. Moreover, we study the network interactions using centrality measures, namely: degree, betweenness, closeness, and eigenvector.

These outputs were useful by creating better visibility of the market using real data from online retailing platforms and making available the first input for future work and further research to improve the pricing strategies.

KEYWORDS

Pricing Network; Price Correlations; Maximum Spanning Tree; Complex Networks

INDEX

1. Introduction 1
   1.1. Context 1
   1.2. Problem Identification 1
   1.3. Objectives and Research Questions 1
   1.4. Structure 2
2. Literature Review 3
   2.1. The Complexity of Pricing Behavior 3
      2.1.1. Network Analysis of Stock Markets 6
   2.2. The Science of Networks 9
      2.2.1. Graph Structures 11
      2.2.2. Methodology 12
      2.2.3. Evaluation Measures 13
3. From Data to Networks 15
   3.1. Data, Collection and Processing 15
   3.2. Price Correlations 19
   3.3. Network Inference & Characterization 20
      3.3.1. Products Network Analysis 22
      3.3.2. Categories Networks Analysis 24
4. Brand Pricing Analysis 27
5. Conclusion 31
   5.1. Challenges and Limitations 32
   5.2. Future Work 32
6. Bibliography 33
7. Appendix A – Product Classification 35

LIST OF FIGURES

Figure 1 - Retail prices for a half gallon of whole milk on the Upper West Side of Manhattan, May 2002 (Phillips, 2005) 5

Figure 2 - Sectors represented by stocks in the network (Smith, 2009) 6

Figure 3 - Edge density for different values of correlation threshold (Huang et al., 2009) 7

Figure 4 - PCA for TSE before and after removing the market mode (Namaki et al., 2011) 8

Figure 5 - Relative changes of vertices for different thresholds (Yin et al., 2017) 9

Figure 6 - Sample Network with 6 nodes and 7 edges (Newman & Girvan, 2004) 10

Figure 7 - Directed Network with weighted edges (Horwood, 2015) 11

Figure 8 - Examples of the most representative network types (Random network, Small-World network and Scale-Free network) – Adapted from (Hodler et al., 2019) 11

Figure 9 - Frequency distribution of products in each time interval of 15 days 15

Figure 10 - Number of products with price observations per week of the year 16

Figure 11 - Temporal variations of prices for a few representative products, covering the period between February 2020 to March of 2021. The price values were obtained from the online webstore of the supermarket chain, using Web Scraper API. 16

Figure 12 - Sample of the two level one categories, using a tree structure from a graph. The category levels are extracted from the website and reflects the retailer categorization of the products used on the website. 17

Figure 13 - A histogram in which distribution of the average price, transformed using the natural logarithm, is plotted on the vertical axis against the total of products on the horizontal axis. 18

Figure 14 - Quartiles Distribution with 95% Confidence Limits for the quantiles 0.25, 0.5 and 0.75 and actual values. The actual values are represented by the average price (Avg Price) per week 19

Figure 15 - Undirected and weighted product network. The strong relationships are associated with a lower distance between the nodes. 22

Figure 16 - Degree distribution, transformed using the natural logarithm, per number of Products. 23

Figure 17 - Undirected and weighted category networks (Category Level 1 + Category Level 2 + Category Level 3). The size of each node is proportional to the number of strong correlations with other nodes (degree), and the strong relationships are associated with a lower distance between the nodes. 24

Figure 18 - Temporal variations of prices for a few representative brands, covering the period between February 2020 to March of 2021. The brand information was obtained from the online webstore of the supermarket chain, using Web Scraper API. 27
Figure 19 - Undirected and weighted brand network. The strong relationships are associated with a lower distance between the nodes. 28
Figure 20 - Degree distribution, transformed using the natural logarithm, per number of Brands... 29

LIST OF TABLES

Table 1 - Summary of the data processing steps and the impact of each in the size of the number of products in the Data Set. Each Step was implemented in a sequential order (from top to bottom). The steps show how the number of products were filtered from raw data to the working data set. 18

Table 2 - Considered tasks for network representation along with the number of edges resulted for each network (Products + Category levels). Each task was implemented in a sequential order (from top to bottom). The total of relevant edges added is identified in brackets after the edges value. 21

Table 3 - Network analysis metrics for Products + Category levels classification, including the number of nodes in the network, the network density, the average degree, and average path length and the cluster coefficient 21

Table 4 - Top 5 products by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value. 23

Table 5 - Top 3 **Level 1 categories** by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value. 25

Table 6 - Top 3 **Level 2 categories** by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value. 25

Table 7 - Top 3 **Level 3 categories** by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value. 25

Table 8 - Categorization measures of the impacted product observed in products network (“Adoçante”). The position is identified in brackets after the centrality value. 26

Table 9 - Network analysis metrics for Brand classification, including the number of nodes in the network, the density, the average degree, and average path length and the cluster coefficient 28

Table 10 - Top 5 brands by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value 29

1. INTRODUCTION

Many analysts argued, in effect, that the Internet would bring about the world of perfect competition, in which sellers would lose control of prices. However, the reality has been quite different.

(Phillips, 2005)

1.1. CONTEXT

The Digital transformation has changed the way that companies work in the different sectors, namely in the retail sector. Retail is continually affected by advances in digital technologies and has been changing the way that customers do shopping through the digital market or the so-called online marketplaces.

The advances in digital technologies and the digital marketing trends lead to significant, sometimes drastic and/or disruptive changes in the competitive landscape with the easiness access to the information about a specific product from different players of the market using a single click by sitting at home (Pandey et al., 2014).

In this context, the online marketplaces increased the transparency of the prices, portfolio information, and retailers' strategies like campaigns or promotions. This makes web scraping a new data-driven strategy to make use of e-commerce data to analyze competitors, monitor price changes, and customize pricing strategies according to the market trends.

1.2. PROBLEM IDENTIFICATION

According to (Phillips, 2005), modern economics claims that in perfectly competitive markets there is no "right price" for a good or service, there are only the actual prices out in the marketplace, based only on the willingness of sellers to sell and buyers to buy. Companies act like price takers, which means that they only control the quantity to produce and whether to enter or exit a market.

However, in our real-world prices are messier than the theory can explain, with unjustifiable price variations that sometimes seems irrational. This gap between the theory and its application to the real world is possibly reflected by the acceleration of business in all fields, supported also by the digital trends, and the increasing need to make timely pricing decisions to respond to competitive actions, market changes, or their own inventory situation (Phillips, 2005).

Understanding the complexity and variation of the price system, powered by the constant move forward and adapt to new market opportunities, is the challenge that retailers are facing to find regularities and/or similar behaviors for better control of the market and human consumption.

1.3. OBJECTIVES AND RESEARCH QUESTIONS

Using the information extracted from the online website of one of the largest supermarket chains in Portugal and considering relevant literature develop to study market prices, our work is focused on the study of price data from a single retail market, by looking at daily price data and using a technical analysis strategy used in stock markets to identify price correlations and reflect market's behavior.

We use a bottom-up approach to analyze individual product prices changes before focusing on product categorizations and retailer's brand partnerships. In this scope, we approach the following questions:

- How do prices differ within a retailer's product portfolio?
- Which products and categories are positively correlated to each other?
- Which products and categories have the greatest influence on the retailer's price structure?
- How do prices differ between retailer's Brand Partnerships?
- Which brands are positively correlated to each other?

With this work, we attend to explore relationships of the e-commerce data and extract insights that could be valued to be used in future work and support business decisions. Additionally, our work provides a methodology based on Pearson Correlation to measure the product relationships and monitor price fluctuations.

1.4. STRUCTURE

In **Chapter 2**, a literature review is conducted for the aim of our work. First, we start by addressing the complexity associated with the price system and its structure. Case studies are presented to explain the pricing behaviors and the strong influence of the customer on price strategies. Additionally, a subsection was created to provide the use of the network approach in stock markets to study and understand the price system. For a better understanding of the work developed, we finalize with the introduction of the main concepts of the network science and methodology used for network inference and characterization.

Chapter 3 presents the exploratory analysis of the available data set and the pre-processing steps used to achieve our working data set. Next, we introduce how we measure the relationships between products and categories, using the Person Correlation, and the steps performed for the network inference. We finalize with an analysis of the network structures of products and categories, using centrality measures and network indicators.

Chapter 4 provides an additional analysis of the retailer's Brand Partnerships, following the same pre-processing steps used in the previous chapter for data preparation and network inference. We finalize with a brief analysis of the main characteristics and provide some suggestions to the main stakeholders.

Finally, in **chapter 5**, we provide some concluding remarks based on the network structures and characterization, and describe the main challenges of this work, along with future work opportunities to enrich our research.

2. LITERATURE REVIEW

In this section, we summarize the main elements of the research to support and understand the work performed. We structure the section in two subsections: in **subsection 2.1.** we introduce the complexity associated with the price structure, including a dedicated section to the study of prices through the network's approach on stock markets. We explain how the networks are used in real work problems and their importance for the study the price behavior; in **subsection 2.2.** we provide a theoretical background of the concept of networks, presenting the main characteristics of a network and the different centrality measures.

2.1. THE COMPLEXITY OF PRICING BEHAVIOR

According to (Phillips, 2005) one of the first applications that was developed based on pricing and revenue optimization was the revenue management systems by the passenger airlines in the 1980s. This application was a success that increased profitability and since then, several studies were conducted, and new trends were raised to give name to a core competency within companies that we call the pricing and revenue optimization.

The search for customer information and the implementation of new processes to collect this data increased along with the online marketplaces and led to new systems that collect customer and transaction data from different channels and make it available in a data warehouse to various business intelligence, data-mining, and analytic systems (Phillips, 2005).

In the industry of the airlines, research was conducted to manage the price of the transportation fare based on the baggage overweight and question the policies currently used by airlines, by one interview to an expert of the area. According to (Okhrimenko et al., 2020), airlines were not considering any customer factors and that could explain why the passengers tried to cheat and avoided the extra charges.

To overcome this problem, it was developed a price optimization model based on determining consumer Willingness-To-Pay (WTP), especially focusing on the company level where only external factors were not considered, for example, ecological fees or local laws in the job security field. The optimization processes required a full understanding of the matter and all the factors that are associated with price variations, sometimes hard to control by the company.

They start to define utility, based on the law on supply and demand, and smoothly considering factors that had better adjusted the model to the reality based on interviews, for example:

- The amount that a customer is willing to spend on a service;
- The limited baggage overweight for a given customer.

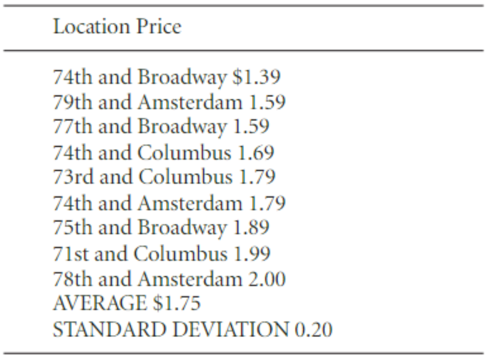
This study proved that the theoretical Willingness to Pay model, expanded and added to the optimization model, can be influenced by airline company. The next steps were to find out the maximum possible price that maximizes revenue and obtain optimized value revenue. Later, to achieve a good approximation of the model, expansions were added, for example, the price for the flight, the flight duration, the aircraft fuel cost.

In the hotel industry, using booking data from 28 United States hotels and presenting five different brands, the authors (Lee, 2011) studied the validity of two key assumptions in hotel revenue management: the first is the customers who book later are WTP higher rates than customer's who book earlier and the second supports that demand is stronger during the week than on the weekend. It has possible to observe the relationships between these assumptions and the behavior of the booking data (average rates paid, occupancy rates and demand) to conclude that further study is required before opting for lower rates during the weekend and/or close to the arrival date. This study suggests that revenue managers need to have a clear understanding of how demand will change under different market conditions and pricing structures to adapt their prices and maximize profit. According to this study, differentiation is a key factor that must be considered than simply matching competitor rates to avoid losing market share.

Studying retailer pricing behavior is an area of interest in the marketing literature (Chintagunta, 2002). Usually when a market player produces a homogenous product and decides to raise its price it probably will lose all its market to its competitors. However, in practice studies have shown that this assumption can be applied profitably. Focus on consumer and industrial markets (Chen et al., 2001) shows us that consumers with a better preference at a particular store will not take advantage of competitor's price cuts. As a result, companies that sell an identical product and choose to institute price-matching strategy by selling their products with the lowest price of the market or based on min{*pA*, *pB*} when considering two competing stores (denoted A and B), can typically suffer the subsequent loss of profits. This happens when a company do not consider consumer segmentation and assumes that all its customers are indifferent to a particular store or specific products.

According to this study, the key determinant in the application of the price-matching strategy profitably is related to the composition of consumers in a market. This means that one company can have higher profit margin when the number of loyals and switchers (consumers that always buy the product that has the lowest price) are the same when compared to its competitors. In this scope, (Chen et al., 2001) shows us the complexity of the price system taking of evidence through their research that price-matching strategy can indeed facilitate competition when considering consumer segmentation and it is applied strategically.

To study the price of a specific product in the west side of Manhattan, (Phillips, 2005) show us the prices of milk at different markets in a 16-block area on a single day in May 2002. The results in Figure 1, show us the price variation of 44% with a range from a low of $1.39 to a high of $2.00 and varied by more than $0.40 for two stores on the same block.



