A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

Field Lab Jerónimo Martins: Optimization of Retail Operations – A Cluster-Based Approach to Market Basket Analysis

SIMON LINK

46018

Work project carried out under the supervision of:

Qiwei Han

17-12-2021

Abstract

In the context of retail analytics, market basket analysis serves as a powerful technique to extract valuable knowledge about consumer preferences and shopping habits. Through its application to a Portuguese retailer, the following study examines purchase transaction data and clusters it based on the product categories in consumers' baskets. With the goal of mining product relationships, this study compares a heuristic and an association rule-based approach for a cluster-based identification of product substitutes and complements. The paper concludes that for finding substitutes, the heuristic fares the best results. For the discovery of product complementarity, an association rule learning-based approach is suited best. Beyond its theoretical contribution, the insights gained through the analyses are utilized to increase customer satisfaction and sales by providing recommendations on managerial decisions, ranging from determining the timing of product promotions, possible improvements to the store's design, informing product placement decisions, to suggestions regarding customer communication.

Keywords: Retail Analytics, Market Basket Analysis, Purchase Transaction Clustering, Association Rule Learning, Products Complements and Substitutes.

This work used infrastructure and resources of Jerónimo Martins SGPS, S.A. and the Pingo Doce & Go Nova Store.

Table of Contents

1 Motivation

In the present decade, the retail industry is without doubt facing ground-breaking disruptions. In the context of the growing influence of online retail and the rise of aggressive competitors in the form of on-demand grocery delivery services, the way of doing classical "brick-and-mortar" retail has to and will change substantially, in the next years (J. Huang, Kohli, and Lal 2018). Supermarkets will define a completely new shopping experience, based on technological advancements such as automated processes, an ever-expanding digital infrastructure, and stores without cashiers. Big disruptions are bound to happen also on the operational side that reach from intelligent planning and procurement, and warehousing automation to new marketing channels that change the way of doing business in a sustained way (Ren, Chan, and Siqin 2020). The data generation and quality of the next generation of supermarkets are immense. Hence, the opportunities presented by this unprecedented variety, volume, and velocity of data are also enormous (Matthew, Kevin, and Brian 2015).

Jerónimo Martins, the leading Portuguese retailer, has opened a store dedicated to testing and implementing technological advancements in the retail sector. The Pingo Doce & Go Nova store in Lisbon is the first of its kind in Europe. It aims at pathing the way for the next generation of supermarkets. The store is testing these new technologies on a very special customer group – students, who are regarded to be tech-savvy and highly adaptive customers. The convenience store offers a broad range of products, including service products like freshly prepared pasta and salads. Handling peak demand for those products that require service, however, is difficult, which leads to long waiting times and a bad customer experience, especially during lunchtime. As customer satisfaction affects profitability, it is critical to improve the customer experience (Anderson, Jolly, and Fairhurst 2007). For this purpose, the far-reaching capabilities of new business analytics tools, in combination with the sheer quantity of data available, are of great

benefit. Decision quality can be improved substantially by data-driven decision-making. This paper suggests an approach to tackle the following question: “How can a cluster-based approach to market basket analysis be used to steer the demand of service and non-service products?”.

First, customer missions are identified. Then, two different approaches to compute substitutes are compared. The purchase transaction data is also leveraged to compute complements following the two approaches. Analyzing the purchase transaction data may be useful in certain tasks, ranging from designing personalized marketing campaigns, informing product placement decisions to determining the timing of product promotions (Adomavicius and Tuzhilin 1999; Agrawal and Srikant 1994; Ho Cho, Kim, and Kim 2000). Based on the insights gained throughout the analyses, suggestions are provided on how to use this information in a fruitful way.

This paper contributes to existing research as, to the end of finding product substitutes and complements, it compares both a novel heuristic and an association rule learning-based approach. To the best of my knowledge, these two approaches have never been compared. Furthermore, the computations are done specifically per shopping mission cluster. What is more, the setting for the data this paper works with is unique, as we deal with an experimentation store that predominantly has students as customers, a customer group with special behavior that has barely been researched.

The remainder of this paper is organized as follows. Section 2 reviews the literature on purchase data clustering and the identification of substitute and complement product relationships. Section 3 presents the methodological approach this paper follows. In Section 4, the results of the analyses are evaluated. Section 5 discusses managerial implications and recommendations inferred by the analyses. Section 6 concludes with a summary, presentation of limitations, and suggestions for future work.

2 Literature Review

2.1 Identification of Customer Missions

One main use case of data analytics for retail data is to systematically perform customer segmentation and to identify customer missions (Sarantopoulos et al. 2016). Customer segmentation is "the process of dividing heterogeneous customers into homogeneous groups on the basis of common attributes and is essential for handling a variety of customers" (Wu and Lin 2005).

There exist various studies that utilize manifold kinds of data in order to identify patterns in customer purchases, ranging from demographics over sales data to technographic data (Griva et al. 2018). Most of these studies identify customer missions with the objective to support the customization of the retail service offering to different customer segments (Sarantopoulos et al. 2016). One branch of research is based on the assumption that individuals move differently in space and tries to determine shopping missions based on motion tracking and interviews (Koch et al. 2009). Another important branch examines large transactional databases in order to find out which products are purchased together frequently (Agrawal, Imieliński, and Swami 1993). This branch can be found under the name "market basket analysis", which mines associations between items.

Current research in this branch utilizes the entire purchase history of a customer, meaning all shopping visits, to determine customer groups (Aeron, Kumar, and Moorthy 2012; Khajvand et al. 2011). As these studies examine the entirety of a customer's shopping history, they omit the shopping purpose of a single customer visit (Sarantopoulos et al. 2016). However, marketing researchers have stressed the need to *understand a single customer visit*, because every single visit carries substantial insights on the shopper's need and, thus, can enable retailers to take

actions to satisfy them (Walters and Jamil 2003; Bell, Corsten, and Knox 2011; Griva et al. 2018).

Carugati, Kokkinaki, and Pouloudi (2014) present a data mining-based framework to identify shopping missions, demonstrating the utility of the framework through the application of the framework to real data of several stores of a Greek retailer (Carugati, Kokkinaki, and Pouloudi 2014). Griva et al. (2018) propose a business analytics approach that mines customer visit segments from market basket data, applying it to a real use case of a major European fast-moving consumer goods retailer. The latter approach extracts knowledge that can be used to support several decisions, ranging from redesigning a store's layout, over marketing campaigns per customer segment to product recommendations (Griva et al., 2018). Sarantopoulos et al. (2016) developed an analytical method for the identification of shopper need states, where they perform clustering for customer mission identification at a store level (Sarantopoulos et al. 2016).

After exploring, cleaning, and preparing the data, all these papers have in common that they use the k-means clustering algorithm to identify shopping missions. Many studies stress the importance of finding an appropriate product category granularity level for clustering because it strongly impacts results.

A research gap has been identified as, to the best of my knowledge, these approaches have never been applied to this new retail format, a next-generation supermarket. In the here presented methodology, clustering will be used to group purchase transactions into exclusive clusters. The characteristics of the single clusters can then be analyzed to gain insight into the composition of the baskets each cluster contains, indicating what type of purchase behavior is associated with the respective cluster.

2.2 Mining Relationships between Products

One popular approach for analyzing market basket data is the discovery of association rules. Association rule learning, at a basic level, is "a technique of machine learning to analyze data for patterns or co-occurrences" (Agrawal, Imieliński, and Swami 1993). Association rules are "if-then" statements that can be mined using transactional data. They appear in the form $X \Rightarrow Y$, X and Y being itemsets, where X is called the antecedent and Y the consequent of the rule. These rules, for example $butter \Rightarrow bread$ ("if butter, then bread"), provide information on how often a combination of products is bought together or if the purchase of one article influences the probability of the purchase of another article (Agrawal, Imieliński, and Swami 1993; Kotsiantis and Kanellopoulos 2006). In this regard, there exist several metrics that can be calculated to evaluate association rules. The *support* of a rule is the fraction of transactions that contain both X and Y, i.e.,

$$sup(X \Rightarrow Y) = P(X \cap Y)$$

The *confidence* of a rule is measured as the fraction of transactions containing X that also contain Y.

$$conf(X \Rightarrow Y) = \frac{sup(X \Rightarrow Y)}{sup(X)} = \frac{P(X \cap Y)}{P(X)}$$

It, thus, reveals the conditional probability that transactions in the database contain Y, given that we know they contain X. Thirdly, the *lift* summarizes the strength of the association between the products on the left and right side of the rule; the larger the lift, the greater the link between the two itemsets.

The equation to calculate the lift is

$$lift(X \Rightarrow Y) = \frac{sup(X \Rightarrow Y)}{sup(X) * sup(Y)} = \frac{P(X \cap Y)}{P(X) * P(Y)}$$

This metric allows us seeing e.g., which itemsets are correlated positively or negatively (Hussein, Alashqur, and Sowan 2015). Lift is a simple, yet one of the most powerful metrics, because it gives insights about the dependency of products, and reveals information if products can be considered complements or substitutes (Puka and Jedrusik 2021).

Other methods for analyzing market basket data are heuristics that have inspiration from other research fields. Mungoli (2020) presented an intuitive and novel method to find complements and substitutes based on the computation of both a complement and a substitute ratio. The big advantage of this approach is that it is simple to implement in scale and easily comprehensible. Mungoli implemented the heuristic on the Instacart data set[1] and concluded that the results are superior to traditional association rule learning algorithms or finding support, confidence, or lift (Mungoli, 2020). In this paper, the heuristic introduced by Mungoli and the traditional association rule learning approach are compared in order to find product complements and substitutes. The next chapter describes the methodology of the approach.

[1] The Instacart data set is famous data set, originally published for a competition on Kaggle.com, containing relational database data from a grocery retailer. It can be found under: https://www.kaggle.com/c/instacart-market-basket-analysis.

3 Methodology

3.1 Overview

This paper proposes an approach to identify product relationships and compute complements/substitutes, that takes into account different customer missions. The presented methodology is structured in two phases. Phase 1 describes the identification of customer missions. For selected customer missions, product substitutes and complements are computed in phase 2.

3.2 Dataset and Preparation

The data for this study was provided by the Portuguese food retailer Jerónimo Martins. It comprises information about purchase transactions and their customer and products in the Pingo Doce & Go Nova Store for a three-month-period from the beginning of September until the end of November 2021. The transactional data was enriched by merging product details onto the transaction records, using the product ID as a key. After removing missing and negative quantity values, recycling incentive products were dropped, as they are not considered as a real product by the supermarket. After data refinement, the data set comprises information about around 150,000 purchase transactions. In order to create the basket data, next, the transaction data was grouped by transaction ID. Then, a binary encoding was applied, creating boolean vectors to indicate for each transaction ID, whether each product division[2] was contained in the transaction or not. As mentioned before, an appropriate level of granularity is important. In our case, the product division seems to have a satisfying level of granularity, where not too much information is omitted, but which is not too granular. A snapshot of the encoded market basket data can be perceived in Table 1.

[2] The product division is the second hierarchy level when grouping products in the Pingo Doce & Go Store. The hierarchy levels are product area > division > family > category > sub-category.

3.3 Dimensionality Reduction

The encoded market basket matrix, however, is very sparse and suffers high dimensionality. When using distance-based similarity measures, as the number of dimensions increases, these types of measures converge to a constant value between any given examples (Google Developers Documentation 2021). Therefore, in order to reduce the dimensionality of the data and to ensure improved clustering results, a Principal Component Analysis (PCA) was conducted. By applying a 90% explained variance ratio threshold, an optimal number of components of nine was determined, according to which the data set was transformed.

3.4 Market Basket Data Clustering

We can now apply a clustering algorithm to group together similar shopping visits. The first choice for clustering the data is k-means, an iterative unsupervised learning algorithm, that tries to partition the dataset into k pre-defined, distinct, non-overlapping subgroups (Z. Huang 1998). K-means is especially suitable for the application with large data sets, as it is not very computationally expensive, comes with linear time complexity (e.g., compared to hierarchical clustering algorithms) and guarantees convergence (Google Developers Documentation 2021). What is more, its cluster centers are easy to interpret, which is important in our business setting. To be more specific, an optimized version of k-means, namely k-means++, is applied, which selects the initial cluster centers in a smart way which results in speeding up convergence. However, for k-means, the number of clusters k has to be determined prior to initializing the algorithm. Using the elbow method in combination with the distortion score, a common heuristic for identifying the optimal number of clusters, our model finds an elbow where the number of clusters equals six (see Figure 1). For the final training of the model, a k-means++ model is used, and the optimal number of clusters is set to six. The clustering output allows dividing the purchase transactions into six sub-groups, that show similar purchase behavior, which will be analyzed in-depth in the following.

3.5 Computation of Complements and Substitutes

Within the established clusters, interesting relationships between products can now be mined, gaining insights that can be used for improving the store's layout or derive interesting marketing campaigns. To this end, we want to know to which degree products can be considered complements or substitutes of each other. In order to do so, the results that both the application of a search heuristic for complements/substitutes and traditional association rule learning yield are compared.

3.5.1 Complement/Substitute Ratio Heuristic

In the first step, we perform a cross-join on the products data set, which combines each product with each product. Then we count how often each product appears and how often the products were bought together, for each combination.

Complements are items that tend to be bought together. Two items X and Y are called complements if they add value to another. In other words, X and Y are two items the consumer uses and buys in conjunction. In order to find such pairs of items, we want to compute the ratio of the number of times X and Y are bought together to that of the number of times X and Y are bought, across all customer baskets (Mungoli 2020).

From set theory, we know

$$|X \cup Y| = |X| + |Y| - |X \cap Y|$$

The Jaccard similarity coefficient, which is a common metric used for gauging the similarity or diversity of a data set, takes the ratio of intersection over union (Jaccard 1912). To calculate the complement ratio of two items X and Y, we apply the following formula, inspired by the Jaccard coefficient (Mungoli 2020).

$$complement\ ratio = \frac{|X \cap Y|}{|X \cup Y|} = \frac{|X \cap Y|}{|X| + |Y| - |X \cap Y|}$$

The complement quality of a pair of two items *X* and *Y* will be higher, the higher the complement ratio is. Thus, we will be sorting the item pairs by descending complement ratio to find the highest quality complements. Thereby, the complement ratio is capped to one.

Substitutes are alternative items used for the same purpose. Two items *X* and *Y* are called substitutes if they are direct competitors for each other and people tend to buy one of the two items but not both (Mungoli 2020). Following the heuristic proposed by Mungoli, to calculate the substitute ratio of two items *X* and *Y*, we apply the following formula.

$$substitute\ ratio\ = \frac{|X \cap Y|}{min(X, Y)}$$

The numerator of the fraction is the intersection of *X* and *Y*, i.e., the count of transactions in which the two products are both present. The denominator is a minimum formula with the overall count of each item *X* and item *Y* as input. Some items can be very popular and thus, will be part of many baskets. The incorporation of a minimum function renders assistance in removing such bias and brings the comparison to a fair standard (Mungoli 2020).

Following this heuristic, we are trying to find item pairs where both items are very frequently bought, but that are very rarely bought together. Hence, we are sorting the data by ascending substitute ratio, because the lower the substitute ratio, the better the quality of the substitute pair.

