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

FIELD LAB JERÓNIMO MARTINS – INTELLIGENT MANAGEMENT INFORMATION SYSTEMS FOR NOVA & GO STORE

Shopper pathway discovery in cashierless stores: a Process Mining approach

Maria Pais dos Reis Delgado Chaves

Work project carried out under the supervision of:

Professor Qiwei Han
Mr. Rui Tomás

25/01/2023

Acknowledgments: To Professor Qiwei Han, for sharing his experience and knowledge and for providing such valuable guidance, encouragement, and motivation. To the Jerónimo Martins group, in particular Mr. Rui Tomás, for the opportunity to work on a real-world problem and for all the support provided throughout this journey.

Abstract: In recent years, the retail industry has undergone a drastic transformation due to the use of the latest technologies. An example of a revolutionary supermarket is Nova & Go, a store that was established by Jerónimo Martins. The present thesis addresses an important challenge relating to improving the daily operations of this Lab store. The analysis of event log data provided valuable insights into the customer journey within the store.

Keywords: Customer Journey, Data Mining, Internet of Things, Management Information Systems, Process Mining, Retail 4.0

This work used infrastructure and resources funded by Fundação para a Ciência e a Tecnologia (UID/ECO/00124/2013, UID/ECO/00124/2019 and Social Sciences DataLab, Project 22209), POR Lisboa (LISBOA-01-0145-FEDER-007722 and Social Sciences DataLab, Project 22209) and POR Norte (Social Sciences DataLab, Project 22209).

INDEX

1. Introduction

In today's world, data is viewed, by most organizations, as a strategic asset *(Delloite, 2021).* In the words of Clive Humby, a British mathematician, it can even be described as “the new oil of the 21st century”, due to its tremendous potential and value when realized through business applications. However, just like oil, data needs to be refined. Raw data is worthless, if valuable insights cannot be extracted from it *(McKinsey, 2018).* Rather, it’s through the acquisition of information that companies are able to respond proactively and intentionally to market conditions. Accordingly, in 2011, Peter Sondergaard, a Gartner’s board member, corrected Humby’s famous quote by stating: “Information is the oil […] and analytics is the combustion engine. Today, analytics is the new oil.”

Among a wide range of industries, Big Data has opened up new opportunities, and Retail is no exception. In fact, the concept of Retail 4.0 was introduced in 2010, with the intention of describing the sector transformation based on Industry 4.0 technologies, such as Internet of Things, Cloud Computing or Artificial Intelligence *(Har et al, 2022).* In fact, technology has enabled the retail sector to create a brand-new shopping experience *(Gazzola et al., 2022).* From self-service kiosks to e-commerce platforms or AR-based tools aimed to enhance customer experience, this new era is characterized by innovation.

An example of an innovative project aligned with this new trend is a cashier-less store created by Jerónimo Martins (JM), the retail leader in Portugal. In October 2019, JM launched Nova & Go store, at the Nova School of Business and Economics campus. Taking advantage of a group of customers predominantly made up of young students, unlike what happens in the regular stores, this project aims to test new technologies and new products. Being primarily visited by Gen Z customers, it is an ideal opportunity to experiment with innovative technologies and new business models. Apart from traditional groceries, the store offers various freshly prepared foods and other novel products, all offered with the students in mind.

Customers' experience in this Lab store starts with the download of the Pingo Doce & Go Nova app, available for both iOS and Android. After registering, a QR code should be displayed to enter the store. Once inside, purchases can be recorded in two ways: either through NFC technology or by pointing the cell phone's camera at the product code. Later, to be able to pay, customers can associate their bank card to the app or, alternatively, use the payment towers. Yet, even in the latter case, only credit or debit cards are accepted, not notes or coins.

During this process, data is generated at various points. So, in order to handle the large volume of data produced by the store, Management Information Systems (MIS) must be incorporated into its daily activities. Essentially, MIS collects and processes data from all sources within an organization to assist managers in their decision-making processes *(Berisha-Shaqiri, 2014).* As a result, JM has access to a detailed database that helps it explore the habits of its customers and analyze critical aspects of its processes.