Location	Price
74th and Broadway	$1.39
79th and Amsterdam	1.59
77th and Broadway	1.59
74th and Columbus	1.69
73rd and Columbus	1.79
74th and Amsterdam	1.79
75th and Broadway	1.89
71st and Columbus	1.99
78th and Amsterdam	2.00
AVERAGE	$1.75
STANDARD DEVIATION	0.20

Figure 1 - Retail prices for a half gallon of whole milk on the Upper West Side of Manhattan, May 2002 (Phillips, 2005)

That said we could conclude that prices can variate between different supplies and emphasize that companies need to implement new ways to deliver richer views of price variations and complement with information about customer behavior. We are facing a new Age in which is necessary to continually monitor demand and update prices along with the segmentation of the customers based on their WTP as a critical piece of pricing and revenue optimization.

Additionally, it was proved that online marketplaces support price differentials in the same way as physical markets. With the growth of digital transformation and the comfort (Phillips, 2005) of checking prices online, price-matching has become an appellant strategy for retailers both in the product and service industries. A recent study, developed with the aim of the price matching on the online sales platforms, point that the online marketplaces take advantage of the Price Matching Guarantees (PMGs) policies, that is the promise to reimburse price differences when competitors offer lower prices in order to increase consumer confidence and brand fidelity (Bottasso et al., 2020). This study was focused on the effect of platforms' PMGs policies on daily consumer electronics prices observed on US online market by comparing the prices before and after the policy shutdown. It was considered 29 products taking 9028 daily price observations (174 days) on NewEgg (a leading online US retailer of consumer electronics products) and on Amazon UK platforms for price control (without the PMGs policies).

In the first stage, it was conducted a descriptive statistics of the prices, analyzing the differences between categories based on the quality and price levels. Based on the results, the causal impact of PMGs on prices was estimated through a Diff-in-Diff research design. The Diff-in-Diff design enables to study the effects of a strategy by observing the differential effects of products exposed to the intervention, in this case, the PMGs policies, identified as a 'treatment group', with the products not exposed to the intervention/policies, called the 'control group', in this case the sample extracted from Amazon UK platforms. In online consumer electronics market, the DiD estimates suggested a harmful competition for high visible products by reduction 3,7% after the PMGs shutdown, while the low visible products have an increase of 3,4%. Currently, the online retailing platforms are a new tool that easily recovers the information on buyers by monitoring competitor's prices through price-tracking systems and easily identify practices of price discriminations of PMGs policies.

To observe the market and the competition behavior, researchers are embedded to study price behaviors through the science of networks, especially on stock markets. In this work, we focus on the power of data on network interactions to improve the pricing strategies and create better visibility of the market using real data from online retailing platforms.

2.1.1. Network Analysis of Stock Markets

Correlation-based networks in the stock market are mostly used to explore the relationship between different stocks prices and provide deeper insights into the internal market structure.

The credit crisis in early 2007 was a phenomenon discussed through a correlation network of stocks and used to highlight the movements of the market as a key of the progress of the crisis. In this context, the authors start by defining a **correlation matrix** of returns between two stocks, where each daily value was the log-return of the closing price from the previous day (Smith, 2009). Based on these correlations, the distance between each stock is defined and used to create the graphical **minimal spanning tree** colored by sector (Figure 2).

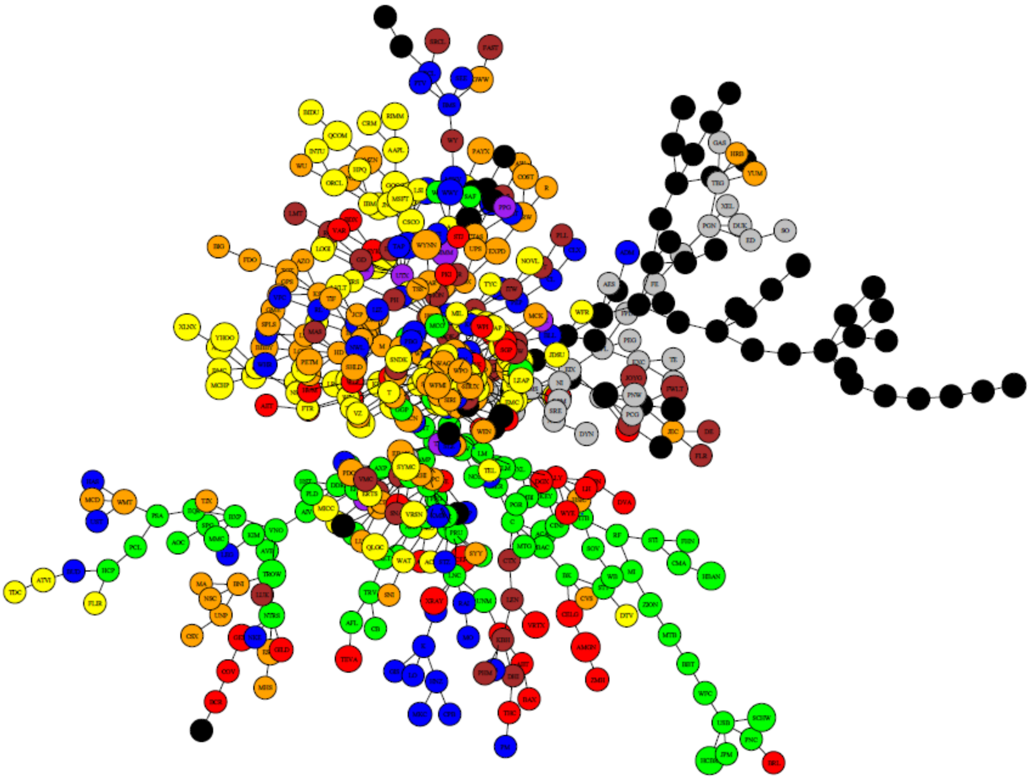


Figure 2 - Sectors represented by stocks in the network (Smith, 2009)

The central part of the network is clary dominated by finance and service sectors (orange for service firms and green for finance firms) corresponding to the first stocks impacted by the crisis. On the other hand, the periphery is dominated by industry specific showing us that the collapse in stock price is recognized and extended for all sectors across the economy (Smith, 2009).

Many complex systems can be described by complex networks that allow us to study the market's behavior. This reality in the Chinese stock market was studied by (Huang et al., 2009) that established the corresponded stock correlation network and found a hierarchical arrangement of stocks. The authors considered 1080 stocks and started by calculating the return of stock-price of daily changes over a period of 1198 days. Based on this information, the cross-correlation between the individual stocks was calculated and analyzed using the **correlation threshold method** (Huang et al., 2009). This study evaluated the degree distribution and the edge density for different threshold values (Figure 3) to evaluate the network structure and identify important stocks in the price fluctuation correlation patterns of stocks.

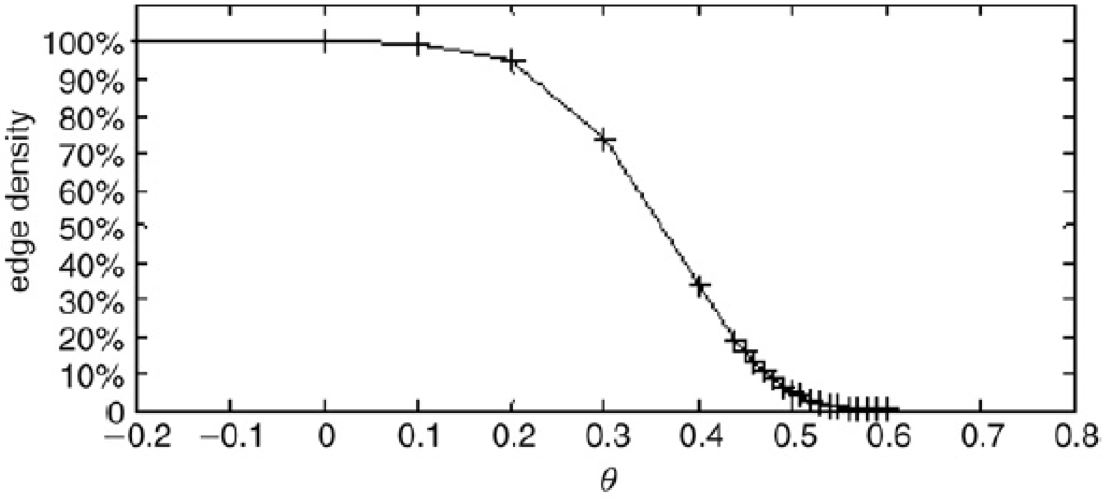


Figure 3 - Edge density for different values of correlation threshold (Huang et al., 2009)

Additionally, the clustering coefficient demonstrated that the stock correlation network presented to be highly clustered when the value of the threshold value increases and even when the edge density decreases. In this scope, the stock correlation network presents to be highly clustered and the statistical analysis of the degree distribution has shown that the power-law model, which is a common network of many real-life massive networks such as the Internet, cell networks and telephone networks, etc., is also applied to the financial market (Huang et al., 2009).

A **power-law model**, also called a scale-free network, are centralized networks where most of the nodes are characterized with a small degree and only a few have a large degree. In the sense of price fluctuation, the larger a stock's degree is, the more stocks it is correlated. From this point of view, stocks characterized by the high degree distributions, also called the "Hub" nodes, are the ones that have a "stronger market influence" and reflect the market behaviors (Huang et al., 2009).

Another example of the network analysis in the financial market was study by (Namaki et al., 2011) using 325 stock prices of Tehran Stock Exchange (TSE) market and analyzing 1291 daily changes. The authors normalized the price change of stocks to standardize the different stock volatilities and compute the cross-correlation matrix C. For a better risk management and to remove the randomness in the measured cross-correlation, an application of **Principal Components Analysis** (PCA) showed the eigenvector with most of the components, in other words, the one that describes the common behavior of all stocks. Based on this analysis, this eigenvector is considered the market mode which effect overwhelms the remaining relationships between the stocks, and for that reason, was removed correlation matrix (Namaki et al., 2011). **Figure 4** shows the effect of pricing before and after removing the market mode.

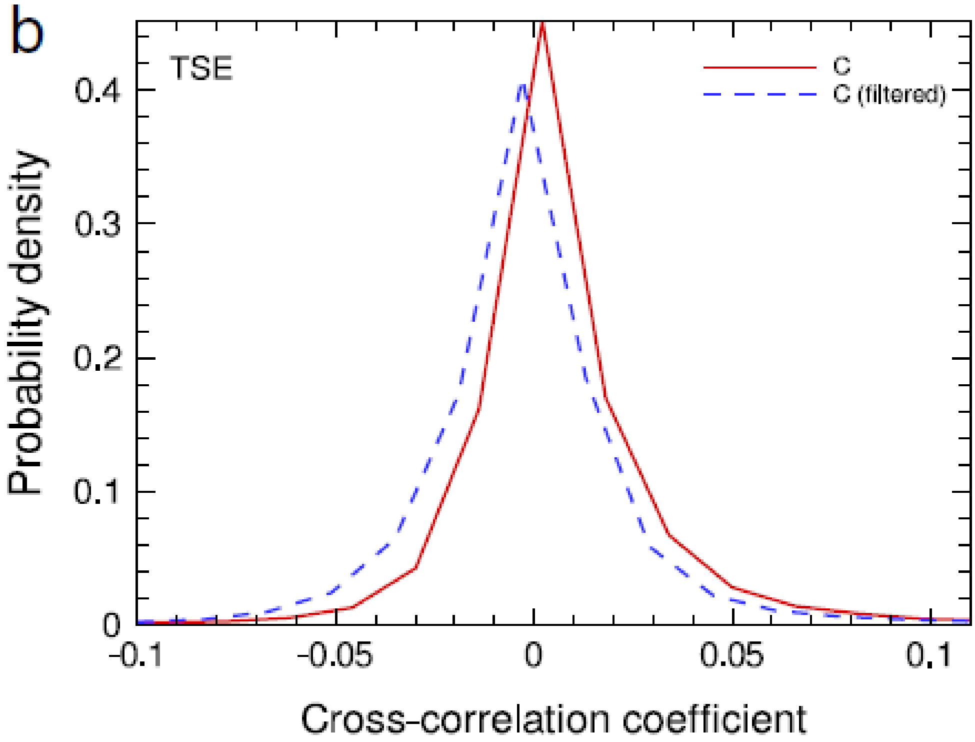


Figure 4 - PCA for TSE before and after removing the market mode (Namaki et al., 2011)

Furthermore, the correlation threshold method showed that after removing this effect, the mean of the resulted matrix was smaller when compared to the average of matrix C. This means that the high correlations of the matrix C were a result of the market effect (Namaki et al., 2011). To conclude, this study also evaluated the degree distribution and the edge density for different threshold values which showed that the TSE correlation network, as many others complex systems, also obeys to the power-law model where most of the stocks are at the same level and only a few have higher influence in the market. The clustering coefficient was also considered and the results revel a highly clustered network when the θ ∈ [0.02, 0.31] (Namaki et al., 2011).

Finally, in the context of financial globalization, daily data of 51 stock indices which covered four regions: America, Europe, Asia Pacific, and Africa, was analyzed by (Yin et al., 2017) through an intensive study of the relationships between international stock market. After transforming the stock indices into log return, the correlations were computed using the DCC-MVGARCH proposed by Engle (2002), a model used to capture volatility and predict future volatilities. As mentioned by the authors, the main characteristic of the dynamic conditional correlation (DCC) model is that the correlation coefficient is time-varying, being widely used on the field of market volatility (Yin et al., 2017)

To simplify the network structure, the correlation threshold method was applied and the relative changes of the nodes for different thresholds was analyzed. This approach chooses the first jump point as a perfect threshold of the network, market as "*" in **Figure 5**, with the value 0.48. Based on this network, the network density and average path length reflected the volatility of the major economic crises and the process of crises recovery from 2005 to 2016 (Yin et al., 2017). In addition, to explore the market changes, the authors used the **Minimum-cost Spanning Tree (MST)** to simplify the network and interpreted the dynamic evolution information without a threshold set. From this output, it was clear the volatile nature of the network and four stages were identified such as the subprime mortgage crisis, the global financial crisis, the European debt crisis and the strong market intervention policies such as Quantitative Easing (QEs) (Yin et al., 2017).