If only one of the two items is bought very frequently, but the other not, the numerator will be very small, but then the minimum function in the denominator will take the very low number of the rarely bought item. This means that a small number is divided by a small number, resulting in a high substitute ratio, which ranks the combination of these two items very low in the ranking.

3.5.3 Association Rule Learning: Lift Metric

An alternative way to mine relationships between items is to apply association rule learning. Association rule learning discovers all association rules that are above a minimum support and minimum confidence level set by the user (Agrawal et al., 1993). In this context, the Apriori algorithm was developed to facilitate the generation of relevant association rules. It can be structured in two phases. In the first phase, it generates all itemsets of size k that satisfy a minimum support. Hereby, it works under the assumption that if an itemset is infrequent, all its supersets must not be frequent (Chui, Kao, and Hung 2007). In the second phase, it generates rules from the set of all remaining, frequent itemsets. This way, it is able to deal with itemsets that contain not only one, but serval possible items, while at the same time working computationally efficiently (Agrawal and Srikant 1994; Agrawal, Imieliński, and Swami 1993). For our dataset, however, the Apriori algorithm only finds itemsets of a maximum length of two, which means association rules that consist of a combination of two products. This outcome is supported by our data, as the average size of unique products in the data set is around 1.8, meaning that most people buy less than two different items per shopping visit.

Similar to the heuristic approach presented in Sections 3.5.1 and 3.5.2, we can compute the lift metric for each product combination pair, following the equation presented in Chapter 2.2. The lift metric provides information on the relationships of the product pairs. A lift greater than one implies a positive correlation between the items, which indicates X and Y show a complementary effect. A lift value of one means that the two items are not correlated. A lift smaller than one indicates that the two products are negatively correlated, which indicates that X and Y can be seen as substitutes (Puka and Jedrusik 2021). The higher/lower the lift value for a rule, the stronger the quality of complementarity/substitutability. The results obtained from this approach can then be compared to the results the complement/substitute ratio heuristic has produced.

4 Analysis

4.1 Cluster Analysis

The cluster analysis shows that the purchase transactions can be grouped into quite well-delimitable groups, which show interesting patterns. These six *customer missions* can be identified: To start with, cluster 1 is a "sweet snack" cluster, which is relatively small and only accounts for 8% of overall transactions. The predominant product divisions in this cluster are confectionery products (contained in 100% of the transactions) and bakery products (25%). The most sold products are cookies, chocolate snacks, and chewing gum. Cluster 1 shows a small peak around 4 and 5 pm, when we look at intraday cluster occurrences (see Figure 2). Cluster 2, the "meal deal" cluster, holds 23% of all transactions. Its predominant product divisions are take-away (100%) and bakery (14%). The most frequently sold products in this cluster are personalized pasta, sandwiches, and pizza, amongst others. This cluster shows a considerable peak around lunchtime (between 12 am and 2 pm), and then a small increase in demand around 8 pm. Cluster 3 has been identified to be the "coffee/pastry snack" cluster. The transactions of this type make up for a total of 31% of overall transactions, making it the largest cluster. The strongest division in this cluster is, by far, the bakery division, which all the transactions associated with this cluster have in common, best-selling products being ham & cheese pastry, cheese bread, and coffees of different types. The cluster shows several peaks in transaction load, at 9 am, 11 am, and 4 and 5 pm, and a small peak at 2 pm (apparently the "after lunch coffee"). Cluster 4, the "alternative lunch" cluster, is with a share of 12% considerably smaller. All its transactions contain soft drinks (100%), and 59% contain the take-away product division. Like cluster 2, it shows peaks around lunchtime, and a small peak at 8 pm. The most frequently sold products in this cluster are Coca-Cola, fresh pasta, smoothies, energy drinks, and iced tea. Cluster 5 encompasses a "healthy fruit snack". It accounts for 13% of overall transactions. The strongest product divisions are fruits and vegetables (100%), bakery (34%), and take-away

(30%). Most sold products are fresh pressed orange juice, and, apart from that, fresh fruits of all kinds: Bananas, apples, pineapple pieces, strawberry pieces, melon pieces, and coffee. This cluster shows a small peak around 9 and 12 am, and some demand in the afternoon around 4 and 5 pm. Lastly, cluster 6, called the "drink break" cluster, is as big as cluster 5. However, it shows quite dispersed product divisions, the strongest being water (28%), packaged goods (25%), beer (22%), and dairy products (15%). It shows a small peak at 4 and 5 pm, but in general, a relatively moderate course. Best-selling products in this cluster are water, beer, and cider, frozen lasagna, and ready-made coffee.

A detailed overview of the intraday cluster occurrences can be found in Figure 2. Insights into the composition of the single clusters are provided in Figure 3 and Table 2.

4.2 Finding Product Substitutes and Complements: Meal Deal Cluster

Taking a look at the customer missions from a higher level, we can perceive that the peak in cluster 2 around lunchtime already suggests that the demand for meal solutions products around this time is very high. This observation is supported by the long waiting times for service products, that customers face between 12 and 2 pm. One of the main pain points for customers in the Pingo Doce & Go Nova Store, is the long waiting time for service products (personalized pasta and salads), especially during lunchtime.

Therefore, we want to identify suitable substitutes for these two products. To this end, the approach presented in Chapter 3 is set into practice. In a first step, for all product combinations in the "meal deal" cluster (cluster 2), which contains the lunch purchases, the substitute & complement ratio, as well as the lift metric is computed. Regarding the goal of finding substitutes/complements, for the two heuristic ratios, in the same way as for the lift metric, it does not matter which one of the products is on the right and left side of the rule, as the metrics are the same for both combinations of two product (i.e. {butter} → {bread}, and {bread} →

{butter} respectively have the same complement/substitute ratio and lift, no matter which item is the antecedent, and which one is the consequent).

4.2.1 Substitutes

In general, the results obtained when sorting the product table both for ascending substitute ratio, as well as ascending lift, are similar. As we deal with an unsupervised learning problem, and there is no uniform way for rating substitutes, domain knowledge has to be applied to evaluate the results.

For freshly prepared, personalized pasta, the top 20 substitutes are very similar for the heuristic and the lift-based approach. Indeed, the top 10 substitutes for both approaches are completely identical. For the following ranks, however, the order differs slightly. Following the heuristic, the tuna baguette, the ham and cottage cheese baguette, ham pizza, and the salmon sandwich seem to be good contenders as substitutes for freshly prepared pasta. Highly ranked are also rice, the Angus burger, and the veggie wrap. Next to these products, sandwiches, and pizzas appear a lot on the table.

Regarding personalized salads, the results of the two approaches do not differ considerably, however, again, we find differences in the order of the top substitutes. Sorting by ascending substitute ratio, the Angus burger, chicken pizza, and the tuna baguette are the top substitutes, while sorting by ascending lift, the mozzarella & tomato sandwich, and the angus burger, and ham and cheese pizza are suggested to be the best substitutes for personalized salads. Notably is the fact that the already prepared Caesar salad (a non-service product) is one of the top substitutes for personalized salads, following both approaches. A detailed overview of the substitutes for both pasta and salads can be found in Tables 3 and 4, respectively.

After the application of the Apriori algorithm (with a minimum support threshold of 0.0008[3]), the possible space of itemsets is reduced to 124. As Table 5 shows, for pasta, now only 13 substitutes are found (being cheese bread, olive bread, beer, and the rest predominantly water products). It is important to note, however, that only cheese bread and olive bread have a lift smaller than one, implying that the other substitute suggestions might rather be of complement character (which can be confirmed when applying domain knowledge). For the salad, only one substitute was identifiable: olive bread, which, however, has a lift greater than one (see Table 6). Thus, the usefulness and quality of these suggested substitutes are rather doubtful.

Consequently, it can be concluded that the application of the Apriori algorithm has caused substantial deterioration of the results for substitutes. In the context of the identification of substitutes, the pruning carried out by the algorithm has the effect that potential, well-suited substitutes are not even included in the analysis.

4.2.2 Complements

For marketing purposes and as insights into possible measures to increase sales, it is also worth finding out which products have a complementary character towards each other. For the heuristic, however, the results are not very satisfactory. The approach has some flaws, as the broad majority of top complements (with complement ratio sorted descending) has a complement ratio of one, coming from the fact that both their item union and intersection is equal to one, so they have been bought one time, and this very time the two products were bought together. These deficiencies can be considerably alleviated by introducing a filter, that e.g., only considers combinations where each item occurred at least 200 times in all transactions. The results of the heuristic, after applying filtering are shown in Table 7.

[3] In order to not prune away too many itemsets, I set the minimum support argument to 0.0008, which was the lowest value I could go until it becomes computationally unhandleable for my local machine.

The results for the application of the lift metric are similarly unsatisfactory. Indeed, some of the suggestions are rather of substitute character. For all top complements, i.e., the ones with the highest lift, the item union is equal to the count of one of the two items, and the other item count is one. This one time the two items were bought together. So, if the item with item count one is bought, the other one is bought, as well, in 100% of the cases in our data set. Again, introducing a filter for minimum occurrence improves result quality (see Table 8). In general, the heuristic and the lift-based approach show relatively similar results. The top complements are similar, only the order is slightly different and some elements in the top 10 list differ from each other.

Applying the Apriori algorithm, however, provides the best results, after applying domain knowledge. The top product complement pairs are the cheese & ham pizza and the oven service, cream cheese and the tuna pasta salad, water and a reusable bottle, meatballs and rice, and, rice and chicken thighs (see Table 9). These results, indeed, are most insightful when we look at the context of the store, as they allow for the broadest selection of complements among all approaches. The first two approaches partly showed similar results, but not the same wide range.

4.3 Validation on the Coffee Snack Cluster

Now the results the different approaches give us are evaluated on the coffee snack cluster, which is proportionally the largest next to the meal deal cluster and out of great interest.

4.3.1 Substitutes

Again, the results for the substitute ratio and the results obtained by sorting by ascending lift are not too different from each other. Following the heuristic, good substitutes are e.g., the croissant with seeds and the chorizo bread, next to cake and various coffee types. Applying the lift metric, the mixed pastry snack and maize & sunflower seed bread or walnut & sultanas pumpkin bread, next to 8-cereals bread and sausage puff pastry are good substitutes. However,

for both approaches, a lot of coffee is suggested as a substitute for pastries. In reality, this substitution will not often be observed, as customers buy either one or the other, or both products, but not as a substitute. After the application of the Apriori algorithm, the results show similar product substitutes, but as expected, they comprise a less broad range. The best substitutes in the coffee snack cluster (the results for the substitute ratio) are shown in Table 10. It is noticeable that "merenda mista" appears in most pairings, which can be traced back to the fact that this product is the best-selling in the store.

4.3.2 Complements

In the coffee snack cluster, following the heuristic, croissants and orange juice, mixed bread and orange juice, and espresso and orange juice are suggested as good complement pairs. All other combinations in the top 20 list have an item count of only one or two, which allows them to have a very high complement ratio, but not being very reliable associations. Similar to the observations in the meal deal cluster, the results for the lift metric do not make much sense under the application of domain knowledge. However, again, they get substantially better after the introduction of the filter of a minimum item count of 200, both for the heuristic, as under the evaluation of the lift metric.

However, after applying the Apriori algorithm, the results are best, compared to the other approaches and can generate a substantial business impact. As one could suspect, when having a deeper look at the cluster, product pairs with a strong relationship are e.g., orange juice + a bakery produce (croissant, bread, etc.), cheese puff pastry + decaffeinated latte coffee, water + reusable bottle, bread + hummus/ham, or orange juice + espresso (see table 11). These results reflect very well what the average customer purchases in the store as a coffee break snack.

4.4 Discussion

Dividing the purchase transactions into clusters at the beginning, according to their customer mission, allowed it to set focus and made it possible to independently examine the single missions. This also helped to make finding product substitutes and complements more efficient.

All in all, after validating the results with domain knowledge, we can say that both the heuristic, as well as the association rule learning approaches yield pretty good results in finding product substitutes. We find the best substitutes to be products that occur very often in transactions, but only seldomly together. This validates the methods in the sense that they do not suggest products that occur very infrequently, but rather provide useful suggestions.

On the contrary, the lift metric has some flaws. With an increasing number of records in the data set (n), the lift increases substantially, when we look at its decomposed formula:

$$lift(X \Rightarrow Y) = \frac{sup(X \Rightarrow Y)}{sup(X) * sup(Y)} = \frac{P(X \cap Y)}{P(X) * P(Y)} = \frac{\frac{|X \cap Y|}{n}}{\frac{|X|}{n} * \frac{|Y|}{n}} = \frac{|X \cap Y| * n}{|X| * |Y|}$$

This behavior is known as the “oddity of lift” (Ramesh 2019). When working with large data sets, like in the present case, a threshold of one to identify if products behave as substitutes or complement to each other is no longer completely adequate. Nevertheless, sorting by ascending lift can reveal insights about product substitutability. Applying the Apriori algorithm helps to set the scope for the evaluation, but for the identification of substitutes, it is counterproductive, as it prunes so many itemsets away, that in the example of pasta and salads, only one item is left over after pruning, despite setting the minimum thresholds very low. To put it in a nutshell, we can conclude, that for the computation of substitutes, the substitute ratio heuristic is best suited.

When intending to find complements, the evaluation has shown that the complement ratio heuristic comes with some major flaws: It does not control for infrequent itemsets, allowing to achieve very high scores for products that occur very seldomly in the market basket data, but in the cases, that they occur, occurring together. The same happens for the lift metric when searching for complements: Many product combinations can be found in high ranks, although or even because they count one to few purchases. This observation can be partly mitigated by the introduction of a filter, leaving out items that do not satisfy a minimum count threshold. The Apriori algorithm, however, automatically carries out pruning, filtering the data set in an intelligent and automated way. Therefore, for the identification of complements, the lift metric in combination with the Apriori algorithm is suggested, in order to find stable results, as it efficiently prunes away infrequent itemsets.

5 Managerial Recommendations

5.1 Elaborating Alternatives to Service Products

In order to derive recommendations from the findings, it is important to recall the initial research question back to our memory. The high demand during peak times for service products (freshly prepared pasta and salads) results in a bad customer experience, as a majority of the customers are faced with long waiting times in order to acquire the aforementioned service products. The question arises how we can use the knowledge gained through the analyses conducted in this paper to improve customer experience, and thus, drive customer satisfaction in order to be able to achieve a long-term increase in sales?

To begin with, one solution is to elaborate on the existing product line and generate alternatives to these service products (freshly prepared foods). In general, it would be advisable to extend the variety and production capacity of non-service products. One possible approach towards dealing with the high demand would be to promote grab-and-go solutions, for example as an alternative to the freshly prepared salad. The prepared and packaged Caesar salad represents a good substitute for fresh salads; however, it is not the number one substitute yet. The quality of this product has to be improved, in order to increase its appeal for more customers. Besides, other salads of the same kind could be provided close to the salad bar, where potential customers drop by and easily spot the alternative offering.

Secondly, the value proposition of sandwiches, baguettes and burgers should be strengthened, as they are strong contenders in terms of substitutes for pasta. Tomato & mozzarella, as well as salmon and guacamole sandwiches or tuna baguettes, for example, have shown to have a high demand and are of a high substitute quality for both pasta, and salads. It is therefore advisable to produce higher quantities of these products. Moreover, it is essential to be aware of out-of-

stock events for these products, to avoid bottlenecks in supply in times of high demand. Intelligent demand planning can help to mitigate this problem.