Based on the exceptional opportunities provided by the Lab store, JM has identified a topic to be developed throughout this thesis. Based on that, a Process Mining tool will be used to monitor the path taken by each customer within the store.

Lastly, it should be noted that Design Science Research (DSR) methodology was chosen to structure the present thesis, since it has proven to offer outstanding benefits when it comes to solving real-life problems within organizations *(Hevner et al., 2004)*. In the appendix, it is possible to find the different steps of this approach *(Fig 1. in the appendix).*

2. Problem Statement and Motivation

Currently, the retail industry is undergoing a dramatic transformation. In the past, when looking for a particular product, customers were restricted by the options available in brick-and-mortar stores. In recent years, however, with the advancement of technology and globalization, consumers became able to request any product or service at any time, regardless of their location. In fact, due to the broad range of substitute products and services available, along with effortless access to information, clients occupy a position of power *(Perkins & Fenech, 2014).*

They have therefore become increasingly sophisticated, demanding, and informed in their decision-making processes. As a result, their loyalty to brands has also become increasingly transient *(Reinartz & Kumar, 2000).* Simultaneously, we are witnessing a shift in what people value most, as the primarily product-oriented model has evolved to a more service-oriented one *(Kindström & Kowalkowski, 2016)*. On the business side, in order to remain competitive over the years, companies must continually adjust their strategies. For instance, as a result of this new paradigm, a customer-centric approach should be adopted. To capture customers' attention, companies should offer a memorable and personalized experience.

To this end, it is essential to identify areas for improvement in their processes *(Alzoubi & Khafajy, 2015)* and key touchpoints during the customer's journey. Customer journey describes the sequential steps that customers go through when interacting with a company *(Følstad & Kvale, 2018).* Through it, companies are able to better understand the viewpoint of their customers . As one of Portugal's largest retailers, JM is also committed to improving its processes. By using Nova & Go store as a case study, it is important to identify the store's main bottlenecks and system failures that may negatively impact the customer's path. By providing a better experience, JM will build trust and loyalty, which may result in an increase in the number of visits to the store and ultimately, an increase in revenue.

3. Objectives of a Solution

As mentioned in the previous section, JM's primary objective with this thesis is to detect the most significant flaws in the customer's journey inside the store. For instance, after identifying process bottlenecks, possible solutions and suggestions can be made in order to reallocate resources that allow a constant improvement in process performance

With the rapid increase of data generated in recent years, JM as any other company, have been faced with one of the most challenging tasks: gaining useful insights from those resources. In fact, Nova & Go store already collects its sales records and in-store events, which are stored in a database. Yet, there is still a need for appropriate tools and techniques that can transform these data into useful information *(Chen et al., 2014).* An example of a powerful tool that is able to strengthen operations within organizations based on data is Process Mining (PM) *(Aalst, 2013).* Process Mining involves analyzing events that are generated during the execution of a business process *(Aalst, 2011).* Later on, for instance, through the analysis of the events data, inefficiencies, productivity losses, and deviations may be detected *(Aalst & Dustdar, 2012).*

Currently, very few studies have been conducted regarding PM's application to retail, highlighting the importance of this analysis. Besides, a complementary analysis of a Data Mining (DM) technique, an Apriori Algorithm, was also performed. In essence, this procedure suggests and evaluates possible relationships between products. Throughout this thesis, introspective knowledge about the store customers' journeys will be extracted using both process and data mining methods. Such insights will represent a strategic and organizational advantage for JM. In other words, the purpose of the present thesis is to deepen the following research question:

Research Question: *How can customer journey maps be discovered, explored, and improved from event logs?*

4. Literature Review

In this section, we will describe the main concepts, definitions, methodologies, and results of previous research conducted, from the perspective of the desired analysis.

4.1. Customer Journey Mapping

In today's market, companies must seek to understand the purchasing process not only from their own point of view but also from the perspective of their customers. Identifying key touchpoints along the journey, as well as the factors that influence consumers in making or refraining from making a purchase, is therefore essential *(Lemon & Verhoef, 2016).* Thus, a customer journey serves as a starting point for organizations to gain a deeper understanding of their customers and business processes, by mapping all interactions and events related to service delivery *(Patrício et al., 2011).* This mapping process involves tracking and describing customer responses and experiences *(Marquez et al., 2015).* Hence, the purpose of Customer Journey Map (CJM) is to gain insights into customer behavior and, ultimately, to improve customer service.