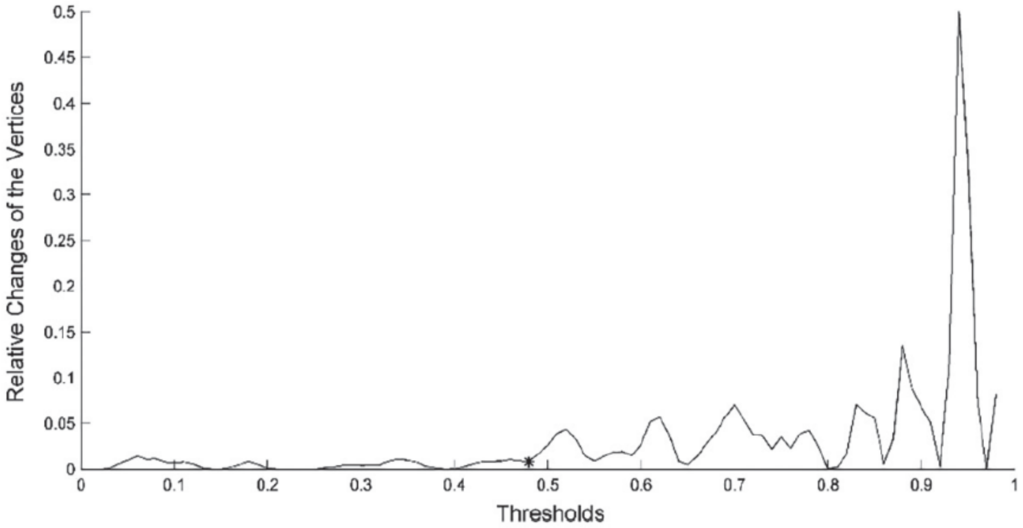


Figure 5 - Relative changes of vertices for different thresholds (Yin et al., 2017)

2.2. THE SCIENCE OF NETWORKS

The network analysis is a data mining technique that allow us to study complex real-world problems and understand the complexity behind datasets with high-dimensional data. This field of analysis is useful to represent many real-world applications such as social networks (Alamsyah et al., 2014), stock markets (Fernando & Dias, 2015), health psychology (Hevey, 2018), among others.

The process provides the capability to observe a group of objects/entities (called nodes) and the relationship between them (called edges). A natural way to represent these links is using a graph, also called a network, that provides deeper insights and helps easily visualize the connectivity between business entities.

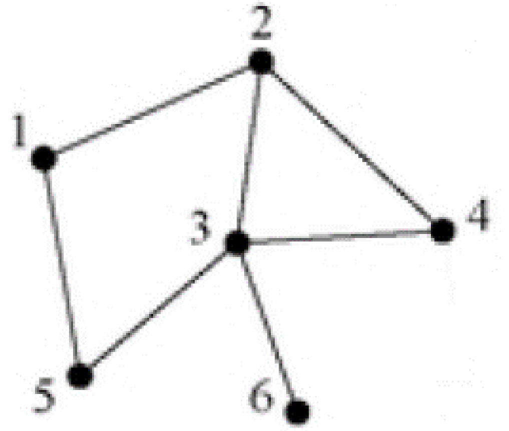


Figure 6 - Sample Network with 6 nodes and 7 edges (Newman & Girvan, 2004)

The **Figure 6** is a simple network, formally defined as G = (V, E), where V is the set of nodes/entities and the E represents the edges/links between them. The nodes and edges are defined as follows:

$$V = \{1, 2, 3, 4, 5, 6\} \qquad (1)$$

$$E = \{(1,2), (1,5), (2,3), (2,4), (3,4), (3,5), (3,6)\} \qquad (2)$$

We can represent different types of information in a network based on the scope of the analysis. This granularity provides details about the linked data which is applied optionally and according to the case study.

The edges can represent not only information about the relationship between the nodes but also the **direction** and **strength** of the link. For example, if we want to deep study the connection between two entities, the associated edge can be weighted, which represents the strength of the related connection. A higher weight will represent a strong relationship which is also associated with a lower distance between the nodes. The edges can also be directed or undirected: directed networks, representing one-way effect of the edges, and undirected networks indicating mutual relationship but without a direction effect (Hevey, 2018).

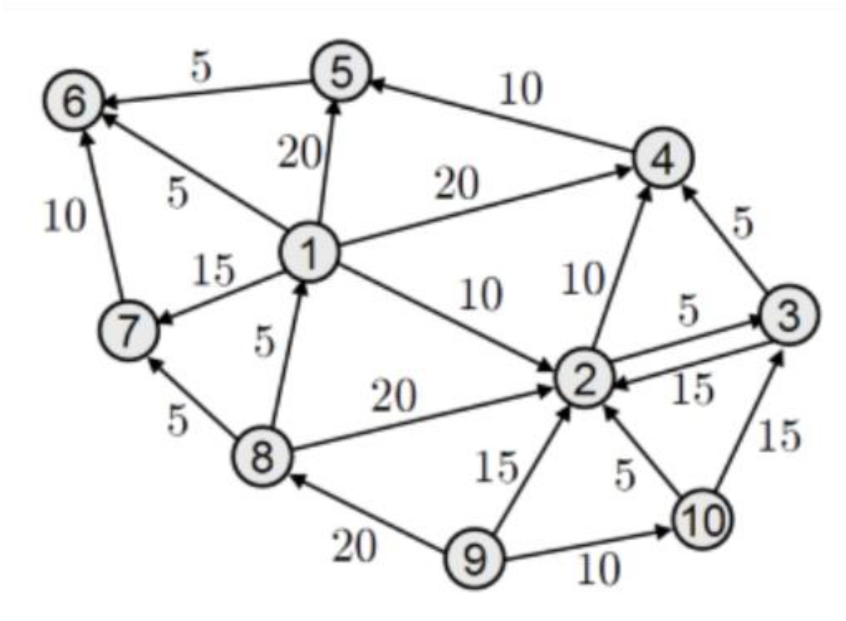


Figure 7 - Directed Network with weighted edges (Horwood, 2015)

The **type** of relationship is may also be represented graphically using different colored lines: positive relationships (e.g. positive correlation/covariance) are typically colored blue or green, and negative relationships (e.g. negative correlation/covariance) are colored red (Hevey, 2018).

Additionally, and depending on the point of view, the networks can be analyzed at a group or individual level. For example, getting a deep dive of the relationships across stock categories (Newman & Girvan, 2004), or if group data revel independent relationships, switch from the analysis of a group to an individual level, and extract deeper insights of specific individuals over time. Furthermore, we can also use networks to compare different populations (Hevey, 2018).

2.2.1. Graph Structures

According to graphs theory, the relationship between the nodes can take many shapes. The most representative network types are known as: Random networks; Small-world networks and Scale-free networks. **Figure 8** shows the different graphic designs associated to each network type.

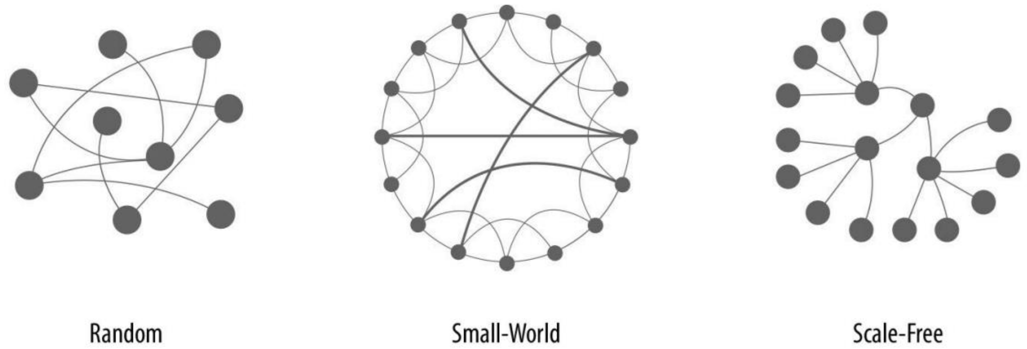


Figure 8 - Examples of the most representative network types (Random network, Small-World network and Scale-Free network) – Adapted from (Hodler et al., 2019)

Random Networks

Traditionally, random networks can be described by probability distributions. In the random graph literature, it is often assumed that each node pair relates to the same probability "p", without creating any hierarchies. The higher the probability, the more connected the graph will be. The degree distribution of these networks follows a normal distribution, with a few nodes with a very low degree and very few nodes with a very high degree. Additionally, the Random Networks generate no clustering, constrained by no-limited rules that may give a distinct pattern or behavior.

Small-World Networks

A small-world network is a type of mathematical graph which is most common in social networks. The network structure shows localized nodes that ensure whole the nodes are connected, even if most of them are not connected to each other. Typically, they are formed by an over-abundance of nodes with a high number of connections, called hubs, and the mean-shortest path length tends to be small.

Scale-Free Networks

Real-world networks are often claimed to be scale free, meaning that an underlying structure is preserved, such as in World Wide Web. The distribution of the number of the nodes follows the power-law distributions, indicating that these networks are formed by a small number of very highly connected nodes.

2.2.2. Methodology

In the retail market, the node can represent a specific product or category (group of products), while edges represent the price changes effect between two products/categories. To represent different populations, the nodes can also be categorized by retailer.

The definition of the network starts by estimating a statistical parameter that quantifies relationships between the entities, e.g. correlations, covariances, partial correlations, among others. In the health psychology field, the most common approach is **partial correlations** to control the effects of confounding interpretations when analyzing different variables (Hevey, 2018). In order words, the relationship of two independent nodes can be a explained by an external variable, creating a causal relationship between them.

In the stock markets, cross-correlations are considered, and the correlation coefficient is computed for all the possible pairs of stocks in a given time period (Smith, 2009). For any period t, we can use Pearson's correlation coefficient, the statistic test that measures the degree of linear correlation between two elements (x_i and x_j), defined as:

$$\rho_{ij} = \frac{\sum_t \left[(x_i(t) - \bar{x}_i)(x_j(t) - \bar{x}_j)\right]}{\sqrt{\sum_t (x_i(t) - \bar{x}_i)^2} \sqrt{\sum_t \left(x_j(t) - \bar{x}_j\right)^2}} \tag{3}$$

Using this matrix and to represent the network, different approaches are used to visualize the patterns and relationships present in the data, for example the **Maximum-cost Spanning Tree** (MST) and/or the **correlation threshold method**.

The **maximum-cost spanning** tree generates a minimal weighted undirected **sub-graph** of G=(V, E), where *w(u, v)* is the weight of edge **(u, v)** (Yin et al., 2017). The network is expressed as:

$$\min w(T) = \sum_{(u,v)\in T} w(u,v), \tag{4}$$

To find the maximum spanning tree, the most known algorithms are Prim and Kruskal, where both starts by an empty spanning tree. The Prim algorithm creates the spanning tree by a start position (**node** with the small weight) and adds the remaining nodes one by one, without creating any cycles, which means that if the next node with the smallest weight forms a cycle, will be discarded. The Kruskal algorithm uses the **edges** to create the spanning tree, starting by sorting the edges based on their weights and builds the tree by adding edges one by one, starting on the smallest until the largest weight, also without creating any cycles (HackerEarth, s.d.). The weight of a spanning tree results in the sum of all the weights given to each edge. The choice of the algorithm can be considered based on the total of nodes and edges presented in the network, which will impact the performance of each algorithm: Prim's algorithm can be adopted when a graph represents a lot of edges and Kruskal's algorithm when a graph represents a lot of nodes.

These models provide the "hierarchical" split of the prices (nodes), showing them in order of distances (weights), calculated based on their correlations (Giudici & Polinesi, 2021).

The **threshold method** removes the weak correlation by defining a static correlation coefficient network threshold (θ), for example θ > 0.7, and consequently remove all the values below the threshold specified. The network is simplified for a better understanding and different levels of threshold can be considered to study and identify the perfect threshold for the network (Yin et al., 2017). Different values of the threshold can be analyzed based for example in the evolution of edge density for different values of the correlation thresholds and/or based on the number of edges presented in the network with the respect threshold (Huang et al., 2009).

2.2.3. Evaluation Measures

The networks can also be evaluated using features that studies the network health and the importance of a node. Some of the important tools to study the networks and discard the noise data are called the centrality measures, namely: degree, betweenness, closeness and eigenvector.

Degree centrality evaluates the nodes that have many connections. In other words, measures the total of edges that are connected to a single node. From the point of view of price fluctuation, the larger the degree is, the stronger is the influence in the market (Huang et al., 2009).

Betweenness centrality is a way of detecting the amount of influence a node has over the flow of information or resources in a graph (Hodler et al., 2019). Typically, the nodes with higher betweenness are called the 'bridges' between nodes that ensure a connected network.

Closeness centrality shows the distance of a node to other nodes by calculating the shortest paths between all pairs of nodes. Nodes with a high closeness score have the shortest distances from all other nodes (Hodler et al., 2019).