Thirdly, in times of high demand, i.e., around lunch time, e.g., the app provided by the store could send out push notifications to customers entering the store, suggesting to them to have a look at other high-ranked non-service products (e.g., sandwiches or baguettes). This means actively promoting non-service products when waiting times are long.

The fourth recommendation suggests that the store layout should be adjusted so that important non-service products are placed more closely to the high in demand service products. Customers may take advantage of the short distances and choose suitable alternatives that are within sight.

All these proposed solutions help to keep a high level of demand and willingness to pay and improve the store's value proposition to the customer: to provide a fast and cheap shopping possibility. In the long term, I am confident that the store will achieve an increase in sales through the realization of these suggestions.

5.2 Cross-Selling and Product Promotion

A second arising thematic block of possibilities created by the analyses is related to cross-selling and the promotion of products. Given that we can identify product complementarity, we can create menus and offer certain product combinations at a lower price when bought in bundles, compared to buying them separately. A strong contender for this case would be pairing personalized pasta with a soft drink, at a bundle price, as this product combination appears with a very high frequency. Freshly prepared pasta is the second most bought product in the store. Very often, it appears in the same basket together with coke, water, or beer. Introducing a bundle at a lower price could incentivize customers to increase their visit frequency and thus, the store's profit. Other promising bundles could be a pastry product paired with coffee or

orange juice, coffee and orange juice or chicken and rice. To this end, the store could use information panels to display the offered bundles and their prices. Likewise, methods of nudge marketing could be brought into action here, influencing customers' decisions indirectly through suggestions or reinforcement (Leonard 2008).

Furthermore, the store layout should be adapted so that complements for the most frequently bought products can be found in close proximity to them. An interesting non-service complement for ham and cheese pastry ("merenda mista"), the best-selling product in the store, could be fruit juice (see Table 12).

6 Conclusion

This paper presents an approach to explore customer missions based on patterns in purchase transactions. It adds to the existing literature in the field as it compares a heuristic method and association rule learning for the mining of product substitutes, based on the customer mission clusters established. For the identification of substitutes, the findings of this paper suggest the use of the substitute ratio heuristic. For finding complements, the best results were achieved by applying the Apriori algorithm and then identifying complements using the lift metric for association rule learning. The applicability and goodness of the approach has been proven through its application to a real-world data set in the context of a big Portuguese retailer.

The presented approach comes with great scalability in terms of automatability and computational cost. Another advantage is the explainability and comprehensibility of the approach and its results. The formulas can be well understood and easily implemented, even for large data sets.

However, the results come with certain limitations, as the study has been carried out with data from an experimental store, the environment and the use cases are very specific. The customers in the Pingo Doce & Go laboratory store are predominantly students, a customer group that displays a specific behavior differing from customer groups and missions observed in regular supermarkets. Building on the existing analyses, it would be profitable to perform customer segmentation for all stores of the company. Even though the data quality was very high, the analyses worked with purchase data collected in the relatively short period of three months. An increase in the quantity of gathered data has the potential to improve and stabilize the results. Furthermore, the findings have not been validated in real-life with the involvement of customers. A possible validation could be A/B-testing by sending out product recommendations to one group of customers and having a control group that does not receive recommendations, to observe actual customer behavior.

Future research in the area could focus on alternative forms for measuring product relationships, e.g., the investigation of lower product embeddings. For the identification of product substitutability and complementarity, representation learning methods with algorithms stemming from Natural Language Processing (NLP), like Word2Vec or Global Vectors for Word Representation (GloVe), could be applied to this type of problem. This holds great potential as these approaches can indirectly account for product characteristics, albeit at the cost of increased computational complexity.

References

Adomavicius, Gediminas, and Alexander Tuzhilin. 1999. "User Profiling in Personalization Applications through Rule Discovery and Validation." In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* . New York, New York, USA: ACM Press. https://doi.org/10.1145/312129.312287.

Aeron, Harsha, Ashwani Kumar, and Janakiraman Moorthy. 2012. "Data Mining Framework for Customer Lifetime Value-Based Segmentation." *Journal of Database Marketing & Customer Strategy Management* 19 (1): 17–30. https://doi.org/10.1057/dbm.2012.1.

Agrawal, Rakesh, Tomasz Imieliński, and Arun Swami. 1993. "Mining Association Rules between Sets of Items in Large Databases." In *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data - SIGMOD '93*, 207–16. New York, New York, USA: ACM Press. https://doi.org/10.1145/170035.170072.

Agrawal, Rakesh, and Ramakrishnan Srikant. 1994. "Fast Algorithms for Mining Association Rules." In *Proceedings of the 20th International Conference On Very Large Data Bases*.

Anderson, Joan L., Laura D. Jolly, and Ann E. Fairhurst. 2007. "Customer Relationship Management in Retailing: A Content Analysis of Retail Trade Journals." *Journal of Retailing and Consumer Services* 14 (6): 394–99. https://doi.org/10.1016/j.jretconser.2007.02.009.

Bell, David R., Daniel Corsten, and George Knox. 2011. "From Point of Purchase to Path to Purchase: How Preshopping Factors Drive Unplanned Buying." *Journal of Marketing* 75 (1): 31–45. https://doi.org/10.1509/jm.75.1.31.

Carugati, L, A Kokkinaki, and A Pouloudi. 2014. "A Data Mining-Based Framework to Identify Shopping Missions." In *Mediterranean Conference on Information Systems(MCIS)*. http://aisel.aisnet.org/mcis2014http://aisel.aisnet.org/mcis2014/20.

Chui, Chun-Kit, Ben Kao, and Edward Hung. 2007. "Mining Frequent Itemsets from Uncertain Data." In *Advances in Knowledge Discovery and Data Mining*, 47–58. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-71701-0_8.

Google Developers Documentation. 2021. "Clustering in Machine Learning." 2021. https://developers.google.com/machine-learning/clustering/algorithm/advantages-disadvantages.

Griva, Anastasia, Cleopatra Bardaki, Katerina Pramatari, and Dimitris Papakiriakopoulos. 2018. "Retail Business Analytics: Customer Visit Segmentation Using Market Basket Data." *Expert Systems with Applications* 100 (June): 1–16. https://doi.org/10.1016/j.eswa.2018.01.029.

Ho Cho, Yoon, Jae Kyeong Kim, and Soung Hie Kim. 2002. "A Personalized Recommender System Based on Web Usage Mining and Decision Tree Reduction." *Expert Systems with Applications* 23. https://doi.org/10.1016/S0957-4174(02)00052-0.

Huang, Jess, Sajal Kohli, and Shruti Lal. 2018. "Winning in an Era of Unprecedented Disruption: A Perspective on US Retail."

Huang, Zhexue. 1998. "Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values." *Data Mining and Knowledge Discovery* 12: 283–304.

Hussein, Nada, Abdallah Alashqur, and Bilal Sowan. 2015. "Using the Interestingness Measure Lift to Generate Association Rules." *Journal of Advanced Computer Science & Technology* 4 (1): 156. https://doi.org/10.14419/jacst.v4i1.4398.

Jaccard, Paul. 1912. "The Distribution of the Flora in the Alpine Zone.1." *New Phytologist* 11 (2): 37–50. https://doi.org/10.1111/j.1469-8137.1912.tb05611.x.

Khajvand, Mahboubeh, Kiyana Zolfaghar, Sarah Ashoori, and Somayeh Alizadeh. 2011. "Estimating Customer Lifetime Value Based on RFM Analysis of Customer Purchase Behavior: Case Study." *Procedia Computer Science* 3: 57–63. https://doi.org/10.1016/j.procs.2010.12.011.

Koch, Daniel, Lars Marcus, Jesper Steen, Jorge Gil, Eime Tobari, Maia Lemlij, Anna Rose, and Alan Penn. 2009. "The Differentiating Behaviour of Shoppers Clustering of Individual Movement Traces in a Supermarket." In *Proceedings of the 7th International Space Syntax Symposium*.

Kotsiantis, Sotiris, and Dimitris Kanellopoulos. 2006. "Association Rules Mining: A Recent Overview." In *GESTS International Transactions on Computer Science and Engineering*, 71–82.

Leonard, T.C. Richard H. Thaler, Cass R. Sunstein. 2008. "Nudge: Improving Decisions about Health, Wealth, and Happiness." *Constitutional Political Economy* 19 (4): 356–60. https://doi.org/10.1007/s10602-008-9056-2.

Matthew, Ridge, Allan Johnston Kevin, and Donovan Brian. 2015. "The Use of Big Data Analytics in the Retail Industries in South Africa." *African Journal of Business Management* 9 (19): 688–703. https://doi.org/10.5897/AJBM2015.7827.

Mungoli, Abhishek. 2020. "Towards Data Science." Retail Analytics: A Novel and Intuitive Way of Finding Substitutes and Complements. March 15, 2020. https://towardsdatascience.com/retail-analytics-a-novel-and-intuitive-way-of-finding-substitutes-and-complements-c99790800b42.

Puka, Radosław, and Stanislaw Jedrusik. 2021. "A New Measure of Complementarity in Market Basket Data." *Journal of Theoretical and Applied Electronic Commerce Research* 16 (4): 670–81. https://doi.org/10.3390/jtaer16040039.

Ramesh, Sudarshan Nadamuni. 2019. “Market Basket Analysis - Why ‘Lift’ Is an Odd Metric.” Https://Www.Linkedin.Com/Pulse/Market-Basket-Analysis-Why-Lift-Odd-Metric-Nadamuni-Ramesh/. August 18, 2019.

Ren, Shuyun, Hau-Ling Chan, and Tana Siqin. 2020. “Demand Forecasting in Retail Operations for Fashionable Products: Methods, Practices, and Real Case Study.” *Annals of Operations Research* 291 (1–2): 761–77. https://doi.org/10.1007/s10479-019-03148-8.

Sarantopoulos, Panagiotis, Aristeidis Theotokis, Katerina Pramatari, and Georgios Doukidis. 2016. “Shopping Missions: An Analytical Method for the Identification of Shopper Need States.” *Journal of Business Research* 69 (3): 1043–52. https://doi.org/10.1016/j.jbusres.2015.08.017.

Walters, Rockney G, and Maqbul Jamil. 2003. “Exploring the Relationships between Shopping Trip Type, Purchases of Products on Promotion, and Shopping Basket Profit.” *Journal of Business Research* 56 (1): 17–29. https://doi.org/10.1016/S0148-2963(01)00201-6.

Wu, Jing, and Zheng Lin. 2005. “Research on Customer Segmentation Model by Clustering.” In *International Conference on Electronic Commerce*, 316–18.

Appendix A: Figures

Figure 1: Distortion Score Elbow for K-Means Clustering

Note: Distortion Score at each value for k between 2 and 20.

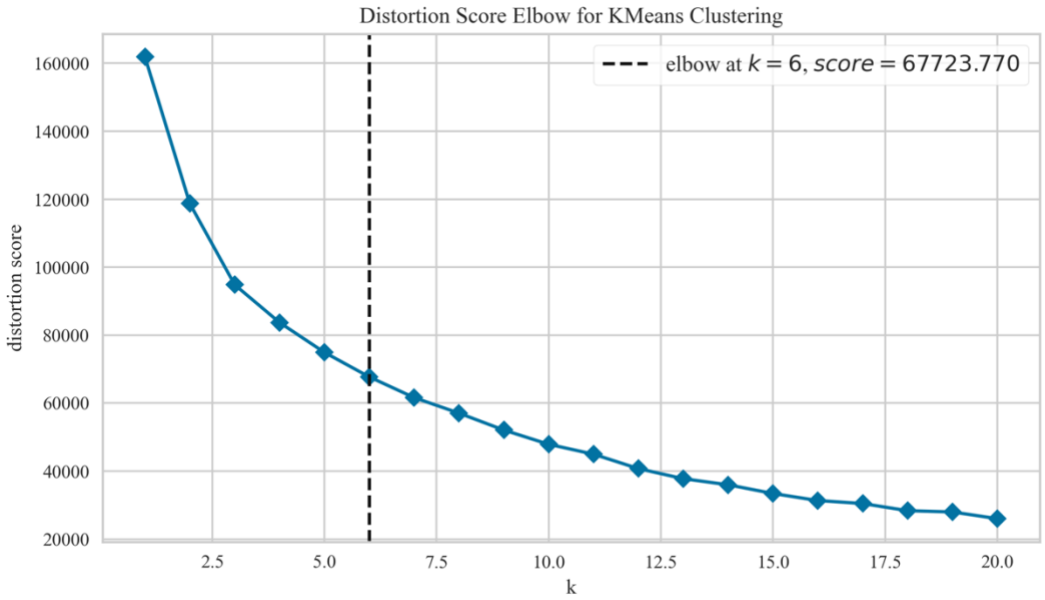


Figure 2: Intraday Cluster Occurrences

Note: Number of Transactions per cluster per hour.

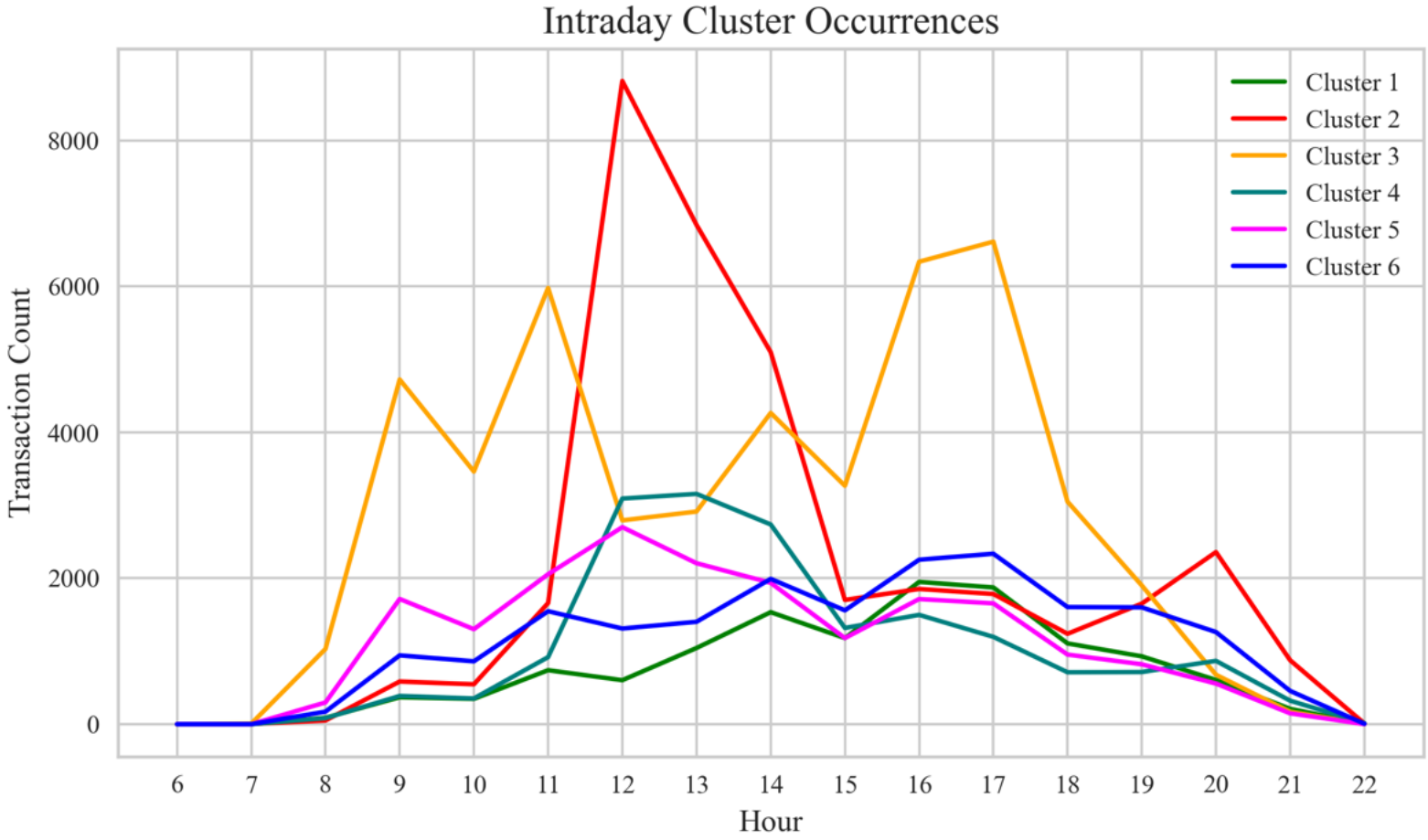
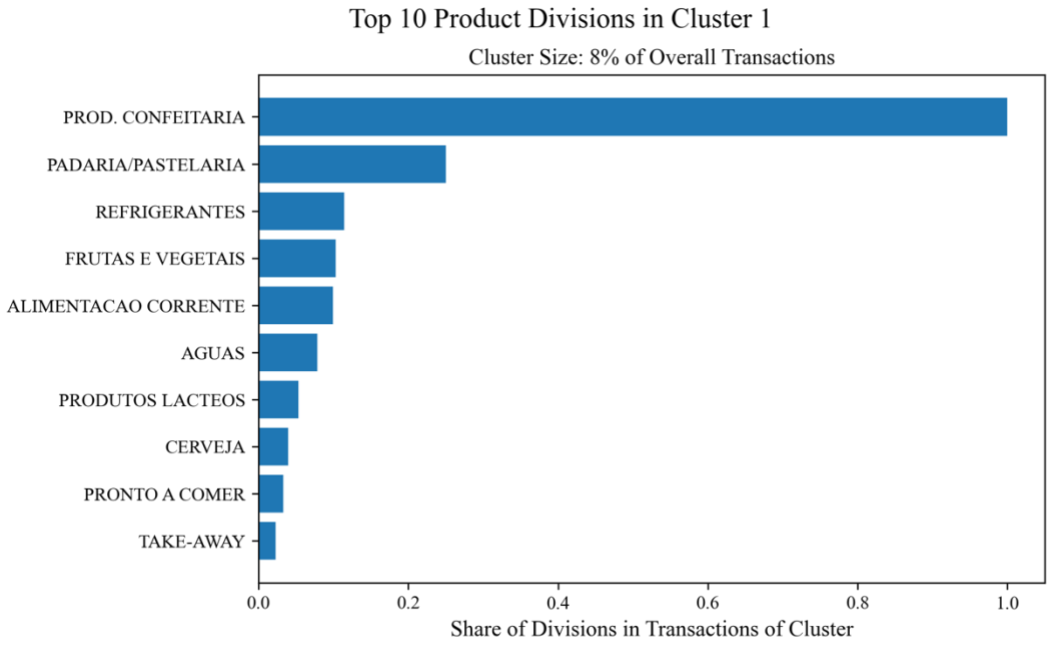
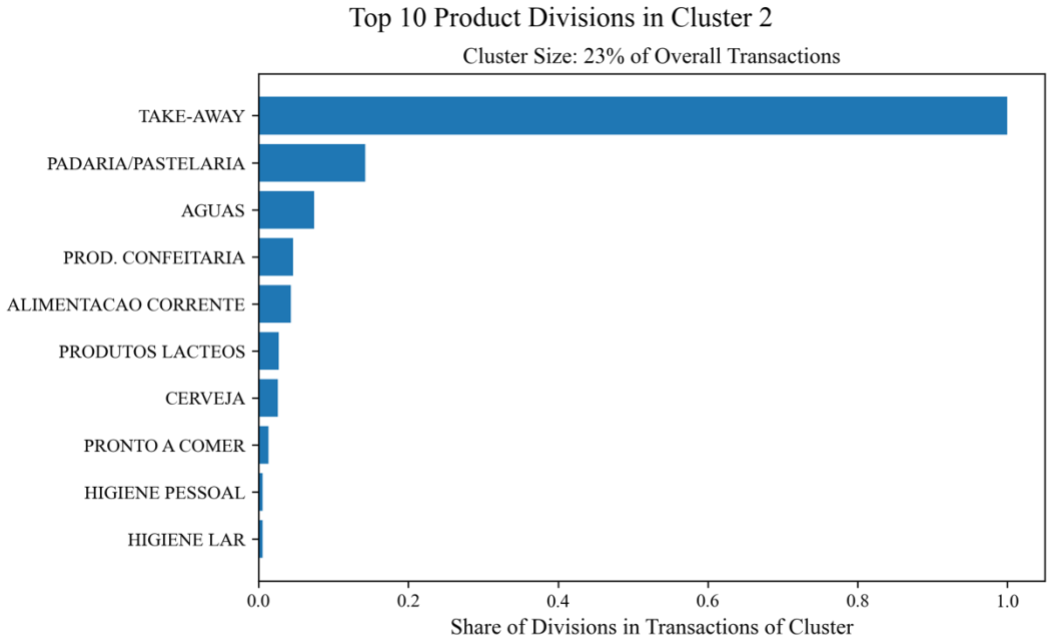
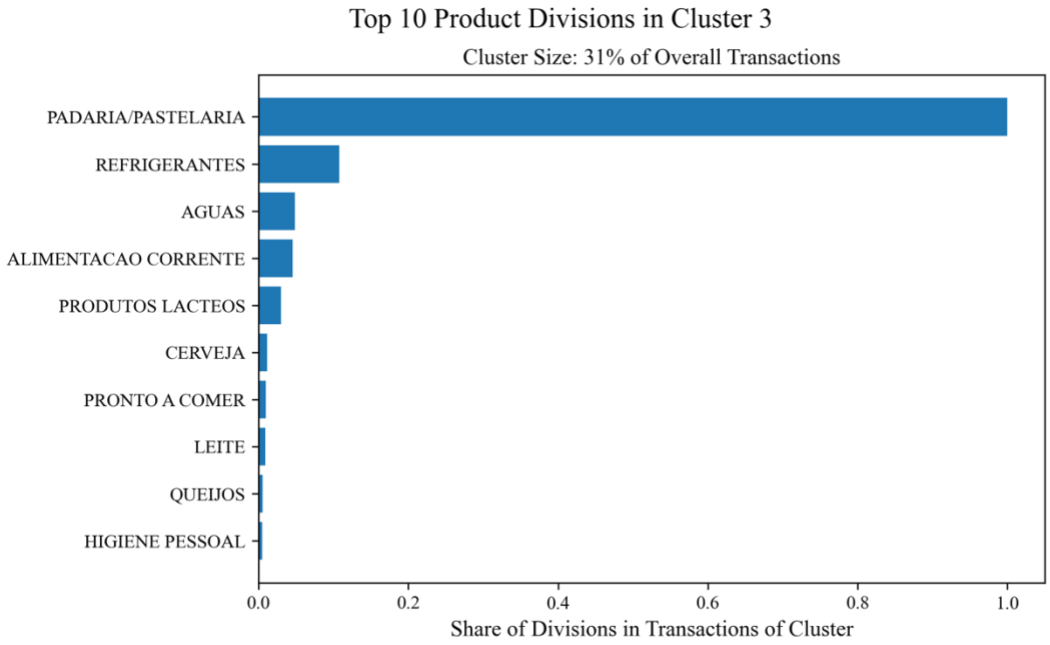


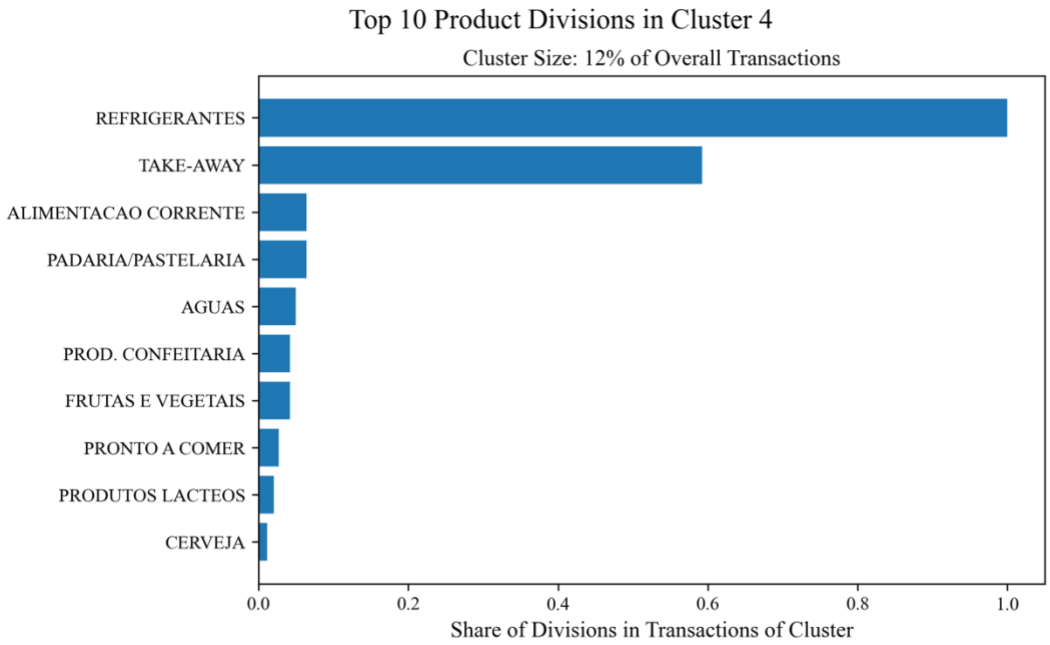
Figure 3: Product Divisions Share per Cluster

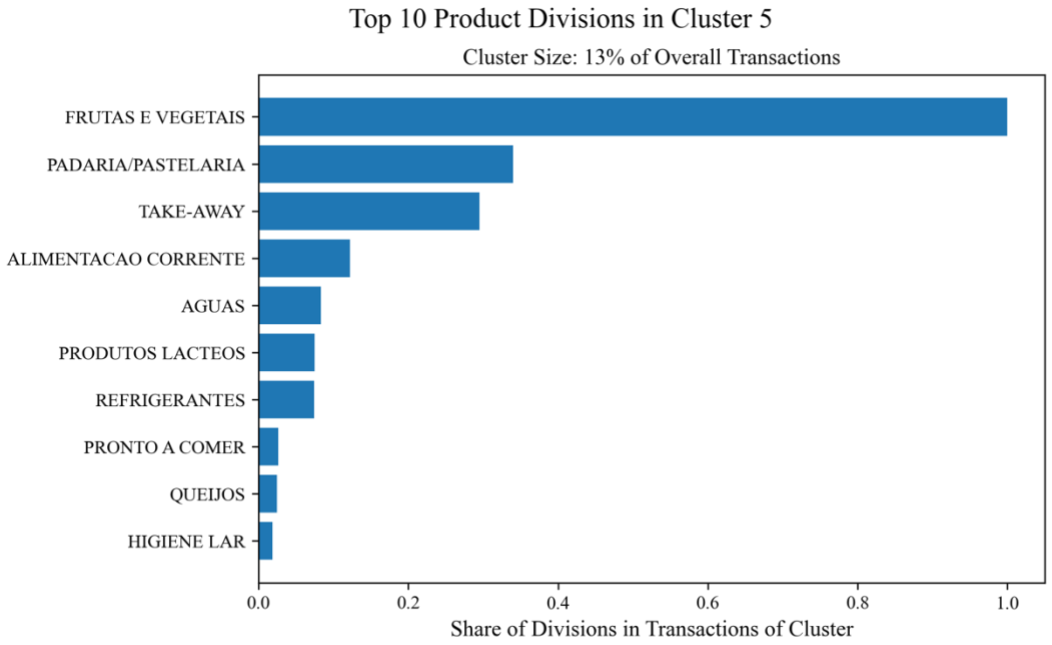
Note: Share of transaction in each cluster that contain a specific product division (most frequent product divisions in transactions per cluster). Values are sorted in descending order. The subtitle of each plot contains information about the relative size each cluster has, looking at all transactions in the data set.

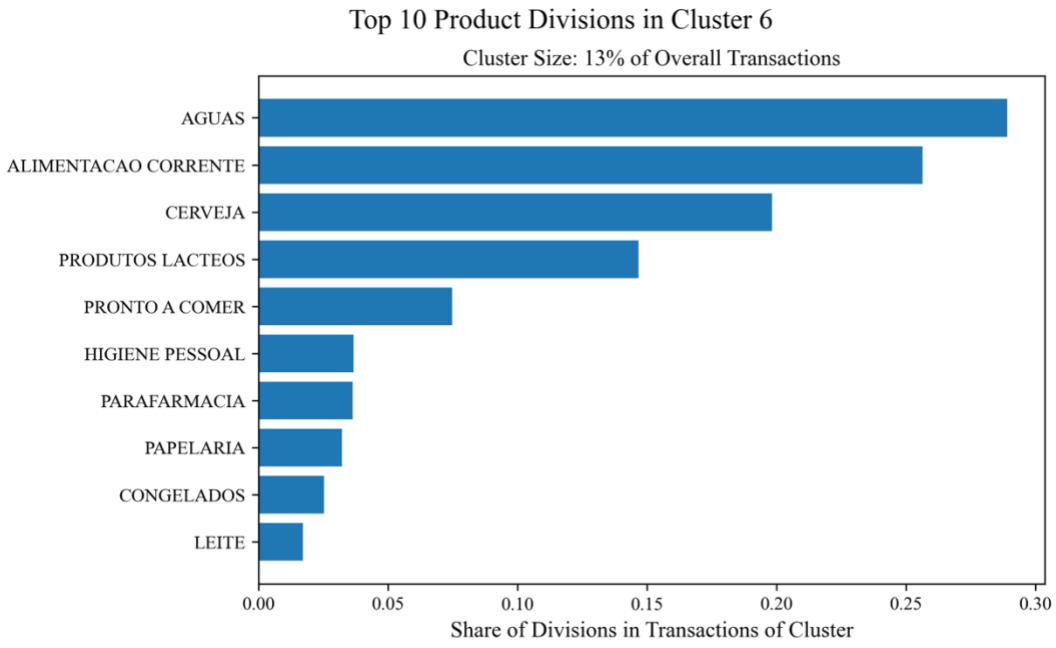





Top 10 Product Divisions in Cluster 3
Cluster Size: 31% of Overall Transactions
PADARIA/PASTELARIA
REFRIGERANTES
AGUAS
ALIMENTACAO CORRENTE
PRODUTOS LACTEOS
CERVEJA
PRONTO A COMER
LEITE
QUEIJOS
HIGIENE PESSOAL
0.0
0.2
0.4
0.6
0.8
1.0
Share of Divisions in Transactions of Cluster


Top 10 Product Divisions in Cluster 4
Cluster Size: 12% of Overall Transactions
REFRIGERANTES
TAKE-AWAY
ALIMENTACAO CORRENTE
PADARIA/PASTELARIA
AGUAS
PROD. CONFEITARIA
FRUTAS E VEGETAIS
PRONTO A COMER
PRODUTOS LACTEOS
CERVEJA
0.0
0.2
0.4
0.6
0.8
1.0
Share of Divisions in Transactions of Cluster


Top 10 Product Divisions in Cluster 5
Cluster Size: 13% of Overall Transactions
FRUTAS E VEGETAIS
PADARIA/PASTELARIA
TAKE-AWAY
ALIMENTACAO CORRENTE
AGUAS
PRODUTOS LACTEOS
REFRIGERANTES
PRONTO A COMER
QUEIJOS
HIGIENE LAR
0.0
0.2
0.4
0.6
0.8
1.0
Share of Divisions in Transactions of Cluster


Top 10 Product Divisions in Cluster 6
Cluster Size: 13% of Overall Transactions
AGUAS
ALIMENTACAO CORRENTE
CERVEJA
PRODUTOS LACTEOS
PRONTO A COMER
HIGIENE PESSOAL
PARAFARMACIA
PAPELARIA
CONGELADOS
LEITE
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Share of Divisions in Transactions of Cluster

Appendix B: Tables

Table 1: Snapshot of the Binary Encoded Market Basket Data Matrix

Note: txs_id denotes the transaction ID, cat1 – cat34 are the 34 product categories (water, beer, take-away, etc.). The single variables give information about whether the specific product category can be found in the respective transaction or not.

txs_id	cat_1	cat_2	cat_3	cat_4	…	cat_32	cat_33	cat_34
1	0	1	0	1	…	0	1	0
2	1	0	1	0	…	0	0	0
3	0	0	1	0	…	1	0	1

Table 2: Top 10 Most Frequently Sold Products per Cluster

Note: Each cluster has its own table, depicting the most frequently bought products in descending order, also listing how often absolutely and relatively they can be found within the cluster.