Eigenvector centrality is a more elaborated version of degree centrality. This measure considers not the quantity of neighbors associated to a node but the quality of the neighbors by considering the centralities the neighbors (Fernando & Dias, 2015). In this context, the node's importance is assumed by the very well-connected neighbors.

3. FROM DATA TO NETWORKS

In this section, we describe the data set and the pipeline used for pre-processing tasks and network inference. The pipeline was used to analyze price behaviors at the product and category level. We conclude with a descriptive analysis of several stylized facts.

3.1. DATA, COLLECTION AND PROCESSING

In this subsection we describe the data set and identify the pre-processing tasks to prepare the working data set for the network analysis. In the aimed or our case study, we started with the exploratory analysis of the products and created versions of our data set by category level.

The data set consists of daily prices of products from the online webstore of one of the largest supermarket chains in Portugal. The data was obtained through web-scrapping and covers the period between February 2020 to March of 2021. In total we have more than 4 million records with information about 12.714 products. Besides the price, additional information such as the product code (EAN), brand, package attributes (size, price, and capacity), and discount attributes (value, type, discount, and price per capacity after the discount) were also extracted.

A first overview of the data shows us a wide variety of categorizations and anomalous behavior in the time series, linked to gaps in the information and missing records for many products. These gaps are the result of operational issues during the data extraction that compromised the information available. **Figure 9** shows the frequency distribution of products in each time interval of 15 days, and **Figure 10** shows the number of products with observations per week of the year.

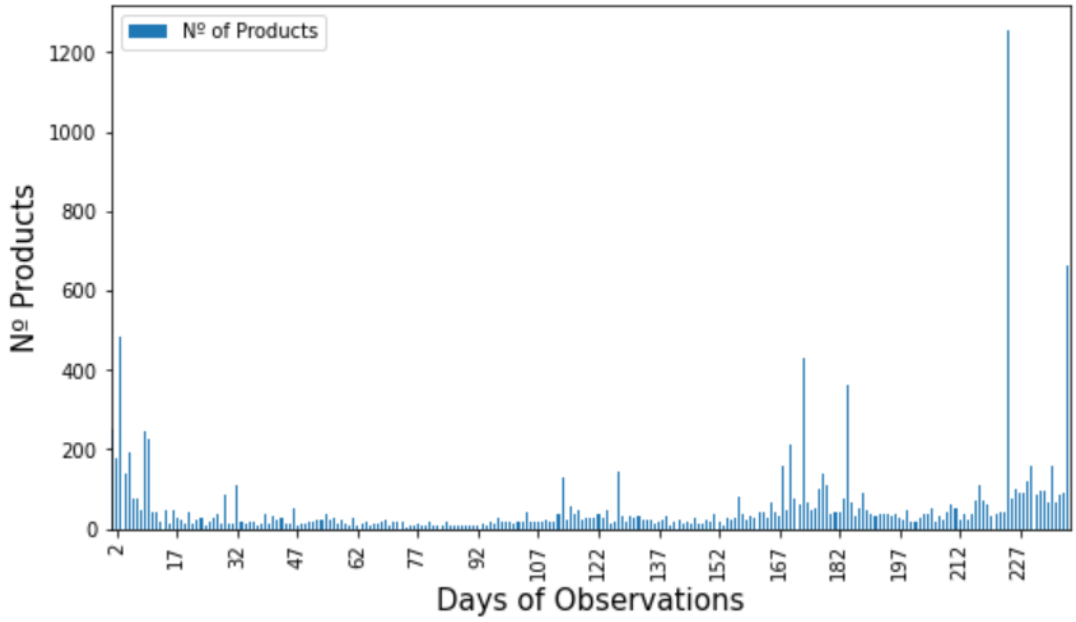


Figure 9 - Frequency distribution of products in each time interval of 15 days

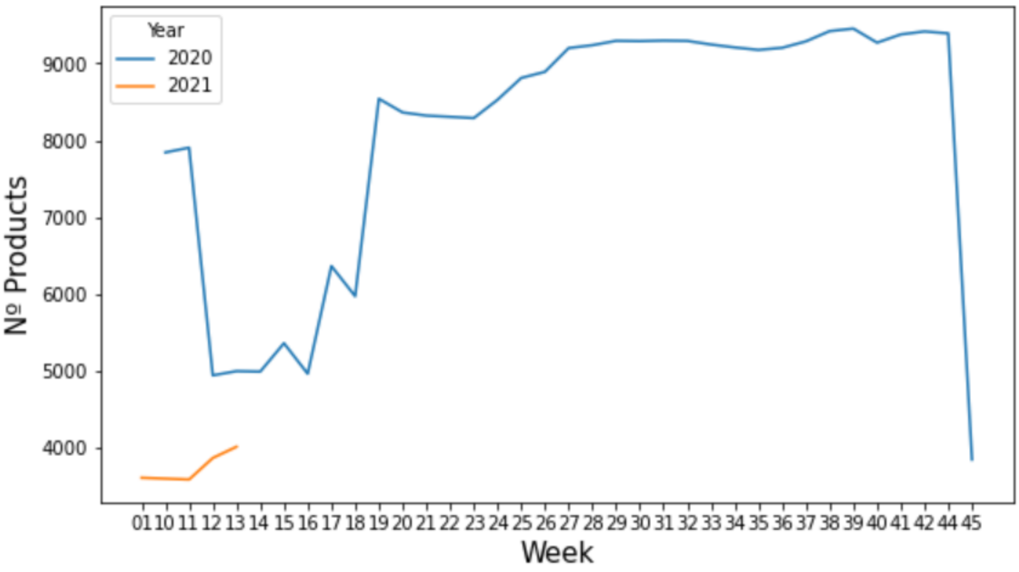


Figure 10 - Number of products with price observations per week of the year

We note an irregular distribution of the observations per product and a decrease in the total number of records after November 2020. Ideally, for each product, the data set would inform of a time-series containing the prices for the 239 days of the time-window of study. In this scenario, missing data can be attributed to products being removed from the webstore or possible changes in the product description on the website that compromised the web-scraping extraction.

To study in more details the time-window of the price observations in study, the **Figure 11** shows an example of the temporal variations of prices for a few representative products.

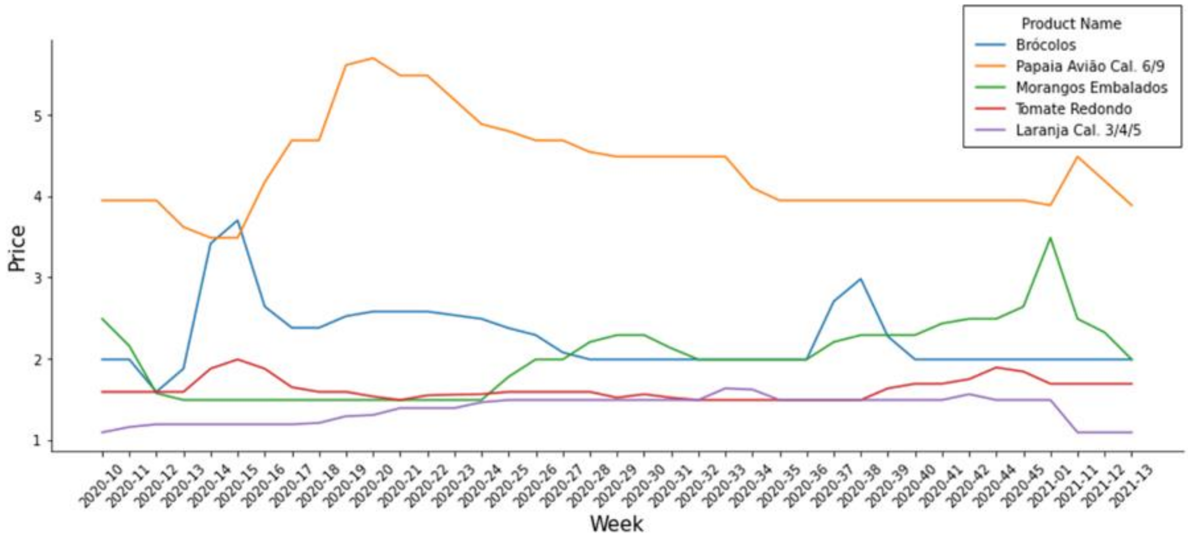


Figure 11 - Temporal variations of prices for a few representative products, covering the period between February 2020 to March of 2021. The price values were obtained from the online webstore of the supermarket chain, using Web Scraper API.

We performed the same exploratory analysis at the categorical level. The product categorization is related to the classification used by the retailer to structure the website and for a better shopping experience. The categorization is made of three hierarchical levels, where the level one represents the most basic classification to the products. Each product is classified according to its characteristics and it's linked with one single classification, while each classification can be associated with one or more products. The data set is composed by a total of 158 categories of level one, 821 categories of level two, and 542 categories of level three.

Figure 12 shows an example of two level one categories and its hierarchical distribution. For a better detail of the product classification, a category list is available in the **Appendix 1.**

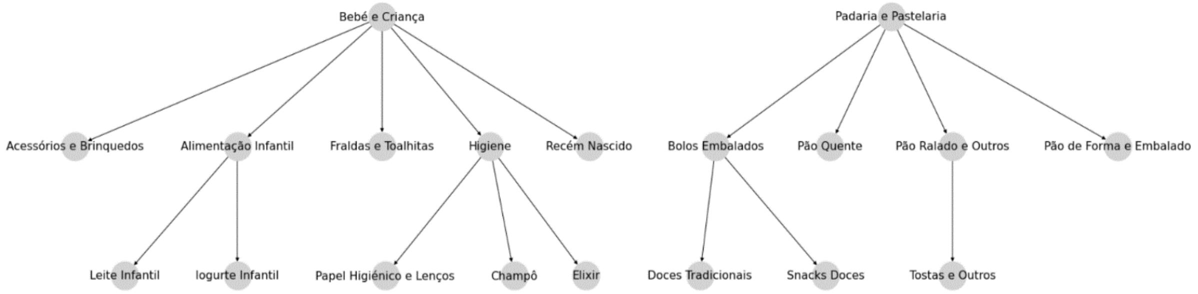


Figure 12 - Sample of the two level one categories, using a tree structure from a graph. The category levels are extracted from the website and reflects the retailer categorization of the products used on the website.

We were also able to identify anomalous behavior in the time-series of the product hierarchies and an irregular distribution of the observations, result of the missing data observed previously in the product analysis.

To clean the data from incomplete/sparse observations, we have performed a set of steps to prepare a working data set. As a first step, we discarded all weeks with data gaps. In step two, we considered a threshold of 3 daily price observations per week and removed the products without week price observations from the time window considered in step one. Finally, we are interested in capturing the information that might be revealed by the common variation in price between products. In that sense, we calculated the average price and computed the temporal week-to-week price variation per product. Hence, for each product i at time t {1, 2, …, N-1, N} we compute

$$r_i(t) = P_i(t) - P_i(t - \Delta t) \tag{5}$$

where P_i (t) is the price of product i at time t (Namaki et al., 2011). Based on this output, we discarded the products without price changes

Table 1 summarizes the above-mentioned steps, where it is also displayed the number of products remained in the data set after that specific pre-processing step along with the fraction of products affected. We also considered the same data processing tasks for all three categorical levels.

Table 1 - Summary of the data processing steps and the impact of each in the size of the number of products in the Data Set. Each Step was implemented in a sequential order (from top to bottom). The steps show how the number of products were filtered from raw data to the working data set.

Step	Description	Number of Products	Fraction of Products
Initial	**Raw Data Set**	**12.714**	
1	Drop Observations with data gaps	11.065	1.649
2	Remove incomplete data	10.711	354
3	Remove products without price changes	766	9.945
Final	**Working Data Set**	**766**	

In doing so, we removed 11948 products while accounting with 24 consecutive trading weeks of price observations, covering 7 months of information about 766 products. The final data set (e.g. the ones to be used in the network analysis) covered the period between May 2020 to November of 2020. **Figure 13** shows the distribution of prices among products, transformed using the logarithm function, while **Figure 14** shows the predictions for quantiles 0.25, 0.5 and 0.75 and actual values represented by the average price (Avg Price) per week.

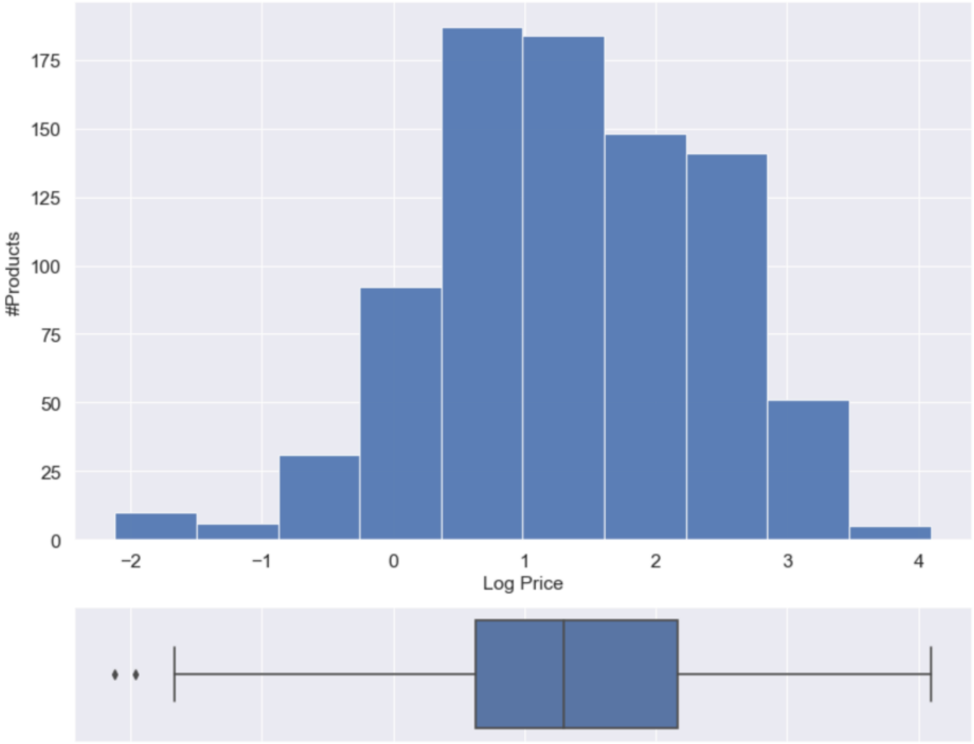


Figure 13 - A histogram in which distribution of the average price, transformed using the natural logarithm, is plotted on the vertical axis against the total of products on the horizontal axis.