Cluster 1	Product Description	Product ID	Count	Relative Frequency
1	BOL COOKIE MINI PD CHOC 125G	533455	666	5.29%
2	PAST EL DR PD PEPPERM 26,1G	700790	652	5.17%
3	BOL COB PD ARGOL CHOC BRANC 150G	661852	384	3.05%
4	TAB CHOC LEITE KINDER CHOC BARR T8 100G	32333	356	2.83%
5	BOL COOKIE PD 150G	662745	317	2.52%
6	PASTILHAS MINI STICK SPEARMINT PD 28G	808852	295	2.34%
7	BANANA UNIDADE LAB	934947	248	1.97%
8	PÃO DE QUEIJO UN	256978	247	1.96%
9	BOL ESP FIN PD PETIT BISCUIT LEITE 150G	533452	246	1.95%
10	MERENDA MISTA 95 G	254381	230	1.83%

Cluster 2	Product Description	Product ID	Count	Relative Frequency
1	MASSA PERSONALIZADA (4 INGREDIENTES)	894334	4556	12.99%
2	SANDES MOZZARELLA E TOMATE	887590	1526	4.35%
3	SANDES DE SALMÃO FUMADO & GUACAMOLE	879460	1408	4.01%
4	HAMBURGUER ANGUS COM QJ CHEDDAR	902975	1069	3.05%
5	SALADA CAESAR 210GR	906262	1044	2.98%
6	PIZZA FIAMBRE E QUEIJO FORNO	884039	938	2.67%
7	BAGUETE LUSITANA DE PASTA DE ATUM	887583	909	2.59%
8	PIZZA FRANGO FORNO	884044	887	2.53%
9	ARROZ BRANCO	10005269	882	2.51%
10	HAMBURGUER ANGUS, CHEDDAR & BACON	902976	853	2.43%

Cluster 3	Product Description	Product ID	Count	Relative Frequency
1	MERENDA MISTA 95 G	254381	5161	10.93%
2	PÃO DE QUEIJO UN	256978	3961	8.39%
3	CAPUCCINO	901250	3500	7.41%
4	AMERICANO	901241	3167	6.71%
5	FOLHADO DE SALSICHA 100 G	902807	2919	6.18%
6	EXPRESSO DUPLO	901239	2880	6.10%
7	MERENDA DE CHOCOLATE 80 G	738198	2532	5.36%
8	CROISSANT COM CHOCOLATE 100 G	254380	2462	5.22%
9	EXPRESSO	901238	2102	4.45%
10	LATTE	901247	1895	4.01%

Cluster 4	Product Description	Product ID	Count	Relative Frequency
1	COCA COLA ORIGINAL LATA 33CL	922602	1785	10.27%
2	COCA COLA SEM AÇUCAR LATA 33CL	720920	1492	8.58%
3	MASSA PERSONALIZADA (4 INGREDIENTES)	894334	1335	7.68%
4	COCA COLA SEM AÇUCAR S/CAFEÍNA LATA 33CL	720921	863	4.96%
5	SMOOTHIE COM FRUTOS VERMELHOS PD 33CL	882914	615	3.54%
6	BEB.ENERGETICA RED BULL 25CL	100463	590	3.39%
7	ICED TEA PESSEGO PD 0,5LT	902737	578	3.32%
8	SMOOTHIE COM FRUTOS TROPICAIS PD 33 CL	882915	561	3.23%
9	BEB.ENERGETICA RED BULL SUGAR FREE 25CL	445129	560	3.22%
10	COCA COLA ORIGINAL PET 50CL	357809	486	2.80%

Cluster 5	Product Description	Product ID	Count	Relative Frequency
1	SUMO LARANJA 330 ML FRUTA	933371	4473	23.24%
2	BANANA UNIDADE LAB	934947	3407	17.70%
3	MAÇA ROYAL GALA CAL 75 BIO UN LAB CPG10	896222	1168	6.07%
4	BANANA IMP UN BIO LAB CCG	896258	1130	5.87%
5	ABACAXI CUBOS LAB	896297	993	5.16%
6	MAÇÃ GRANNY SMITH BIO UN LAB CPG10	896223	868	4.51%
7	MORANGOS CORTADOS LAB	896313	823	4.28%
8	TOMATE MINI ALONGADO SNACK 250GR CCP	920311	800	4.16%
9	MELAO VERDE CUBOS LAB	896316	732	3.80%
10	SUMO LARANJA 1000 ML	804534	600	3.12%

Cluster 6	Product Description	Product ID	Count	Relative Frequency
1	CERVEJA C/ALC SUPER BOCK LATA 33CL	826004	1568	8.12%
2	ÁGUA LUSO 50CL	534504	960	4.97%
3	ÁGUA ECO 1,5LT	902886	807	4.18%
4	AGUA PINGO DOCE 1,5LT	598441	791	4.10%
5	ÁGUA LUSO SPORT 75CL	52508	590	3.05%
6	AGUA MONCHIQUE 1,5LT	52631	564	2.92%
7	LASANHA REFRIGERADA PDOCE BOLONHESA 400G	647928	560	2.90%
8	CERVEJA C/ALC SAGRES LATA 33CL	60636	523	2.71%
9	ÁGUA LUSO 1,5LT	700041	502	2.60%
10	ÁGUA MONCHIQUE PET 100% RECICLADO 72CL	923977	378	1.96%

Table 3: Substitutes for Freshly Prepared Pasta

Note: Substitutes for Pasta, ordered in descending order by the substitute quality, which means, ascending substitute ratio (the first element is the best substitute). Columns 2 and 3 show the results of the heuristic approach, columns 4 and 5 are the results obtained by applying the lift metric (in ascending order, as well). Lower values are better.

	Substitute Ratio	Substitute Ratio	Substitute Lift	Lift
1	BAGUETE LUSITANA DE PASTA DE ATUM	0.00110	BAGUETE LUSITANA DE PASTA DE ATUM	0.0138
2	BAGUETE LUSITANA PRESUNTO&QUEIJO FRESCO	0.00132	BAGUETE LUSITANA PRESUNTO&QUEIJO FRESCO	0.0166
3	PIZZA PRESUNTO FORNO	0.00156	PIZZA PRESUNTO FORNO	0.0197
4	SANDES DE SALMÃO FUMADO & GUACAMOLE	0.00213	SANDES DE SALMÃO FUMADO & GUACAMOLE	0.0268
5	ARROZ BRANCO	0.00227	ARROZ BRANCO	0.0285
6	HAMBURGUER ANGUS, CHEDDAR & BACON	0.00234	HAMBURGUER ANGUS, CHEDDAR & BACON	0.0295
7	WRAP VEGGIE	0.00253	WRAP VEGGIE	0.0319
8	SALADA CAPRI	0.00260	SALADA CAPRI	0.0327
9	CROQUETE DE CARNE 2 UN	0.00278	CROQUETE DE CARNE 2 UN	0.0350
10	WRAP DE ATUM	0.00281	WRAP DE ATUM	0.0354
11	SANDES MISTA	0.00282	SANDES MISTA	0.0355
12	COXINHA DE FRANGO 2 UN	0.00300	PIZZA CALZONE FORNO	0.0379
13	PIZZA 4 QUEIJOS FORNO	0.00300	SANDES MOZZARELLA E TOMATE	0.0412
14	SANDES DE PRESUNTO, BRIE E COMPOTA	0.00300	PIZZA 4 QUEIJOS FORNO	0.0425
15	PIZZA FIAMBRE E QUEIJO FORNO	0.00300	HAMBURGUER VEGETARIANO	0.0448
16	PIZZA CARBONARA FORNO	0.00300	ARROZ DE PATO UNI	0.0479

17	SALADA PERSONALIZADA (4 INGREDIENTES)	0.00300	SANDES DE FIAMBRE DE FRANGO	0.0480
18	POKÉ DE SALMÃO & MANGA	0.00300	SANDES DELÍCIAS DO MAR	0.0520
19	PERNA FRANGO ASSADO	0.00300	SANDES FRANGO	0.0543
20	WRAP VEGGIE	0.00300	LEGUMES Á BRÁS UNI	0.0577

Table 4: Substitutes for Freshly Prepared Salad

Note: Substitutes for salads, ordered in descending order by the substitute quality, which means, ascending substitute ratio (the first element is the best substitute). Columns 2 and 3 (in grey) show the results of the heuristic approach, columns 4 and 5 are the results obtained by applying the lift metric (in ascending order, as well). Lower values are better.

	Substitute Ratio	Substitute Ratio	Substitute Lift	Lift
1	HAMBURGUER ANGUS COM QJ CHEDDAR	0.00139	SANDES MOZZARELLA E TOMATE	0.0523
2	PIZZA FRANGO FORNO	0.00139	HAMBURGUER ANGUS COM QJ CHEDDAR	0.0747
3	BAGUETE LUSITANA DE PASTA DE ATUM	0.00139	PIZZA FIAMBRE E QUEIJO FORNO	0.0852
4	SANDES MOZZARELLA E TOMATE	0.00139	BAGUETE LUSITANA DE PASTA DE ATUM	0.0879
5	PIZZA FIAMBRE E QUEIJO FORNO	0.00139	PIZZA FRANGO FORNO	0.0901
6	SANDES DE FIAMBRE DE FRANGO	0.00190	SANDES DE FIAMBRE DE FRANGO	0.1522
7	PERNA FRANGO ASSADO	0.00206	SALADA CAESAR 210GR	0.1530
8	CHAMUÇA DE CARNE 2 UN	0.00255	PERNA FRANGO ASSADO	0.1647
9	SALADA CAESAR 210GR	0.00279	WRAP VEGGIE	0.2023
10	WRAP VEGGIE	0.00279	CHAMUÇA DE CARNE 2 UN	0.2038
11	PIZZA QJ/FIAMB/COG FORNO	0.00279	FRANGO ASSADO METADE	0.2041
12	FRANGO ASSADO METADE	0.00279	PIZZA QJ/FIAMB/COG FORNO	0.2105
13	MASSA PERSONALIZADA (5 INGREDIENTES)	0.00300	MASSA PERSONALIZADA (4 INGREDIENTES)	0.2279
14	DOSE CALDO VERDE	0.00408	MASSA PERSONALIZADA (5 INGREDIENTES)	0.2399
15	SANDES DELÍCIAS DO MAR	0.00413	SANDES DE SALMÃO FUMADO & GUACAMOLE	0.2645
16	WRAP DE ATUM	0.00421	DOSE CALDO VERDE	0.3261
17	IOGURTE COM MIRTILOS	0.00575	SANDES DELÍCIAS DO MAR	0.3301
18	PIZZA CARBONARA FORNO	0.00649	WRAP DE ATUM	0.3366
19	SANDES MISTA	0.00649	ARROZ BRANCO	0.4223
20	SANDES DE PRESUNTO, BRIE E COMPOTA	0.00649	IOGURTE COM MIRTILOS	0.4591

Table 5: Substitutes for Pasta (Apriori-Pruned)

Note: Substitutes for freshly prepared pasta, applying the lift metric, Apriori-pruned; descending quality (first line depicts best substitute). Lower values are better.

	Substitute	Lift
1	PÃO DE QUEIJO UN	0.57
2	PÃO COM AZEITONAS E AZEITE 55 G	0.89
3	CERVEJA C/ALC SAGRES LATA 33CL	1.71
4	SACO DE PAPEL COMIDA FRESCA FSC MISTO	1.84
5	ÁGUA LUSO 50CL	1.85
6	ÁGUA LUSO 1,5LT	2.00
7	ÁGUA LUSO SPORT 75CL	2.00
8	AGUA PINGO DOCE 1,5LT	2.05
9	ÁGUA ECO 1,5LT	2.08
10	AGUA MONCHIQUE 1,5LT	2.18
11	AGUA EVIAN 75CL	2.23
12	CERVEJA C/ALC SUPER BOCK LATA 33CL	2.27
13	BEB PROTEIN LEMON LIMA LIMÃO PROZIS 500M	2.38
14	ÁGUA MONCHIQUE PET 100% RECICLADO 72CL	2.67

Table 6: Substitutes for Salads (Apiori-Pruned)

Note: Substitutes for freshly prepared salads, applying the lift metric, Apriori-pruned.

	Substitute	Lift
1	PÃO COM AZEITONAS E AZEITE 55 G	6.19

Table 7: Top Complement Pairs in the Meal Deal Cluster (Heuristic, Filtered)

Note: Top 10 complement pairs in the meal deal cluster, following the complement ratio; descending quality (first line depicts best complement pair). Higher values are better.

	Product 1	Product 2	Complement Ratio
1	FRANGO ASSADO METADE	ARROZ BRANCO	0.095
2	DOSE PANADOS	ARROZ BRANCO	0.075
3	PERNA FRANGO ASSADO	ARROZ BRANCO	0.059
4	ARROZ BRANCO	CORDON BLEU 2 UN	0.043
5	COXINHA DE FRANGO 2 UN	ARROZ BRANCO	0.040
6	MERENDA MISTA 95 G	FOLHADO DE SALSICHA 100 G	0.032
7	PIZZA QJ/FIAMB/COG FORNO	SERVIÇO DE FORNO	0.032
8	CHAMUÇA DE CARNE 2 UN	ARROZ BRANCO	0.032
9	ARROZ BRANCO	EMPADA DE GALINHA 75GR 2 UN	0.029
10	PÃO COM AZEITONAS E AZEITE 55 G	PÃO DA AVÓ 100 GR	0.028

Table 8: Top Complement Pairs in the Meal Deal Cluster (Lift-Based, Filtered)

Note: Top 10 complement pairs in the meal deal cluster, following the lift metric; descending quality (first line depicts best complement pair). Higher values are better.

	Product 1	Product 2	Lift
1	DOSE PANADOS	ARROZ BRANCO	19.41
2	MERENDA MISTA 95 G	FOLHADO DE SALSICHA 100 G	12.22
3	FRANGO ASSADO METADE	ARROZ BRANCO	11.96
4	PÃO COM AZEITONAS E AZEITE 55 G	PÃO DA AVÓ 100 GR	11.04
5	PERNA FRANGO ASSADO	ARROZ BRANCO	10.19
6	PIZZA QJ/FIAMB/COG FORNO	SERVIÇO DE FORNO	9.56
7	ARROZ BRANCO	CORDON BLEU 2 UN	9.48
8	DOSE SOPA JULIANA	EMPADA DE GALINHA 75GR 2 UN	9.27
9	RISSOL DE CAMARÃO 2 UN	DOSE SOPA DE LEGUMES	8.49
10	RISSOL DE CAMARÃO 2 UN	ARROZ BRANCO	8.47

Table 9: Top Complement Pairs in the Meal Deal Cluster (Lift-Based, Apriori-Pruned)

Note: Top 10 complement pairs in the meal deal cluster, Apriori-pruned, following the lift metric; descending quality (first line depicts best complement pair). Higher values are better.

	Product 1	Product 2	Lift
1	SERVIÇO DE FORNO	PIZZA FRS S/GLUT S/LACT QJ/FIAMB PD 420G	73.48
2	QJ FRS PINGO DOCE 200G	SALADA DE MASSA & ATUM	38.15
3	ÁGUA ECO 1,5LT	GARRAFAO REUTILIZAVEL ECO 1,5LTS	23.05
4	ARROZ BRANCO	DOSE ALMONDEGAS	12.05
5	ARROZ BRANCO	DOSE ESPETADA DE COXA DE FRANGO	9.48
6	DOSE CHILLI VEGETARIANO	ARROZ BRANCO	8.63
7	SERVIÇO DE FORNO	PIZZA QJ/FIAMB/COG FORNO	8.28
8	DOSE PANADOS	ARROZ BRANCO	6.91
9	SALADA VITA MINUTE GREGA	SALADA VITA MINUTE GREGA	6.19
10	PÃO COM AZEITONAS E AZEITE 55 G	SALADA PERSONALIZADA (4 INGREDIENTES)	6.18

Table 10: Top Substitute Pairs in the Coffee Snack Cluster (Heuristic)

Note: Top 10 substitute pairs in the coffee snack cluster, following the substitute ratio; descending quality (first line depicts best substitute pair). Lower values are better.

	Product 1	Product 2	Substitue Ratio
1	CROISSANT COM SEMENTES	PÃO C CHOURIÇO E QUEIJO UN 135 G	0.001637
2	QUEQUE 55 G	FLAT WHITE	0.001742
3	QUEQUE 55 G	LATTE DE SOJA	0.001742
4	QUEQUE 55 G	CAPUCCINO COM LEITE DE SOJA	0.001742
5	DAWNUT'S RECHEADO C/ CHOCOLATE 75 G	FLAT WHITE	0.001751
6	DAWNUT'S RECHEADO C/ CHOCOLATE 75 G	PASTEL NATA GOURMET 70 G	0.001751
7	CROISSANT COM CREME 100 G	SUMO LARANJA NATURAL 250 ML	0.001754
8	BOLO DE ARROZ	2 X EXPRESSO	0.001887
9	QUEQUE 55 G	2 X EXPRESSO	0.001887
10	2 X EXPRESSO	FLAT WHITE	0.001887

Table 11: Top Complement Pairs in the Coffee Snack Cluster (Lift-Based, Apriori-Pruned)

Note: Top 10 complement pairs in the coffee snack cluster, Apriori-pruned, following the lift metric; descending quality (first line depicts best complement pair). Higher values are better.

	Product 1	Product 2	Lift
1	SUMO LARANJA NATURAL 250 ML	CROISSANT BRIOCHE MISTO	66.33
2	SUMO LARANJA NATURAL 250 ML	PÃO DA AVÓ MISTO	55.76
3	LATTE DESCAFEINADO	FOLHADO COM QUEIJO E GOIABA 90 G	29.30
4	ÁGUA ECO 1,5LT	GARRAFAO REUTILIZAVEL ECO 1,5LTS	19.65
5	PÃO DA AVÓ 100 GR	AGUA ECO 6LTS	17.12
6	HUMMUS PINGO DOCE180G	PÃO 8 CEREAIS 90 G	13.96
7	FIAMBRE PERNA EXTRA PDOCE FAT FINAS 150G	PÃO DA AVÓ 100 GR	13.33
8	SUMO LARANJA NATURAL 250 ML	EXPRESSO	12.22
9	CROISSANT BRIOCHE MISTO	EXPRESSO	12.02
10	EXPRESSO	PÃO DA AVÓ MISTO	10.61

Table 12: Complements for “Merenda Mista” (Lift-Based, Apriori-Pruned)

Note: Top 10 complements for “merenda mista”, Apriori-pruned, following the lift metric; descending quality (first line depicts best complement). Higher values are better.

	Complement	Lift
1	SUMO 100% MAÇÃ PINGO DOCE 20CL	3.30
2	SUMO ESPREMIDO ANANAS PD 750ML	2.68
3	ICED TEA MANGA PD 0,5LT	2.19
4	ICE TEA LIPTON PESSEGO LATA 25CL	2.10
5	LEITE PASTEURIZADO VIGOR CHOCOLATE 200ML	1.98
6	NECTARÍSSIMO PINGO DOCE PÉSSEGO 25CL	1.84
7	COCA COLA ORIGINAL LATA 33CL	1.53
8	NECTARÍSSIMO PINGO DOCE MANGA 25CL	1.50
9	COMPAL VITAL MANGA/LARANJA 33CL	1.50
10	COMPAL CLASSICO NECTAR PESSEGO 33CL	1.48

A Work Project, presented as part of the requirements for the Award of a Master's degree in Business Analytics from the Nova School of Business and Economics.

FIELD LAB JERÓNIMO MARTINS: OPTIMIZATION OF RETAIL OPERATIONS

JOAO RAIMUNDO

JACOB LIND

SIMON LINK

LUKAS BERNDT

Work project carried out under the supervision of:

Qiwei Han

17-12-2021

1 Field Lab

Driven by increased competition, low margins are the key environment merchants are facing nowadays when doing business in the retail industry (Vaja 2015). Therefore, retailers are faced with an urgent need for continuous improvements, fostering disruptions in products, services, and operations. Among the most recent innovations are cashierless supermarkets, leveraging on the latest technologies including IoT devices, numerous sensors, cameras, and machine learning techniques (Ponte and Bonazzi 2021). This next generation of brick-and-mortar stores will compete in this fought-over environment together with online retailers and on-demand delivery services that have been emerging quickly in recent years.

Stores like the Pingo Doce & Go Nova laboratory store, brought up by the Portuguese retail giant Jerónimo Martins, utilize these new kinds of technologies, and start exploiting the new information in the data generated. This allows to detect, address, and solve problems, inefficiencies, and areas with further opportunities for improvement. Data-driven decision support systems thereby directly translate into bottom-line profitability improvements (Vassakis, Petrakis, and Kopanakis 2018).

After launching the experimentation store at Nova SBE, Jerónimo Martins identified two major areas for improvement, namely gathering knowledge about customer behavior and the shopping process, and issues that hamper the retailer's operations from running smoothly. In detail, this field lab covers four areas of interest. First, process mining techniques are applied to get a deeper understanding about the purchase processes of in-store shopping. Second, market basket analysis is performed on shopping mission clusters in order to identify substitute and complement products with specific focus on freshly prepared food items. Third, a demand planning tool is developed to support the day-to-day production planning for ultra-perishable

items. Fourth, a semi-supervised learning fault detection model for oven is developed to identify false preparations.

The motivation for the store to focus on these aspects are manifold. Identifying and modelling a standard shopping process and gaining insights into shopping habits of customers enables the business to respond more appropriately to customer needs, behaviors, and preferences, ultimately enhancing the in-store experience for increased customer satisfaction. By revealing relationships between products, marketing actions can be taken to steer demand and provide recommendations for the store's assortment and replenishment strategy. Introducing a data-driven demand planning tool, comprising demand forecasting and operational planning, allows to optimize the trade-off between product availability and food waste. Detecting faults and ensuring conformity with food safety regulations is fundamental for the store's operationability and reduces inefficiencies with regards to energy consumption and food waste.

All proposed measures are likely to contribute to sales or costs in a favorable manner, implying significant upside potentials for the store's profitability.

2 Institutional Background

With the Pingo Doce & Go Store opening at Nova School of Business and Economics (SBE) in Carcavelos in October 2018, the Portuguese retailer Jerónimo Martins has introduced his vision of the future retail to the market (Salgueiro 2019). Initially inspired by the requests and needs of students for an extended range of cheap and convenient meal offerings at the old campus at Nova SBE, Jerónimo Martins started with the planning and opening of the store already before the new campus of Nova SBE in Carcavelos was inaugurated.

Having predominantly students – young, tech-savvy people – as customers, the store presents an optimal opportunity to experiment with innovative technologies and business models.

Solutions that have proven themselves to be demanded and accepted in this laboratory-style store qualify for being potentially rolled out to other stores of the group.

Fully packed with the latest technologies, a completely new shopping experience is provided on over 250 sqm in the brick-and-mortar store at Nova SBE (Salgueiro 2019). It is designed to serve the customer needs for convenience and freshness, providing a pleasant shopping experience in less than a minute. Besides traditional grocery articles for everyday use, the store offers a wide variety of both on-demand and takeaway freshly prepared food items at low prices, which directly compete with other on- and off-campus food options (Caetano 2019).

The typical shopping process starts and ends within the app, a key component throughout the purchase. After a self-check-in with their mobile devices, the customers use the app to self-reliantly add their desired articles to their virtual baskets through either scanning or NFC, creating a quick and convenient customer experience. Having finished their item selection, the customer does not finish his purchase through a physical checkout but either by in-app payment or at the payment stand.

Throughout the process and beyond, data is created at various points. Apart from gathering data about customers at the point of registration, every action done through the app is saved by the system as an event log. Moreover, the store generates production data through its IoT devices, ranging from ovens to kitchen devices that are connected to the internet. By the fact that every purchase can be assigned to a specific user, we have access to a detailed transactional purchase database that can investigate customer habits in an unprecedented way in the current retail context. Furthermore, the system stores data about article details, inventory movements, as well as cleaning and hygiene log data. This never-seen-before variety and volume of data allow us to profoundly dive into various analyses about operational pain points and the area of customer behavior.

3 Exploratory Data Analysis

In the following we will briefly introduce the dataset provided by Jeronimo Martins from different perspectives. Each

3.1 Product Assortment

The product assortment comprises 2,878 articles in total. Hereby, the assortment follows a hierarchical structure, assigning each product to a product area, that can further be broken down into product divisions, families, categories, and sub-categories, where sub-category poses the most granular level. An overview over the number of unique levels per product hierarchy level is provided in Table 1.

3.2 Basket Exploration

The average shopping basket comprises 1.8 different products. Examining the product frequencies in shoppers' baskets, products from five dominating areas are most often found in the transactions: "pereciveis especializados" (contained in 69% of the transactions), "meal solutions" (40%), "bebidas" (32%), "mereceria + pet food" (22%), and "pereciveis não especializados" (12%). All other product areas occur in less than 2% of transactions (see Figure 1).

Increasing the level of granularity, a similar pattern is visible, where few mainly bought divisions dominate. The most frequent product divisions can be perceived in Figure 2. The product division "padaria/pastelaria" is contained in 42% of the transactions, "take-away" in 34%, "refrigerantes" in 17%, "frutas e vegetais" in 14%, and all others appear in 10% of total transactions or less.

When we have a look at the best-selling products in the store, we obtain a first indication of the store's customer preferences: The most bought products, in descending order, are "merenda

mista", "massa personalizada", "pão de queijo", "sumo laranja", and "cappuccino" (see Figure 3). In general, students seem to be seeking either fully complete lunch solutions or, next to coffee and orange juice, coffee break snacks.

3.3 Sales patterns

Exploring the number of purchases per day from September to November 2021 (91 observations) reveals a strong weekly seasonality with a period of seven days (see Figure 4). In the first weeks, we can further perceive a growing trend in September until the start of the exam period in mid-October. Even though the exams were over, the number of purchases did not pick up in the week after. Furthermore, an extremely low number of purchases is recorded on holidays (e.g., 5th October and 1st November), as well as at the day of the career fair on the 22nd of September. This suggests that customer visits and consequently purchases are strongly dependent to whether classes take place or not.

Excluding Sundays and holidays from the further analysis leaves us with 76 observations for the examination into the intraweek purchase patterns. Here, the boxplot in Figure 5 shows the weekly seasonality. The number of purchases tend to increase until Wednesday and Thursday before it decreases on Fridays towards its weekly low on Saturdays. This opposes the common customer buying habit and preference of shopping on weekends (Ehrenthal, Honhon, and van Woensel 2014). However, the laboratory grocery store's customer base may not conform to the norm, as most of the students usually do not visit the campus on weekends. This underlines the special positioning of the store that does not aim to be a full-fledge supermarket.

The special positioning of the store towards offering convenient meal products can also be derived from Figure 6, which shows the intraday purchase pattern across the store's opening hours (988 day-hour observations). Interestingly, the boxplots rather suggest a wave-like purchasing behavior that seems to again correlate with the class schedules. The daily peak

demand times are concentrated on the lunchtime hours from 12 pm to 2 pm. In addition, smaller local demand maxima are detected at 9 am in the morning and in the afternoon from 4 to 5 pm. Towards dinnertime and later, the demand drops significantly, potentially pointing towards the customers purchasing less because they are either not on campus anymore, or the store does not offer the products serving the customers' needs in the evening.

Plotting the number of average hourly purchases against the average basket size in units of items and the average basket value in terms of Euros results in Figure 7. One can observe relatively stable basket sizes (3-4 items) throughout the day, while increase in the two hours before store closure in the evening. (4-5 items) While the basket value also increases towards the evening, suggesting that during evening purchases, people might also buy products not only for immediate consumption, but also to take them back home, the basket value is also increased during lunch time, while basket size rather slightly decreases. This points towards less products bought, but of higher value, during lunch time.

3.4 Customer Data

As we could expect, most customers of the store (84%) are of Portuguese nationality, taking the users' phone country code as a proxy. 6% of the customers registered with a German country code, followed by 2% with Italian and, 1% with French (see Figure 8). All other country codes form less than 1% of the customer data base. When exploring the dates when users initially registered in the app, the data indicates interesting patterns as depicted by Figure 9. Many customers did their registration in October of 2019, when the Bachelor students' semester began, which was the last pre-Covid intake. Only few registrations then happened during the semester. After Covid came up, the registrations reduced to almost zero in spring 2019. A next big spike was at the beginning of the fall semester of 2020, when all activities returned to be back on campus. After another lock-down, students only came gradually back to campus and

thus, registrations happened in a similar fashion in spring 2021. The last big peak was at the beginning of the fall semester 2021.