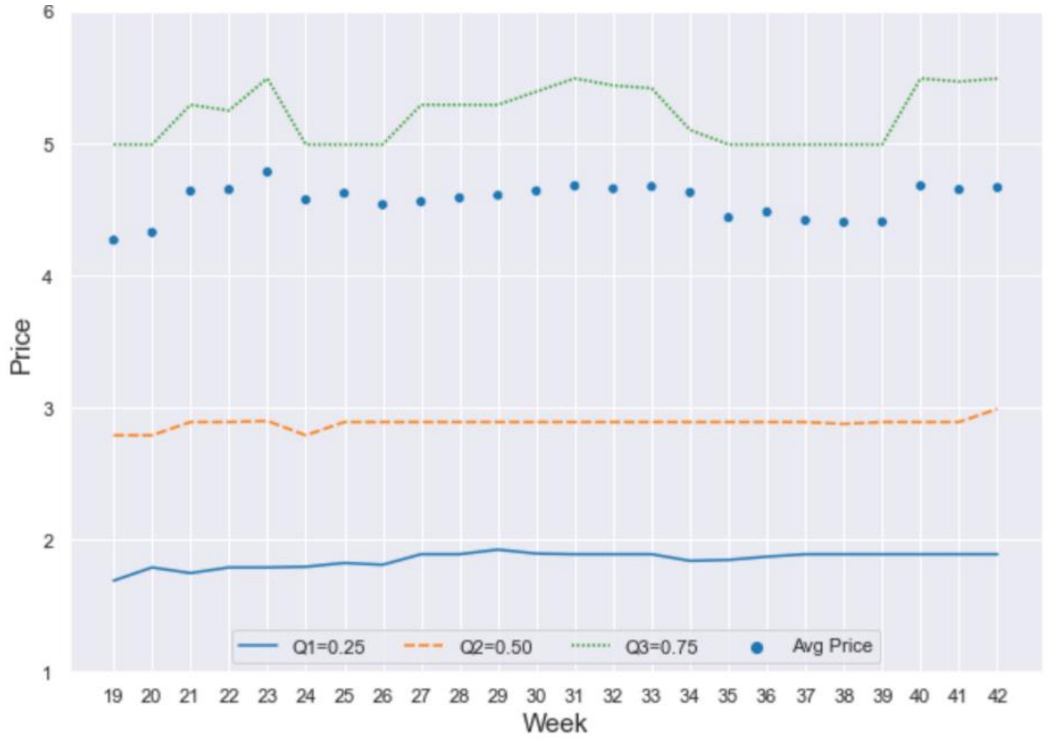


Figure 14 - Quartiles Distribution with 95% Confidence Limits for the quantiles 0.25, 0.5 and 0.75 and actual values. The actual values are represented by the average price (Avg Price) per week

We can observe a shape of a normally distributed set of the average price among products, which suggests that the major products share the same scale of price values, measured with the official Eurozone currency (euro). The quartiles also show us that the major products follow the average price of the whole data set, with a few representative products measured in a lower scale of values

3.2. PRICE CORRELATIONS

In this subsection, we aimed to calculate the relationships between products/categories, based on the correlation coefficient between their weekly price changes. Following the same approach from to study the stock markets, the networks will be constructed based on the price correlations to understand correlation patterns among products/categories.

We use the temporal week-to-week price variation calculated in the step 3 of the data processing procedure and calculated the pairwise correlations between the price evolutions of all pairs of products, say *i* and *j*. To that end, we use the standard Pearson Correlation Coefficient, which is given by the following formula:

$$\rho_{ij} = \frac{\sum(r_i - \mu_i)(r_j - \mu_j)}{\sqrt{\sum(r_i - \mu_i)^2 \sum(r_j - \mu_j)^2}} \tag{6}$$

Where μ_i represents the average of the price variations for product i and μ_j represents the average of the price variation for product j. The pairwise correlation ρ_{ij} can vary between [-1, 1] where ρ_{ij}= 1 or ρ_{ij} = -1 means that two products are perfect correlated or anti-correlated and ρ_{ij} = 0 means that they are uncorrelated (Huang et al., 2009).

The value of the correlation depends on the length of the time series that is used in the study. In smaller time series, one correlation has a higher importance when compared with larger series (Fernando & Dias, 2015). Due to the reduced number of liable data, we consider a hypothesis test of the significance of the Correlation Coefficient to decide if the correlation between the product i and j is strong enough to use for modelling. We discard relationships that fail at a significance level of 5% (α = 0.05).

3.3. NETWORK INFERENCE & CHARACTERIZATION

In this subsection, we describe the steps performed for the network inference, where each product/category is represented by a node in the network and the correlation between products/categories (nodes) are represented by a link.

To perform the network analysis, we used different samples (data sets) that went through the same data pre-processing steps described in **Table 1** of Subsection 3.1. Additionally, we considered that the networks are **undirected** as for our case study the correlation between products/categories is the same in both directions; and **weighted** by the value of the correlation between two products/categories.

To start, the matrix of correlations was converted in an edge list, containing the information for the correlation between each pair of products/categories, with both positive and negative correlations (edges). From this list, we removed the self-correlations and performed a network analysis to study individually the positive and negative correlations. It is possible that certain economic or political events cause the increase of the price of some products, and at the same time the decrease of other product prices, as an event can be positive for some companies, and at the same time negative to others (Fernando & Dias, 2015). In our context, we considered only positive relationships between products/categories.

For each analysis, we created an undirected network G = (V, E), where V represents the products/categories (nodes), and E represents the connections (edges). Furthermore, we extracted the maximum spanning tree and created a sub-graph G', using the Kruskal’s algorithm to find the largest edge and form a sub-graph where no cycles (or loops) are formed. A maximum spanning tree (MST) or maximum weight spanning tree is a tree composed of all the vertices and some (or perhaps all) of the edges of G (Helmi et al., 2012). The MST model provide us more insights by splitting the products in “hierarchical” way and show them in order of distances (weights), calculated from their correlations (Giudici & Polinesi, 2021).

To enrich and reach a network with an average degree four, we considered a sample of the relevant correlations, based on edge list, and combined with the sub-graph G'. The result not only provide the MST, but also shows additional relationships that are relevant for our study. To visualize the network, we used the ‘Fruchterman-Reingold’ algorithm which sets the positions of the nodes using the cost function, allowing the quick identification of groups of products with similar properties (West 2019).

Table 2 summarizes the above-mentioned steps for network representation, where it is also displayed the number of edges for each network (Products + Category Level 1 + Category Level 2 + Category Level 3) after that specific step.

Table 2 - Considered tasks for network representation along with the number of edges resulted for each network (Products + Category levels). Each task was implemented in a sequential order (from top to bottom). The total of relevant edges added is identified in brackets after the edges value.

	Task	Description	Product Edges	Level 1 Edges	Level 2 Edges	Level 3 Edges
1	Create the undirected network	Representation of the correlation matrix without self-correlations	103.457	74	2.240	1.395
2	Extract Maximum Spanning Tree	Extract the sub-graph with the largest edge	764	14	105	89
3	Add relevant links	Composed graph based on the MST and further links from the correlation matrix	1.563 (+799)	39 (+25)	215 (+110)	188 (+99)

To study the network structure, we calculated four measures: density, average degree, average path length and cluster coefficient. **Table 3** summarizes the main network metrics that characterize the networks for each data set. Existing literatures, particularly in the stock market, focus on basic topological properties and structure characteristics of the network as an important factor for portfolio selection and risk management (Zhuang & Xiu, 2015).

Table 3 - Network analysis metrics for Products + Category levels classification, including the number of nodes in the network, the network density, the average degree, and average path length and the cluster coefficient

	Nodes	Density	Average Degree	Average Path Length	Cluster Coefficient
Products	765	0.35	270.47	1.65	0.19
Level 1	19	0.43	7.78	1.58	0.47
Level 2	106	0.40	42.26	1.60	0.44
Level 3	90	0.34	31.0	1.65	0.41

The average degree is higher in the products network and level 2 categories, which means that these networks registered a higher number of common price behaviors in the period of study. These networks exhibit a very connected structure, mostly related to the diversification of nodes of the networks when compared to level 1 and level 3 categories.

The clustering coefficient measures the degree to which nodes in a graph tend to cluster together. The output results suggest that the graph tends to form two communities when the products are aggregated by category. Such values lead us to a better understanding of the price behaviors by category and support the complexity associated with the product level/detail.

The average path length, which measures the average shortest path between all pair of nodes in the network, was approximately two in all the data sets analyzed. This output highlights the two main players of the supermarket chain price fluctuations for each network. These two observations are supported through the analysis of the network centralities, namely: Degree, Betweenness, and Eigenvector, described in the network analysis subsections 3.3.1 and 3.3.2.

3.3.1. Products Network Analysis

This subsection provides the details of the network analysis for the data set focused on the Product level/detail. **Figure 15** represents the product network, based on the MST and further links from the correlation matrix, and **Figure 16** shows the degree distribution of the products. The network includes the identification of the most relevant products (nodes) with the highest degree (identified in brackets). The network results in a total of 765 nodes, which means that approximately 6% of the products presented signs price fluctuation during the period of analysis.

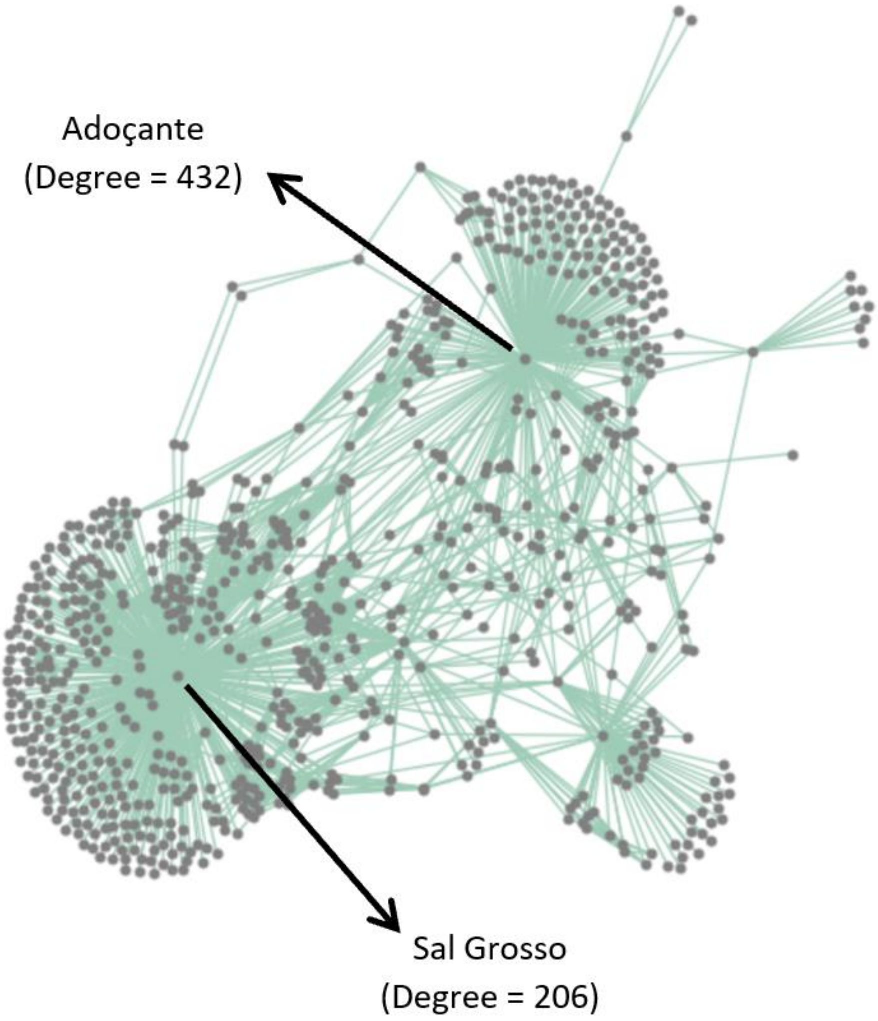


Figure 15 - Undirected and weighted product network. The strong relationships are associated with a lower distance between the nodes.

We observed that the influencer product, e.g., the product that whose price impacts the price of other products, is the sweetener (“Adoçante”), followed by the salt (“Sal Grosso”), defined as the group of vertices densely connected to the other products of the network. **Figure 16** supports this analysis and shows that majority of the product interact with these two nodes and such scenarios with highly central nodes might be more susceptible to a high risk to the entire price system.

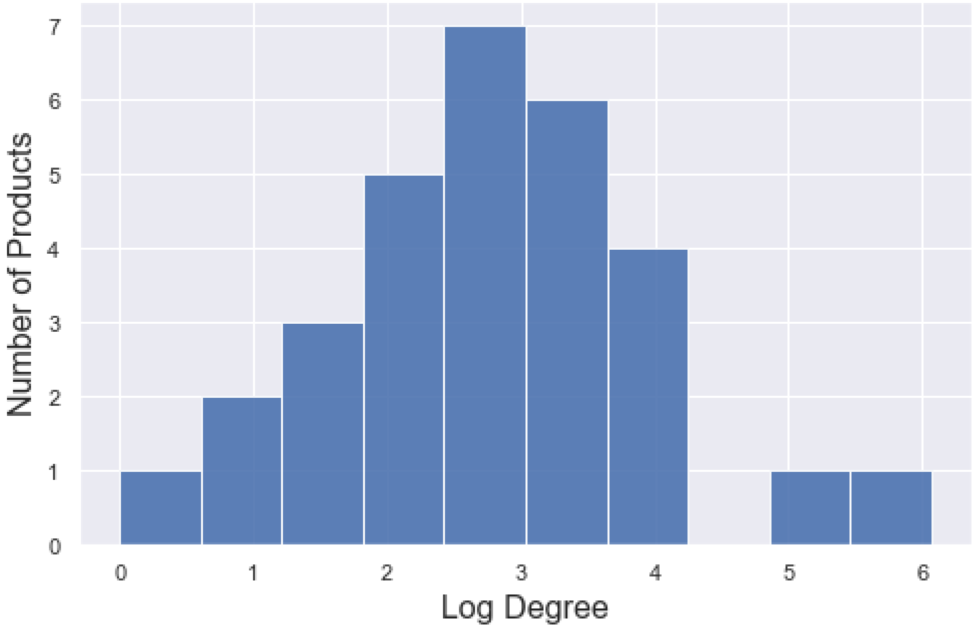


Figure 16 - Degree distribution, transformed using the natural logarithm, per number of Products.

The **Table 4** provides a better overview of the top 5 products per degree distribution. For a better understanding of the network topology, we added the analysis of other three centrality measures, namely: Betweenness (shows which nodes are 'bridges' between nodes in the network), Eigenvector (shows the influence of a node in the network), and Closeness (shows us the distance of a node to other nodes). We observed that the most impacted product in the network is the sweetener (“Adoçante”) with the highest centrality measures (degree, betweenness, eigenvector and closeness).

Table 4 - Top 5 products by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value.

	Product	**Degree**	**Betweenness**	**Eigenvector**	**Closeness**
1	Adoçante com Doseador Canderel	432	0.73 (1)	0.67 (1)	0.64 (1)
2	Sal Grosso	206	0.50 (2)	0.08 (4)	0.57 (2)
3	Mandarina	66	0.11 (3)	0.01 (445)	0.41 (6)
4	Desparasitante para Cão	48	0.01 (7)	0.08 (3)	0.39 (115)
5	Penso Diário Maxi	45	0.001 (51)	0.10 (2)	0.40 (100)

3.3.2. Categories Networks Analysis

This subsection provides details of the network analysis for the data set focused on the Category levels, e.g., level 1, level 2 and level 3, used for product classification. For this analysis, we aggregated our data set based on the category and calculated the average price per week.

Figure 17 represents the categorical networks, based on the MST and further links from the correlation matrix. The network includes the identification of the most relevant categories (nodes) with the highest degree based on the size of the node.

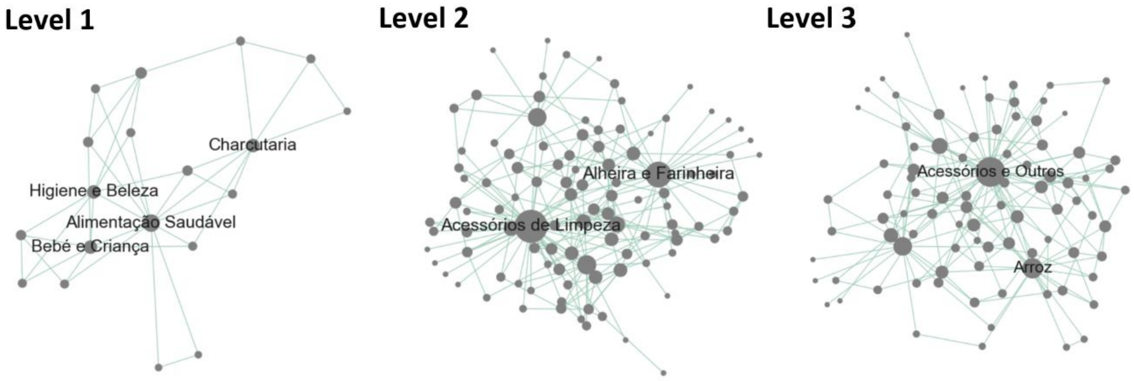


Figure 17 - Undirected and weighted category networks (Category Level 1 + Category Level 2 + Category Level 3). The size of each node is proportional to the number of strong correlations with other nodes (degree), and the strong relationships are associated with a lower distance between the nodes.

We observed the random network structure among category levels, when using the same data set and time window. **Tables 5, 6 and 7** provides a closer look to the top 3 categories for each level, by degree distribution, along with the analysis of other three centrality measures: Betweenness (shows which nodes are 'bridges' between nodes in the network), Eigenvector (shows the influence of a node in the network), and Closeness (shows us the distance of a node to other nodes).

We observed that the most impacted categories in the Level 1 network is Healthy Eating ("Alimentação Suadável") with the highest centrality measures (degree, betweenness, eigenvector and closeness). Looking to the Level 2 network, the impacted category is Cleaning Accessories ("Acessórios de Limpeza") with the highest centrality measures and on the Level 3 network, we identified the category Accessories and Others ("Acessórios e Outros") as the main player of the network and subcategory of Cleaning Accessories.

Table 5 - Top 3 **Level 1 categories** by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value.

	Category	Degree	Betweenness	Eigenvector	Closeness
1	Alimentação Saudável	12	0.52 (1)	0.49 (1)	0.75 (1)
2	Charcutaria	7	0.23 (2)	0.26 (4)	0.58 (2)
3	Higiene e Beleza	7	0.14 (3)	0.35 (2)	0.58 (3)

Table 6 - Top 3 **Level 2 categories** by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value.

	Category	Degree	Betweenness	Eigenvector	Closeness
1	Acessórios de Limpeza	43	0.62 (1)	0.58 (1)	0.63 (1)
2	Alheira e Farinheira	26	0.30 (2)	0.21 (2)	0.51 (2)
3	Café e Chá	14	0.08 (5)	0.20 (3)	0.44 (4)

Table 7 - Top 3 **Level 3 categories** by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value.

	Category	Degree	Betweenness	Eigenvector	Closeness
1	Acessórios e Outros	39	0.61 (1)	0.58 (1)	0.64 (1)
2	Arroz	18	0.21 (2)	0.21 (2)	0.50 (2)
3	Azeitonas e Tremoços	15	0.17 (3)	0.18 (3)	0.47 (3)

Additionally, is possible to identify a mismatch between the categories network and the product network, considering the classification of the impacted product analyzed in the section 5.3.1 (“Adoçante” with the classification “Merciaria; Açúcar, Farinha e Ovos”). The **Table 8** shows the categorization of the impacted product per centrality measure and its position in the category levels networks (Level 1 + Level 2). The position is identified in brackets after the centrality value.

Table 8 - Categorization measures of the impacted product observed in products network ("Adoçante"). The position is identified in brackets after the centrality value.

	Network	Category Name	Degree	Betweenness	Eigenvector	Closeness
1	Category Level 1	Merciaria	3 (10)	0.01 (8)	0.19 (9)	0.49 (9)
2	Category Level 2	Açúcar, Farinha e Ovos	4 (31)	0.01 (16)	0.06 (53)	0.38 (46)

4. BRAND PRICING ANALYSIS

In this section, we aimed to study the price behaviors using a product categorization at the brand level. To do so, we aggregated the price observations by brand and performed an exploratory analysis.

The raw data counts with a total of 1.377 brands, covering the period between February 2020 to March of 2021. To study in more details the time-window of the price observations by brand, the **Figure 18** shows an example of the temporal variations of prices for a few representative brands.

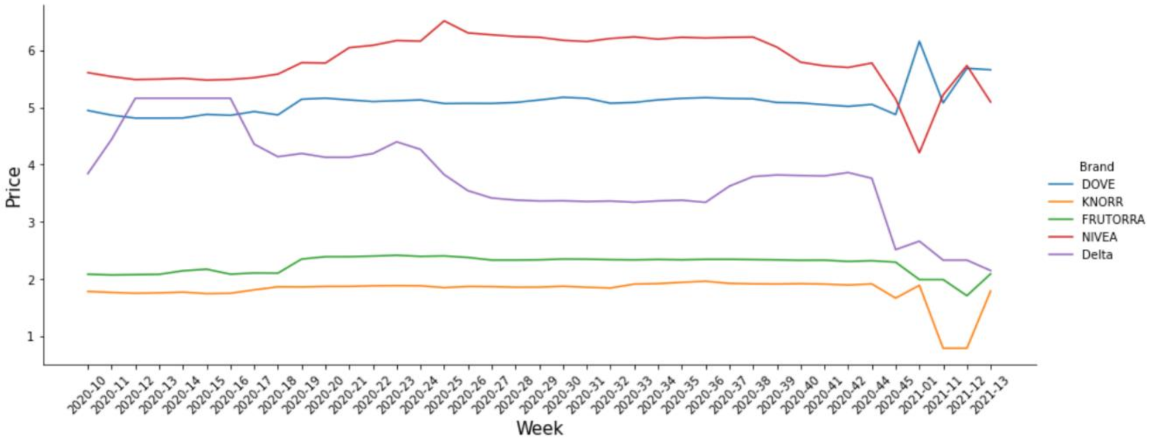


Figure 18 - Temporal variations of prices for a few representative brands, covering the period between February 2020 to March of 2021. The brand information was obtained from the online webstore of the supermarket chain, using Web Scraper API.

We can observe a different scale of values of the average price among brands, which suggests that a set of brands are premium, presenting higher average prices, e.g. Nivea and Dove, when compared to other brands like Knorr, presenting a lower average price.

To observe price behaviors, our working data set results of the same processing steps performed in **Table 1**. In doing so, we removed 455 brands while accounting with 25 consecutive trading weeks of price observations, covering 7 months of information. To perform the network analysis, we followed the same approach used in the products and categories network, using the correlation coefficient to study price relationships and the tasks described in **Table 2** for network representation.

The network results in a total of 608 nodes, which means that approximately 44% of the brands are significantly relevant and presented signs price fluctuation during the period of analysis. To study the network structure, we calculated the same four measures used in the Chapter 3: density, average degree, average path length and cluster coefficient. **Table 9** summarizes the main network metrics that characterize the network, along with the number of nodes.

Table 9 - Network analysis metrics for Brand classification, including the number of nodes in the network, the density, the average degree, and average path length and the cluster coefficient

	Nodes	Density	Average Degree	Average Path Length	Cluster Coefficient
Brand	608	0.41	251.68	1.59	0.44

The average degree shows a similar structure presented by the products network, which means that this network is highly connected and registered a higher number of common price behaviors in the period of study.

The average path length, which measures the average shortest path between all pair of nodes in the network, was approximately two, suggesting the presence of two possible main players of the network. These observations are supported through the analysis of the network centralities, namely: Degree, Betweenness, and Eigenvector, described in the **Table 10.**

Figure 19 represents the brand network, based on the MST and further links from the correlation matrix, and **Figure 20** shows the degree distribution of the brands. Additionally, the network includes the identification of the most relevant brand (node) with the highest degree (258).

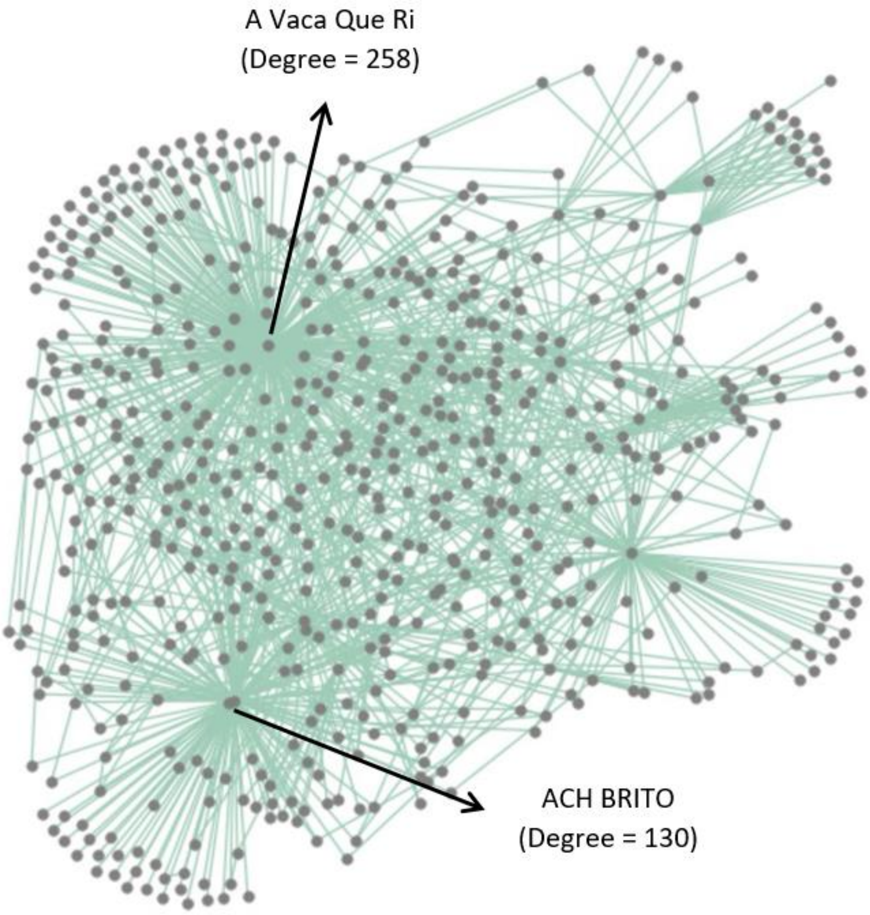


Figure 19 - Undirected and weighted brand network. The strong relationships are associated with a lower distance between the nodes.