3.5 IoT Oven Data

A thorough analysis of the data generated by three IoT ovens in the store expose clear issues in the preparation process that characterize faulty preparations. This study will focus on the five food types that have the most observations in the dataset. These are frango, misto, pao de queijo, pao and burger congelado. The amounts of observations and labels can be found in Figure XX. Around 19% of the observations taken into consideration are labeled with being a good or faulty preparation. The remaining observations are unlabeled. Each observations resembles a batch of in-store prepared products and consists of data points describing the preparation process of the batch. For this study, these data points were translated into 8 features that describe an observation from different perspectives, namely: total_seconds, max_temp, door_open_cnt, door_open_seconds, kWh_consumed, final_temp, avg_food_temp, total_peeks. Features were created to have readable input for machine learning models. The data collected by the ovens and translated into features shows that a large fraction of the observations deviate from the optimal preparation process, which can have several implications for the store. For example, as shown in the Figure 10, the available labeled observations for frango show that false preparations have a much higher energy consumption on average than good ones. Simultaneously, the energy consumption levels measured are much more volatile, indicating that some baking processes labeled as false do not comply with the expected standard and are inconsistent. The other features like total duration and the total amount of seconds an oven door is open, also show large discrepancies between good and false preparations for frango as seen in Figure 10. Differences in distributions of feature values like these are present among all food types. Faulty prepared products pose a food safety risk and lower the quality perceived by customers. Likewise, incorrect preparations increase food waste and often show excessive

amounts of energy consumption, leading to increased costs. Most issues in the baking process that were revealed during the analysis seem to be caused by human error and can be avoided by adjusting the preparation steps performed by individuals operating the oven. Common characteristics for such errors are leaving a batch in the oven for too long, having the door open during preheating or leaving the oven on after finishing. By identifying the underlying reasons for false preparations through machine learning, these errors can be flagged and resolved in an automated fashion.

3.6 Visits Patterns and Durations

The original event log – which had 469,893 records – denotes all events generated either by the users' interactions with the store via the Pingo Doce & Go Nova app or by the system itself. Case in point: a customer introducing a new payment method on the app, entering the store, and scanning a product, the opening of the Go 24/7 cabinet door, the coffee machine storing in queue demanded products, every little operation is recorded on the event log. Naturally, not all records are worth studying – for example, knowing that the brews queue on the coffee machine is not relevant for analyzing the shoppers' paths. As so, after cleaning the data and considering that the maximum duration for a visit is 30 minutes and that no entries are recorded on Sundays and national holidays, one was left with a total of 271,742 records in the month of October, - the total period of this data – corresponding to 55,415 visits and averaging 2,217 visits per day and an approximately 4.39 events by visit. The first analysis performed – figure 11 – took into consideration the thought that the consumers' desires when going to the PD&Go Nova store are likely to vary with the time of the day. By looking at the top products added to shoppers' baskets, one can easily detect differences in the customers' missions. For example, a client entering the store on a weekday early morning typically goes for coffee machine products, pastries, and bakery products, whilst on lunchtime is looking for pasta or a pizza, and on a Saturday afternoon, he or she is longing for a beer.

When examining the distribution of the visits across the different days of the week – figure 12 –, and throughout the time of the day, one can observe that during the weekdays, the customer presence within the store tends to follow the same behavior, which in turn differentiates itself from the visits on the weekends. Additionally, the peaks occur at break times between classes, which was expected. Regarding the duration of the visits, the data indicates that the common customer spends an average of 4 minutes and 39 seconds within the store and that more than 50% of the visits last between 1 to 4 minutes. Moreover, the average time inside the Pingo Doce & Go Nova varies throughout the time of the day and whether the visit occurs on weekdays or on the weekend, as Table 2 suggests, and therefore it is strongly believed that the shoppers have different behaviors according to the different times of the day and different days of the week.

Another aspect worth mentioning is the number of visits where customers enter and leave the store without recording any other event. According to the log, these correspond to a considerable 38.6% of the total visits, and represent a reality that cannot be measured in terms of customers' paths.

References

Caetano, Edgar. 2019. “Na Nova Loja Pingo Doce, Não Há Caixas, Filas Nem Dinheiro. Há Sensores, Câmaras e (Claro) Uma ‘App.’” *Observador*, October 3, 2019.

Ehrenthal, J.C.F., D. Honhon, and T. van Woensel. 2014. “Demand Seasonality in Retail Inventory Management.” *European Journal of Operational Research* 238 (2): 527–39. https://doi.org/10.1016/j.ejor.2014.03.030.

Ponte, Diego, and Stefania Bonazzi. 2021. “Physical Supermarkets and Digital Integration: Acceptance of the Cashierless Concept.” *Technology Analysis & Strategic Management*, November, 1–13. https://doi.org/10.1080/09537325.2021.1994942.

Salgueiro, Maria. 2019. “Pingo Doce & Go: Experimentámos o Novo Conceito de Supermercado Sem Filas.” *NiTfm*, October 4, 2019.

Vaja, Bankim R. 2015. “RETAIL MANAGEMENT.” *IJRAR- International Journal of Research and Analytical Reviews* 2 (1).

Vassakis, Konstantinos, Emmanuel Petrakis, and Ioannis Kopanakis. 2018. “Big Data Analytics: Applications, Prospects and Challenges.” In , 3–20. https://doi.org/10.1007/978-3-319-67925-9_1.

APPENDIX A: Tables

Table 1: Product Hierarchy

Hierarchy Level	Level Count
Area	14
Division	37
Family	108
Category	284
Sub-Category	737

Table 2: Average Shopping Duration

	Weekdays	Weekend
Morning	3 minutes and 41 seconds	3 minutes and 42 seconds
Lunch Time	5 minutes and 22 seconds	6 minutes and 4 seconds
Afternoon	3 minutes and 57 seconds	4 minutes and 26 seconds
Dinner Time	6 minutes and 16 seconds	5 minutes and 3 seconds