We observed that the influencer brand, e.g., the brand that whose price impacts the price of other brands, is “A Vaca que Ri”, a cheese brand born in France. This result is also supported through the analysis of the network centralities, namely: Degree, Betweenness, and Eigenvector, described in the **Table 10**.

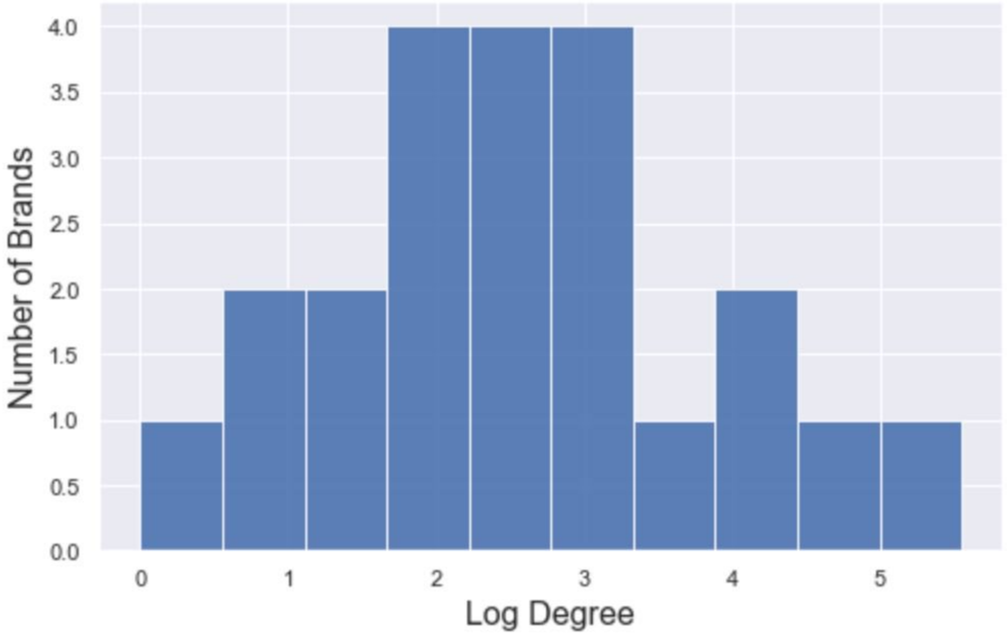


Figure 20 - Degree distribution, transformed using the natural logarithm, per number of Brands

The **Table 10** provides a better overview of the top 5 brands per degree distribution. For a better understanding of the network topology, we added the analysis of other three centrality measures, namely: Betweenness (shows which nodes are 'bridges' between nodes in the network), Eigenvector (shows the influence of a node in the network), and Closeness (shows us the distance of a node to other nodes). We observed that the most impacted brand in the network is “A VACA QUE RI”, followed by “ACH BRITO”, with the highest centrality measures (degree, betweenness, eigenvector and closeness).

Table 10 - Top 5 brands by degree distribution and its position for the betweenness, eigenvector and closeness. The position is identified in brackets after the centrality value

	Brand	Degree	Betweenness	Eigenvector	Closeness
1	A VACA QUE RI	258	0.74 (1)	0.67 (1)	0.63 (1)
2	ACH BRITO	130	0.28 (2)	0.12 (2)	0.49 (2)
3	AGROS	72	0.15 (3)	0.07 (3)	0.45 (3)
4	AFTER EIGHT	50	0.09 (4)	0.06 (4)	0.43 (4)
5	AIRWICK	36	0.05 (5)	0.06 (5)	0.42 (6)

We observed high connected structures when analyzing the products and brands network, and random structures when supporting our analysis with the product hierarchy to study the relationship between prices. The statistical analysis of the networks shows the complexity of the price system taking of evidence the random structure, generated by no-limited rules that, when considered may give a distinct pattern or behavior and offer a better understanding of market behavior during the period of study.

5. CONCLUSION

In this work, we explore the daily prices of products, extracted from the online webstore of one of the largest supermarket chains in Portugal, using the network science approach, defined as a data mining technique that allow us to study complex real-world problems. To better understand the network structure, we present detailed study of the properties of the pricing correlations networks used for products, categories, and brands.

We started by preparing the working data set with the implementation of pre-processing steps to clean up the data gaps identified in the time-series and filter out the observations with no price changes in the period study. To explore our data set, we used a weekly window, and considered a threshold of 3 daily price observations per week.

Following the literature used to support our study, we used the correlation coefficient to measure the relationships between products, and considered a hypothesis test of the significance of the correlation to consider only strong relationships, and discard those that fail at a significance level of 5% ($\alpha = 0.05$). For the network inference, we extracted Maximum Spanning Tree and added additional relationships that are relevant for our study.

Our analysis suggested that we are facing a very connected random network with a sweetener player showing the highest centrality measures (degree, betweenness, eigenvector and closeness). Additionally. the network categorization shows aggregated products classified as "Healthy Eating" with the highest centrality measures (degree, betweenness, eigenvector and closeness) and "Cleaning Accessories" as the main players of the category networks.

At the brand categorization level, we can observe a similar network structure when comparing with the products network, which means that we are facing a highly connected random network. The most impacted brand in the network is "A VACA QUE RI", a Portuguese brand known for their cheese products, followed by "ACH BRITO", with the highest centrality measures (degree, betweenness, eigenvector and closeness).

The statistical analysis of the networks brings as evidence the complexity of the price system, characterized by several types of relations and many complex features (Rešovský et al., 2013), especially when considering the random structure of the networks and looking to the cross-correlation among products, categories, and brands. The modeling of our data allowed us to identify different marked players, that present a risk and most impact the network.

To conclude, our methodology approach allowed us to get a better view of the retailer market behavior during the period of study and generated a start point to explore future network changes and generate input to the main stakeholders to support business decisions. To finalize our work, we describe the main challenges and limitations identified during the process and future work suggestions to enrich our research.

5.1. Challenges and Limitations

Reviewed literature has shown that building a predictive model for pricing requires a full understanding of the matter and all the factors that are associated with price variations. Besides this limitation that out data set presented, we addressed challenges during the network inference that also impacted the network structure, namely: the limitations of the e-commerce data, and the regular changes in the website structure.

E-commerce data limitations

Online Marketplaces are not only changing the way customers do shopping but also increasing the competition among retail players by the availability of a wide variety of products on a single platform. To face this, companies tend to strategically implement the hybrid model (online store + physical store) and make only part of their product portfolio available online. With this in mind, the domain knowledge might be incomplete, and the online marketplaces do not ensure access to the entire portfolio, creating a bias while detecting correlations and understanding price fluctuations.

Regular changes in the website structure

Additionally, websites regularly change their HTML structure for a better user interface and to enhance customer experience. These changes are anti-scraping which led us to data gaps in the whole data set and resulted in poor data. Scrapers were built according to the website layout and required regular maintenance that was not guaranteed for the correct execution of the pipeline.

5.2. Future Work

We can observe that the amount of data used in algorithms plays a critical role when creating the output and analyzing results. Currently, our dataset covers a period of 7 months of daily price information, which could be improved if we consider a larger time window. In this scope, we propose to enrich the results by analyzing the same product range for the period of 2021, covering the same time window used during our work. Considering the price fluctuations of 2021, we can consider a larger time series and analyze the network scape over time.

As pointed in the stock markets literature, the statistical analysis of the graph properties over time is convenient for studying the impact of relevant government and policies measures in the price fluctuations and provides insightful conclusions, supported by the social environment of the retailer (Yin et al., 2017).

Additionally, it may be interesting to analyze the network structure of the remaining players of the market and explore price correlations between similar products. To achieve this, we propose to collect data sets from other players in the retail market and implement our methodology to study price correlations for match products and brands. This could lead us to significant conclusions in the space of market competition and identify similar behaviors between the retailers.

For more research and model enrichment, we also propose to analyze price fluctuations, based on customer reviews posted on social media. Nowadays, the customer shopping experience is also based on product reviews and other customer's feedback. In this scope, it becomes useful to extract insightful review data and monitor customer satisfaction to adapt price structures accordingly.

6. BIBLIOGRAPHY

Alamsyah, A., Rahardjo, B., & Kuspriyanto. (2014). Community detection methods in social network analysis. *Advanced Science Letters*, *20*(1), 250–253. https://doi.org/10.1166/asl.2014.5301

Bottasso, A., Marocco, P., & Robbiano, S. (2020). *Price Matching and Platform Pricing Munich Personal RePEc Archive Price matching and platform pricing Bottasso , Anna and Marocco , Paolo and Robbiano , Simone*. *December*. https://doi.org/10.13140/RG.2.2.27982.08000

Chen, Y., Narasimhan, C., & Zhang, Z. J. (2001). Research note: Consumer heterogeneity and competitive price-matching guarantees. *Marketing Science*, *20*(3), 300–314. https://doi.org/10.1287/mksc.20.3.300.9766

Chintagunta, P. K. (2002). Investigating category pricing behavior at a retail chain. *Journal of Marketing Research*, *39*(2), 141–154. https://doi.org/10.1509/jmkr.39.2.141.19090

Fernando, R., & Dias, L. (2015). *Monitoring Evolving Stock Networks*.

Giudici, P., & Polinesi, G. (2021). Crypto price discovery through correlation networks. *Annals of Operations Research*, *299*(1–2), 443–457. https://doi.org/10.1007/s10479-019-03282-3

HackerEarth. (s.d.). Obtido em 11 de July de 2021, de https://www.hackerearth.com/practice/algorithms/graphs/minimum-spanning-tree/tutorial/

Hevey, D. (2018). Network analysis: A brief overview and tutorial. *Health Psychology and Behavioral Medicine*, *6*(1), 301–328. https://doi.org/10.1080/21642850.2018.1521283

Hodler, A. E., Farnham, B., Tokyo, S., Boston, B., Sebastopol, F., & Beijing, T. (2019). *Mark Needham Graph Algorithms Practical Examples in Apache Spark and Neo4j*.

Huang, W. Q., Zhuang, X. T., & Yao, S. (2009). A network analysis of the Chinese stock market. *Physica A: Statistical Mechanics and Its Applications*, *388*(14), 2956–2964. https://doi.org/10.1016/j.physa.2009.03.028

Lee, S. (2011). *Study of Demand Models and Price Optimization Performance Study of Demand Models and Price Optimization Performance*. *December*.

Namaki, A., Shirazi, A. H., Raei, R., & Jafari, G. R. (2011). Network analysis of a financial market based on genuine correlation and threshold method. *Physica A: Statistical Mechanics and Its Applications*, *390*(21–22), 3835–3841. https://doi.org/10.1016/j.physa.2011.06.033

Newman, M. E. J., & Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *69*(2 2), 1–15. https://doi.org/10.1103/PhysRevE.69.026113

Okhrimenko, D., Kovalev, K. P., & Titov, S. A. (2020). *Airlines Baggage Price Optimization Model*. *April*, 1–18.

Pandey, A., Jamwal, M., & Soodan, V. (2014). Transforming physical to digital marketplace-E-retail: An Indian Perspective. *International Journal Of*, *4*(4), 11–18.

https://www.researchgate.net/profile/Vishal_Soodan/publication/264155245_Transforming_physical_to_digital_marketplace-E-retail_An_Indian_Perspective/links/53cfb3d10cf25dc05cfb0f14.pdf

Phillips, R. L. (2005). *PRICING AND REVENUE OPTIMIZATION* (Stanford Business Books, Ed.; 1st Edition). Board of Trustees of the Leland Stanford Junior University.

Rešovský, M., Horváth, D., Gazda, V., & Siničáková, M. (2013). *Minimum Spanning Tree Application in the Currency Market*.

Smith, R. D. (2009). *The Spread of the Credit Crisis: View from a Stock Correlation Network*. 1–8.

West, J. (2019, September 7). *Visualising asset price correlations*. Https://Julian-West.Github.Io/Blog/Visualising-Asset-Price-Correlations/.