APPENDIX B: Figures

Figure 1: Distribution of Product Areas in Transactions

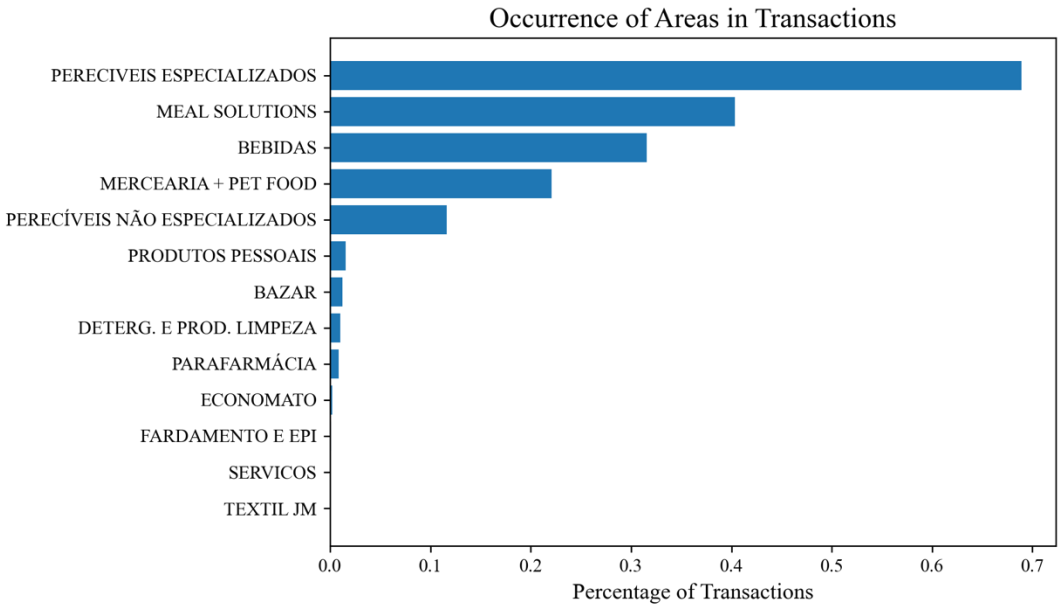


Figure 2: Distribution of Product Divisions in Transactions

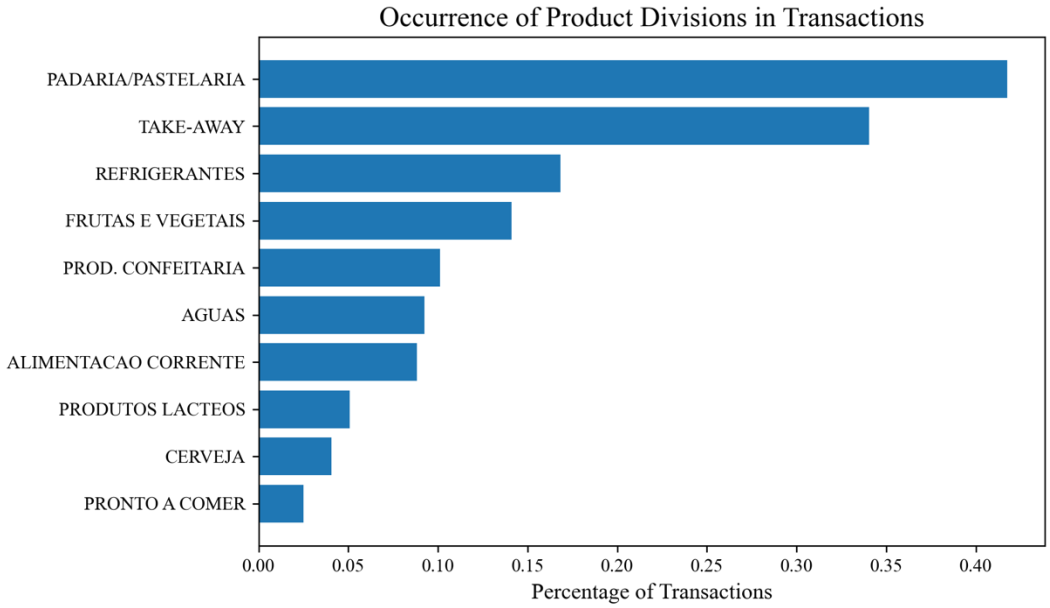


Figure 3: Top 10 Most Sold Products

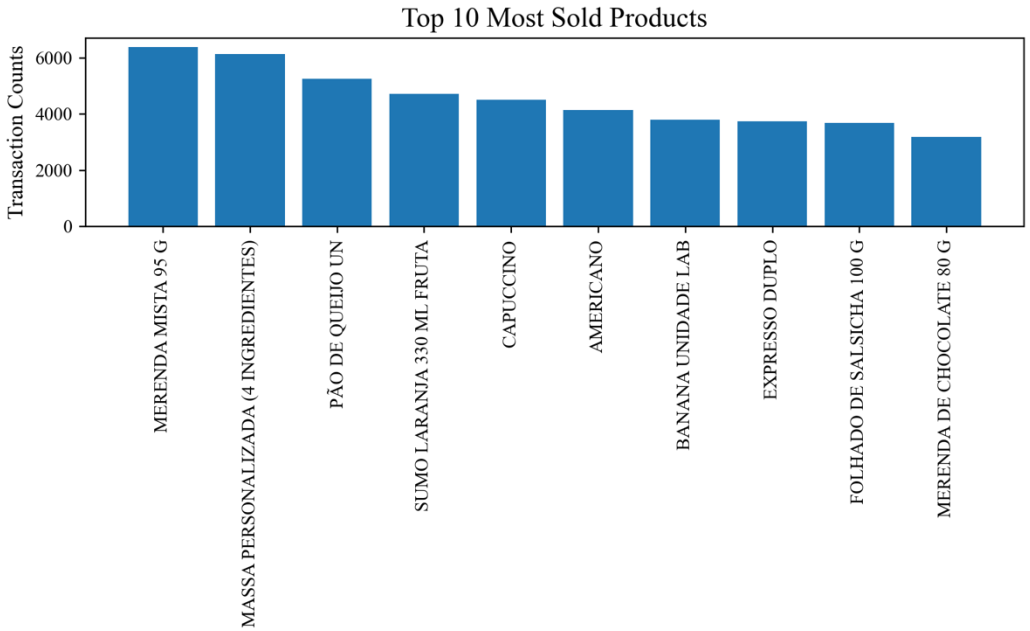


Figure 4: Purchase Patterns over the Whole Period

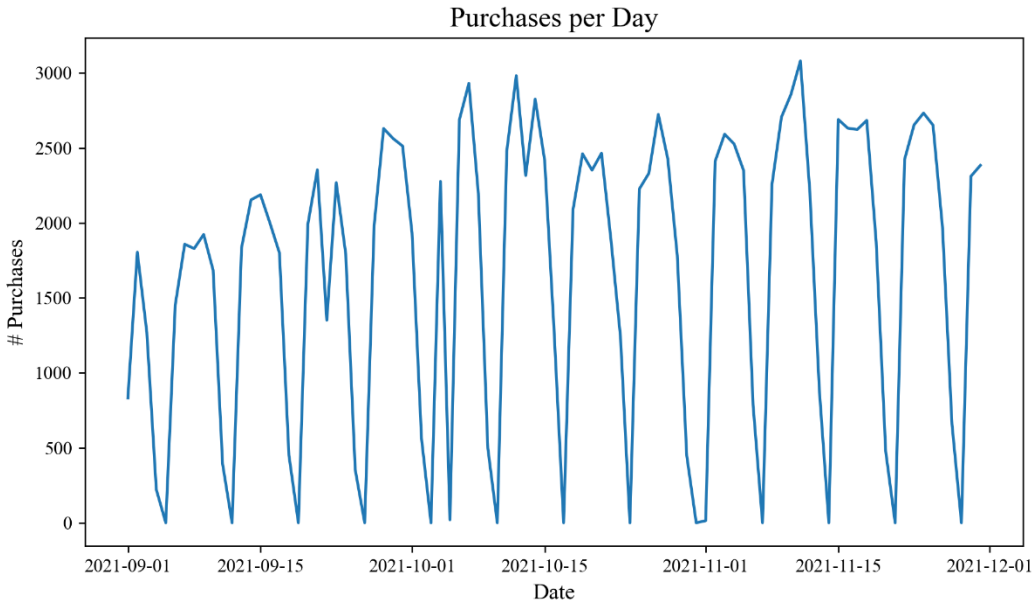


Figure 5: Intraweek Purchase Patterns

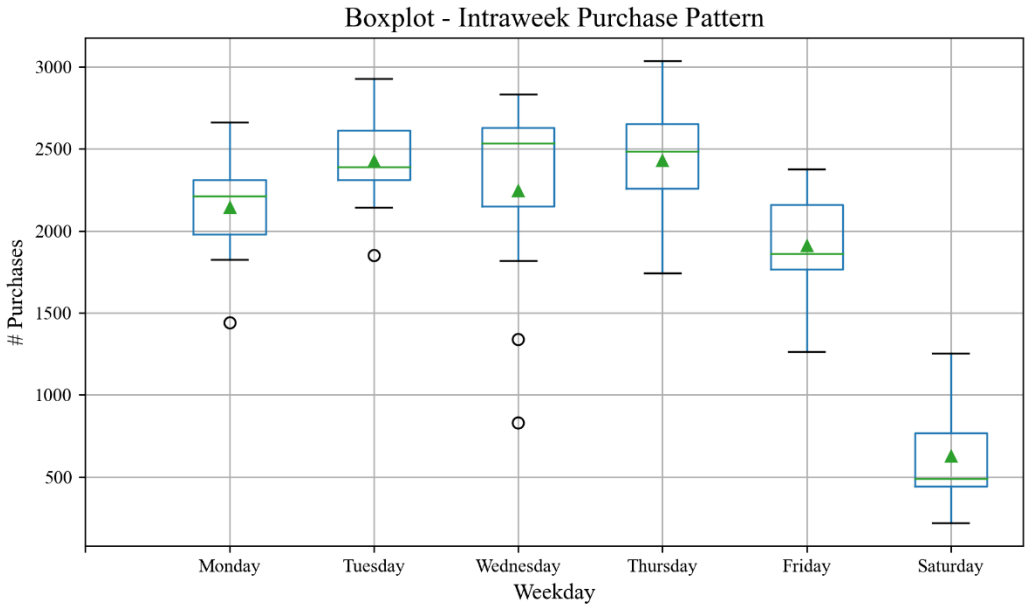


Figure 6: Intraday Purchase Patterns

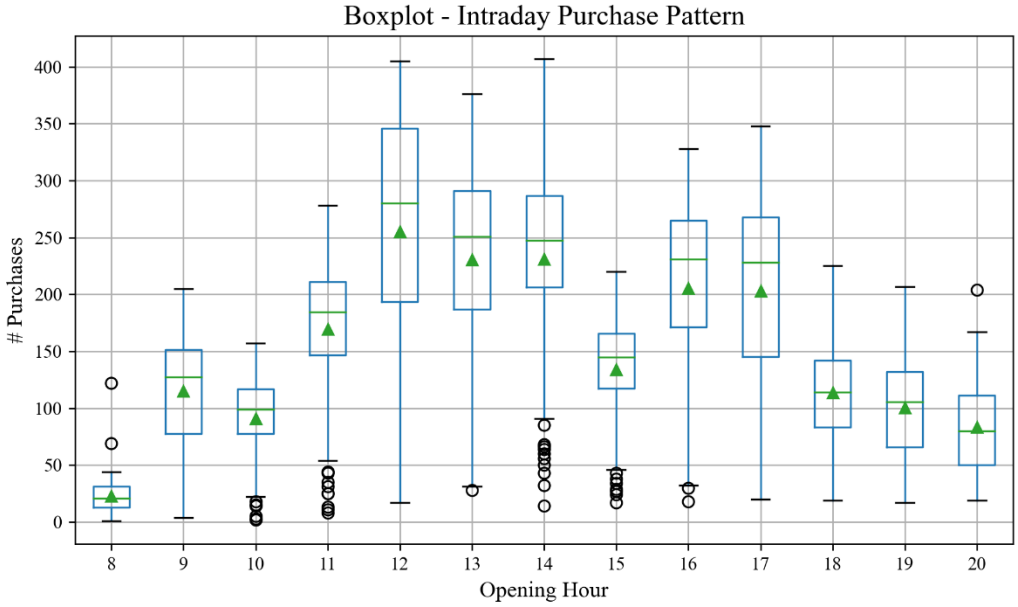


Figure 7: Average Basket Size, Basket Value and Purchases

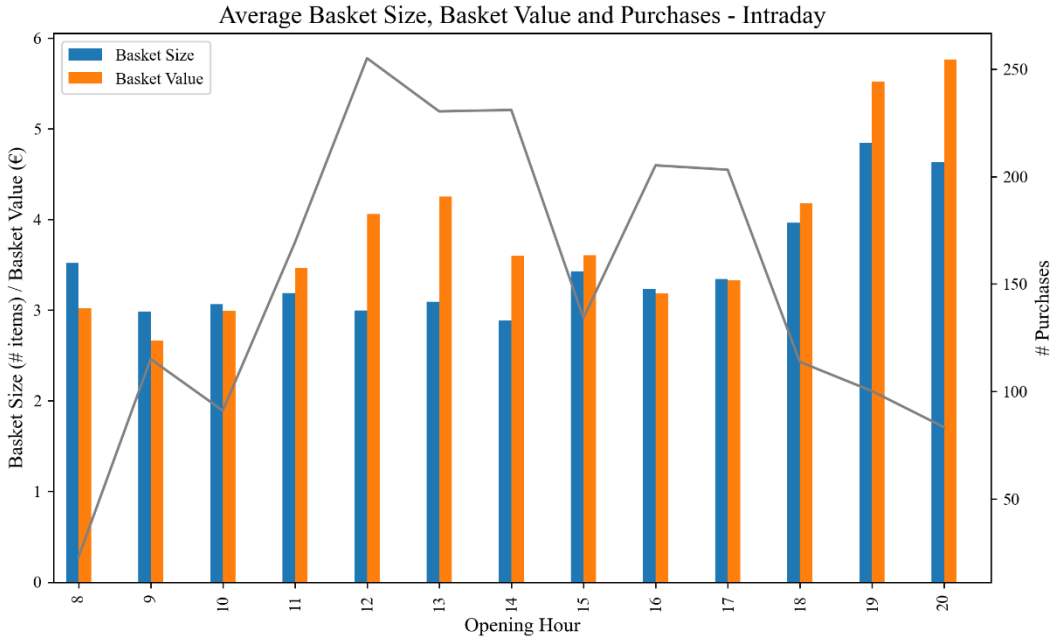


Figure 8: Distribution of Country Phone Codes in Customer Data Base

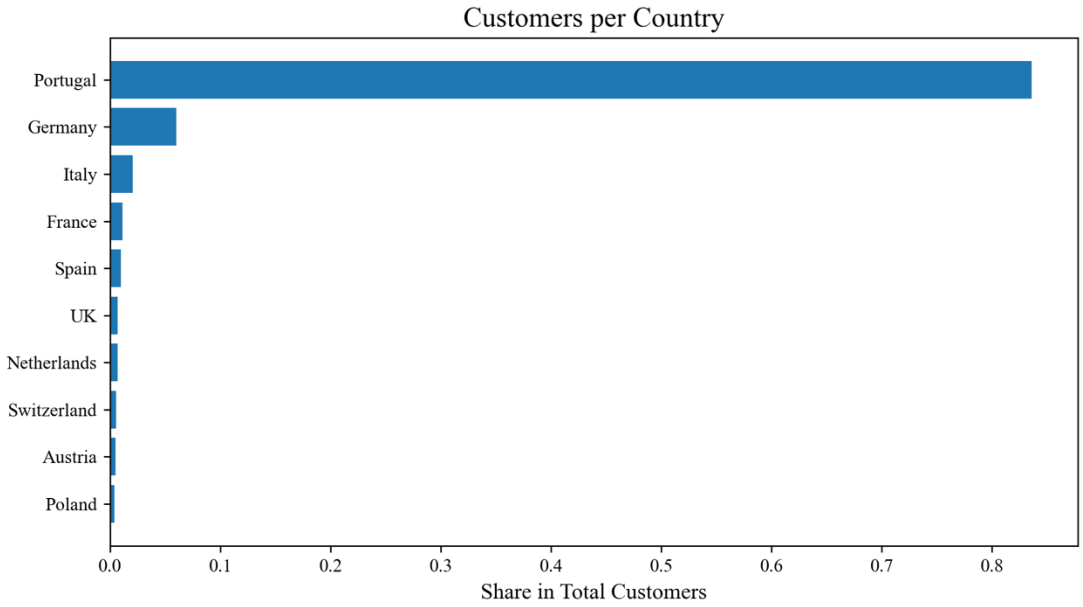


Figure 9: Daily Customer Registrations

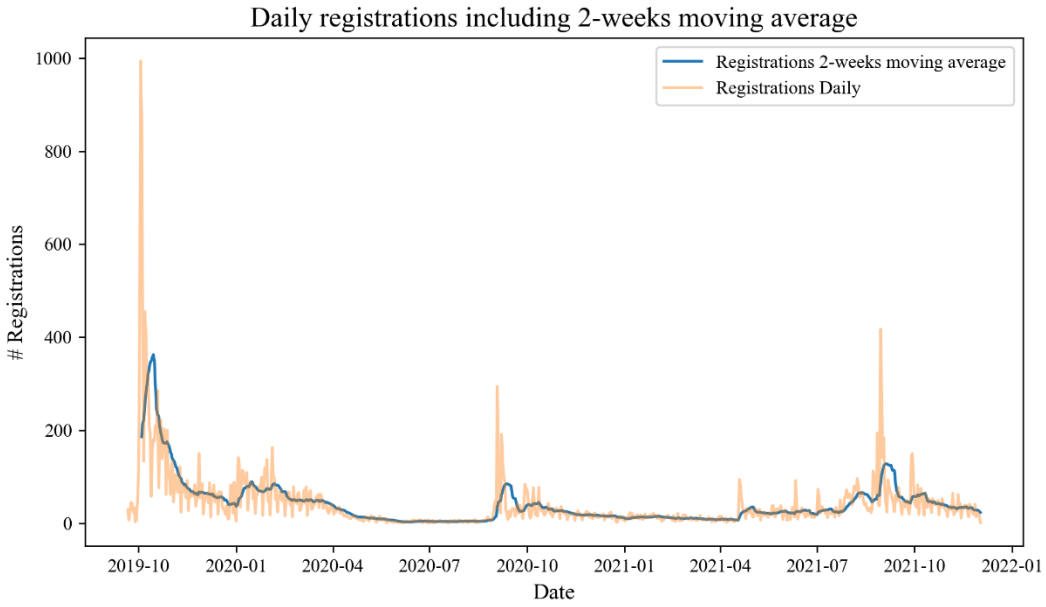


Figure 10: Energy consumed for good and bad frango preparations

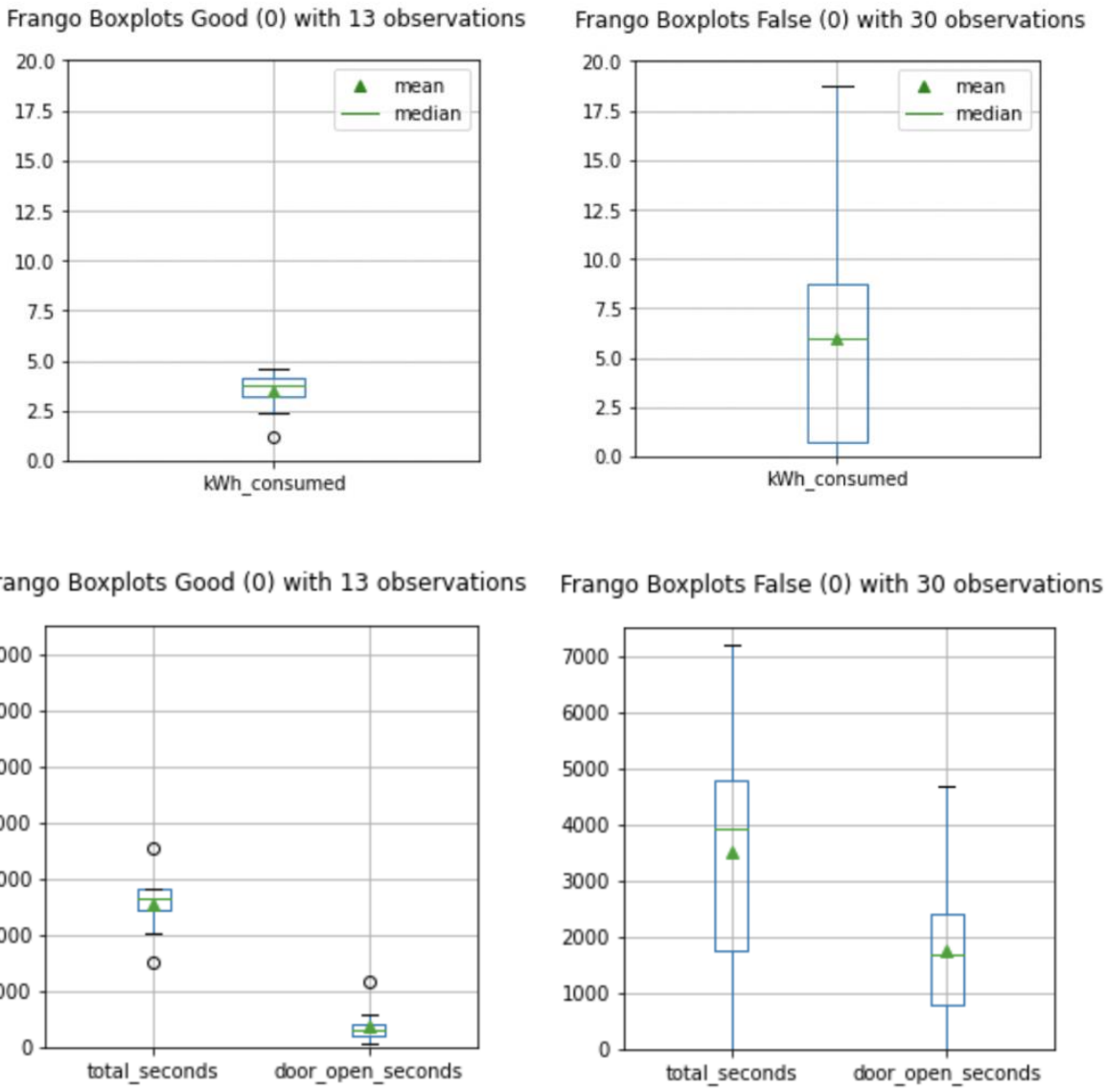


Figure 11: Number of visits by Hour and Minute, during Weekdays and Weekends

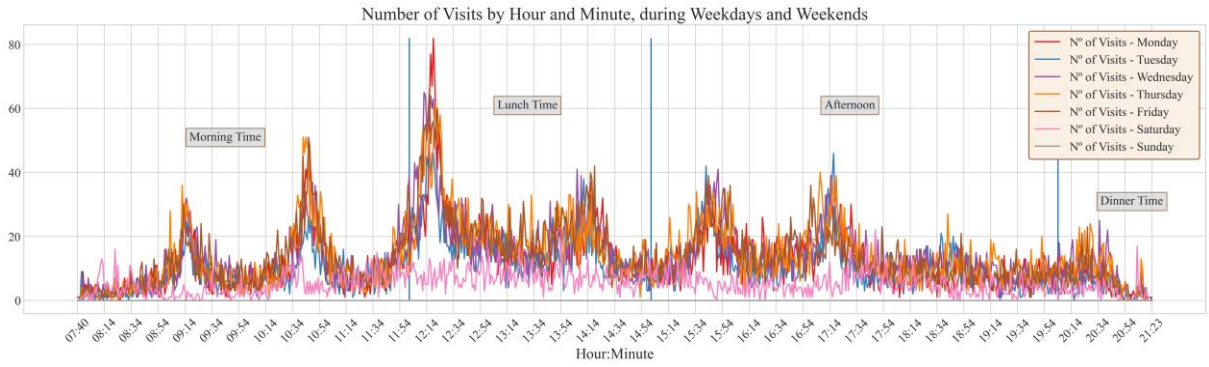


Figure 12: Frequency Distribution of Visits' Duration