Yin, K., Liu, Z., & Liu, P. (2017). *TREND ANALYSIS OF GLOBAL STOCK MARKET LINKAGE BASED ON A DYNAMIC CONDITIONAL CORRELATION NETWORK*. *18*(4), 779–800. https://doi.org/10.3846/16111699.2017.1341849

Zhuang, X. W., & Xiu, J. (2015). Correlation analys is between topological properties and market volatility of stock network based on complex network. *Proceedings of the 2015 27th Chinese Control and Decision Conference, CCDC 2015*, 2903–2906. https://doi.org/10.1109/CCDC.2015.7162422

7. APPENDIX A – PRODUCT CLASSIFICATION

Category 1	Category 2	Category 3
Alimentação Equilibrada	Biológicos	Chás e Infusões
		Outras Conservas de Peixe
		Polpas e Concentrados
		Sumos e Néctares
	Cereais e Pequeno-Almoço	Aveia
		Cereais
		Muesli e Granola
		Superalimentos
	Chocolates, Bolachas e Snacks	Barras
		Bolachas
		Chocolates e Guloseimas
		Snacks e Outros
		Tortitas
	Nutrição Desportiva	Bebidas
		Pequeno-Almoço
		Snacks
		Suplementos
	Sem Glúten	Farinhas e Pão
		Snacks e Bolachas
	Vegetariano e Vegan	Cappuccino e Outros
Alimentação Saudável	Bolachas e Snacks	Biscoitos
		Com Chocolate
		Integrais e Digestivas
		Maria e Manteiga
		Outras Bolachas
		Recheadas e Waffers
		Snacks Doces
		Snacks Salgados
		Tabletes de Chocolate
		Tostas e Outros
		Água e Sal e Cream Cracker
	Chás e Infusões	Chá e Infusões
	Go Bio	Arroz
		Biscoitos
		Chá e Infusões
		Com Chocolate
		Iogurte Aromas
		Massa
		Outras Bolachas
		Snacks Doces
		Tabletes de Chocolate

		Tostas e Outros
	Pura Vida	Arroz
		Bebidas
		Biscoitos
		Chocolates de Culinária
		Chás e Infusões
		Integrais e Digestivas
		Iogurte Aromas
		Iogurte Grego
		Iogurte Líquido Aromas
		Iogurte Líquido Magro
		Iogurte Líquido Natural
		Iogurte Líquido sem Lactose
		Iogurte Natural
		Iogurte sem Lactose
		Maria e Manteiga
		Massa
		Outras Bolachas
		Outros
		Queijo Fatiado
		Queijo Flamengo
		Queijo Fresco e Requeijão
		Queijo Regional
		Queijo de Barrar
		Snacks Salgados
		Tudo para Sobremesas
		Água e Sal e Cream Cracker
	Sem Glúten	Arroz
		Com Chocolate
		Iogurte Líquido Aromas
		Maria e Manteiga
		Massa
		Outras Bolachas
		Outros
		Recheadas e Waffers
		Snacks Doces
		Snacks Salgados
		Tabletes de Chocolate
		Tudo para Sobremesas
	Sem Lactose	Iogurte Aromas
		Iogurte Grego
		Iogurte Líquido Aromas
		Iogurte Líquido Natural
		Iogurte Líquido sem Lactose

		Iogurte Natural
		Iogurte Pedaços
		Iogurte sem Lactose
		Leite Achocolatado
		Leite Gordo e Meio Gordo
		Leite Magro
		Leite Sem Lactose
		Maria e Manteiga
		Outras Bolachas
		Outros
		Queijo Fatiado
		Queijo Fresco e Requeijão
		Queijo Regional
		Água e Sal e Cream Cracker
	Vegetarianos	Iogurte Líquido Aromas
		Salsichas
		Tudo para Sobremesas
Animais	Cão	Higiene e Acessórios Cão
		Ração Húmida para Cão
		Ração Seca para Cão
		Snacks e Biscoitos para Cão
	Gato	Areia para Gato
		Ração Húmida para Gato
		Ração Seca para Gato
		Snacks e Leite para Gato
	Outros animais	Peixes
		Pássaros
		Roedores
Bazar	Livraria	Livros Adulto
		Livros Criança
		Livros Escolares
		Revista Sabe Bem
Bebidas	Cerveja	Cerveja Estrangeira
		Cerveja Nacional
		Cerveja Sem Álcool
		Cerveja Sem Álcool
	Leite	Outros Leites
	Molhos	Tudo para Sobremesas
	Sumos e Refrigerantes	Bebidas Energéticas
		Concentrados
		Néctares e 100%
		Refrigerantes Com Gás
		Refrigerantes Sem Gás
		Refrigerantes com Gás

		Refrigerantes sem Gás
		Sumos e Néctares
	Vinho	Espumante
		Vinho Branco
		Vinho Rosé
		Vinho Tinto
		Vinho Verde
		Vinho do Porto
	Vinho Branco	Alentejo
		Corrente
		Douro
		Dão
		Espumante
		Lisboa
		Setúbal
		Tejo
		Verde
	Vinho Tinto	Beira
		Bairrada
	Vinhos	Outros Vinhos
		Vinho Branco
		Vinho Rosé
		Vinho Tinto
		Vinho do Porto
	Água	Outros
		Água Aromatizada e Tisanas
		Água Com Gás
		Água Sem Gás
		Água Tónica
Bebé e Criança	Alimentação Infantil	Iogurte Infantil
		Leite Infantil
	Higiene	Champô
		Elixir
		Escova de Dentes
		Outros
		Papel Higiénico e Lenços
		Pasta de Dentes
Bem Estar	Cosméticos	Creme de Corpo
Boas Festas	À Mesa	Iguarias Pingo Doce
Bolachas e Doces	Bolachas e Biscoitos	Biscoitos
		Com Chocolate
		Infantis
		Integrais e Digestivas
		Maria e Manteiga

		Outras Bolachas
		Para Levar
		Recheadas e Waffers
		Água e Sal e Cream Cracker
	Bolos e Sobremesas	Gelatina
		Pudim e Mousse
		Tudo para Sobremesas
	Chocolates e Guloseimas	Chocolates de Culinária
		Guloseimas
		Outros Chocolates
		Tabletes de Chocolate
	Compotas e Cremes de Barrar	Iguarias Pingo Doce
		Outros
Casa	Limpeza da Casa	Abrasivos
		Blocos Sanitários
		Lava Tudo
		Lixívia
		Madeira e Vidro
		Multiusos
		Outros Limpeza da Casa
		Tira Gorduras
	Loiça	Acessórios de Limpeza
		Aditivos, Sal e Abrilhantadores
		Detergente Manual
		Detergente Máquina
		Esfregão
	Roupa	Aditivos e Outros
		Amaciador
		Detergente
Charcutaria	Presunto	Iguarias Pingo Doce
Congelados	Bacalhau Congelado	Iguarias Pingo Doce
	Carne	Almôndegas, Hambúrgueres e Carne Picada
		Crocantes e Panados
	Gelados e Sobremesas	Iguarias Pingo Doce
	Marisco	Delícias e Preparados Marisco
	Marisco Congelado	Iguarias Pingo Doce
	Outras Refeições	Iguarias Pingo Doce
	Peixe	Crocantes, Panados e Hambúrgueres
		Filetes, Medalhões e Outros
	Refeições	Entradas e Salgados
		Pizzas, Massas e Outros
Frigorífico	Iogurtes	Iogurtes Aromas
		Iogurtes Bifidus e Funcionais

		Iogurtes Gregos
		Iogurtes Líquidos
		Iogurtes Magros
		Iogurtes Naturais
		Iogurtes Pedaços e Mix-In
		Iogurtes Vegetais
		Iogurtes com Gelatina
		Skyr, Kefir e Proteína
	Queijos	Queijo Barra e Fatiado
		Queijo Estrangeiro
		Queijo Flamengo
		Queijo Fresco e Requeijão
		Queijo Nacional
		Queijo Snacking
		Queijo para Barrar
		Queijo para Culinária
Frutas e Legumes	Legumes	Legumes Simples
		Misturas e Pré-Cozinhados
	Legumes Preparados	Snacks Salgados
Higiene e Beleza	Cabelo	Acessórios
		Champô
		Coloração
		Condicionador
		Laca, Espuma e Outros
		Máscara
		Máscaras e Tratamentos
	Casa de Banho	Acessórios e Outros
		Gel de Banho
		Papel Higiénico e Lenços
		Sabonete
		Tratamento de Rosto
	Corpo	Acessórios de Higiene
		Cremes, Loções e Óleos
		Depilatórios e Descolorantes
		Desodorizante
		Gel de Banho
		Mãos e Pés
		Sabonetes
	Cosméticos	Creme de Corpo
		Maquilhagem
		Tratamento de Mãos e Pés
		Tratamento de Rosto
	Depilatórios	Barba
		Cabelo e Corpo

	Desodorizantes	Homem
		Mulher
	Higiene Infantil	Champô
		Elixir
		Escova de Dentes
		Outros
		Papel Higiénico e Lenços
		Pasta de Dentes
	Higiene Oral	Elixir
		Escova de Dentes
		Pasta de Dentes
	Higiene Íntima Feminina	Pensos Diários
		Pensos Higiénicos
		Tampões
		Toalhitas Íntimas
	Homem	Barba
		Barbear
		Cabelo e Corpo
		Desodorizante
	Preservativos e Lubrificantes	Lubrificantes
		Preservativos
	Primeiros Socorros	Acessórios e Outros
		Outros
		Tratamento de Rosto
	Rosto	Creme de Rosto
		Lábios
		Tratamento de Rosto
	Solares	Creme de Corpo
		Gel de Banho
		Tratamento de Rosto
Infantil	Higiene Oral	Cuidados Para Próteses
		Elixir
		Escova de Dentes
		Fio e Fita Dentária
		Pasta de Dentes
Lacticínios	Bebidas Vegetais	Bebidas
	Iogurtes Líquidos	Iogurte Líquido Aromas
		Iogurte Líquido Infantil
		Iogurte Líquido Magro
		Iogurte Líquido Natural
		Iogurte Líquido sem Lactose
		Kefir
	Iogurtes Sólidos	Iogurte Aromas
		Iogurte Grego

		Iogurte Infantil
		Iogurte Magro
		Iogurte Natural
		Iogurte Pedaços
		Iogurte Proteico
		Iogurte sem Lactose
		Kefir, Proteína e Skyr
		Soja e Vegetal
	Leite	Leite Achocolatado
		Leite Gordo e Meio Gordo
		Leite Infantil
		Leite Magro
		Leite Sem Lactose
		Outros Leites
	Manteigas e Cremes Vegetais	Iguarias Pingo Doce
	Natas, Bechamel e Chantilly	Outros
	Queijo	Outros Queijos
		Queijo Estrangeiro
		Queijo Fatiado
		Queijo Flamengo
		Queijo Fresco e Requeijão
		Queijo Regional
		Queijo de Barrar
	Sobremesas	Iogurte Aromas
Leite, Ovos e Natas	Leite Magro	Leite Sem Lactose
	Outros Leites	Tudo para Sobremesas
Mercearia	Aperitivos e Snacks	Bolsas de Fruta
		Frutos Secos e Desidratados
		Pipocas
		Snacks Variados
		Tostas e Croutons
	Arroz, Massa e Feijão	Arroz
		Feijão e Grão
		Massa
		Massa Chinesa e Noodles
	Azeite, Óleo e Vinagre	Outros
	Açúcar, Farinha e Ovos	Outros
	Batatas Fritas	Iguarias Pingo Doce
		Originais e Gourmet
		Originais e Light
		Palha
		Sabores
	Bolachas e Bolos	Bolos, Tortas e Queques

	Com Chocolate
	Com Recheio
	Croissants e Pães de Leite
	Integrais e Digestivas
	Maria, Manteiga e Clássicas
	Para Criança
	Waffers e Outras
	Água e Sal e Cream Cracker
Café e Chá	Café
	Café em Cápsulas
	Cevadas e Misturas
	Chá e Infusões
	Descafeinado
	Outros
Cereais	Snacks Doces
Cereais e Barras	Barras de Cereais
	Cereais Linha e Fibra
	Cereais Muesli e Granola
	Cereais Variados
Chocolates	Bombons e Trufas
	Snacks de Chocolate
	Tablete de Chocolate
Chá, Café e Achocolatados	Achocolatados
	Bebidas de Cereais
	Café Moído e Grão
	Café Solúvel
	Cappuccino e Outros
	Chás e Infusões
	Cápsulas e Pastilhas de Café
	Descafeinado
Compotas e Doces	Creme de Barrar
	Marmelada
	Mel
Conservas	Atum
	Atum e Sardinha
	Azeitonas e Tremoços
	Azeitonas, Pickles e Tremoços
	Frutas e Legumes
	Outras Conservas
	Outras Conservas Vegetais
	Outras Conservas de Peixe
	Patês
	Patês e Pastas
	Salsichas

	Guloseimas	Gomas
		Outras Guloseimas
		Pastilhas Elásticas
		Rebuçados
	Massa	Outros
	Molhos	Frutas e Legumes
		Tudo para Sobremesas
	Pré-Cozinhados	Massa
		Massa Chinesa e Noodles
	Sal, Temperos e Caldos	Outros
	Snacks	Bolsa de Fruta
		Frutos Secos
		Snacks Doces
		Snacks Salgados
		Tostas e Outros
	Sobremesas	Leite Condensado e Evaporado
		Polpas e Frutas em Calda
		Preparado de Gelatina
		Preparados de Mousses e Bolos
		Tudo para Sobremesas
	Temperos	Caldos
		Especiarias e Ervas Aromáticas
		Maionese, Mostarda e Ketchup
		Outros Molhos e Temperos
		Polpas e Concentrados
		Sal
Padaria e Pastelaria	Bolos Embalados	Doces Tradicionais
		Snacks Doces
	Pão Ralado e Outros	Tostas e Outros
Tudo para Casa	Acessórios de Limpeza	Acessórios e Outros
		Esfregão
	Ambientador	Outros
	Festa	Outros
	Lixívias	Acessórios e Outros
	Loiça	Esfregão
	Roupa	Amaciador
		Detergente