4.2. Technology Evolution

In recent years, technological advancements, specifically on Internet of Things (IoT), have allowed retailers to adopt new business models and innovative approaches *(Hopping, 2000; Bok, 2016).* In fact, IoT refers to the latest technological revolution, following the computer and the Internet. The concept was introduced by *Kevin Ashton,* in 1998, during an official presentation to the board of Procter & Gamble, where he was a brand manager. During that time, the British technology pioneer sought a solution that would ease the monitoring of the high volume of products in the company's supply chain. Simply, the main idea behind IoT was a world where technology would be able to define a physical network of connected objects, information, and people.

4.3. Process Mining and Data Mining

Since the volume of data produced has increased over the past few years, it became necessary to develop tools that allow an automatic interpretation and exploitation of this abundant and accessible amount of data *(Fayyad, 1996).* As a matter of fact, these resources, coupled with the urgency of converting it into valuable knowledge, has led to the creation of Data Mining tools *(Tiwari et al., 2008).* The goal of DM is to uncover hidden, unpredictable, and useful patterns from large databases, by utilizing a variety of methods and algorithms *(Han et al., 2012).* Moreover, DM is considered an important step of Knowledge Discovery in Databases (KDD), which can be described as the process of creating insightful information from raw data. There are five steps involved in the KDD process, including data selection, preprocessing, transformation, data mining, and, finally, interpretation *(Fayyad et al., 1996).* KDD was defined in 1996 by *Fayyad, Piatetsky-Shapiro, and Smyth.*

As time passed, the need to fill the gap between traditional process models and Data Mining methods arose, resulting in the creation of Process Mining *(Aalst, 2016).* PM was coined by Wil van der Aalst and, essentially, relies on data in information systems, to discover process models, monitor model compliance, and improve and automate processes, providing at the same time the necessary support *(Aalst & Dustdar, 2012).* In this regard, the main difference between Data Mining and Process Mining is that while the former is concerned with the orderly organization of raw data and its patterns, PM focus on ensuring the orderly arrangement of activities within business processes. In fact, PM techniques use information systems and event-log records. The prospect value of PM for the customer was first highlighted by Bernard and Andritsos *(Bernard & Andritsos, 2017).* Through PM, the researchers demonstrated how the total set of events that customers experience, when interacting with an organization, could be analyzed.

Furthermore, according to the Process Mining Manifesto *(Aalst et al., 2011)* created by the IEEE Task Force on PM, it has three main types: Discovery, Compliance, and Enhancement. The most common PM technique corresponds to Process Discovery, which seeks to build a model that represents the actual processes of an organization. On the other hand, a Compliance Check associates events with the activities visualized in the model. This allows a comparison and analysis of the differences between a defined model and a registered behavior. Finally, Process Enhancement refers to correcting, strengthening, or improving processes by analyzing the process event log. Nowadays, there are many tools available on the market that can be used for PM. A comparative study by *Çelik & Akçetin (2018)* identified the following top five tools: Apromore, ProM, Celonis, myInvenio, and Disco.

4.4.Applications of Data Mining and Process Mining to the Retail Sector

Market Basket Analysis is considered one of the oldest areas of DM *(Raeder & Chawla, 2010).* Simply, it uses Association Rules to assess whether the presence of a set of items in the customers' basket leads to the subsequent addition of a different item to the same basket *(Agrawal & Srikant, 1994).* So, Association Rules consist of two sets of items: an antecedent (X) and a consequent (Y), which are described as Antecedent (X) → Consequent (Y). Association rules were introduced by *Agrawal & Srikant (1994)*, aiming to find associative patterns in supermarket basket data. Specially, Apriori is an algorithm used in Association Rule Learning for transactional data, and its creation has contributed significantly to the advancement of research in Data Mining. In fact, the Apriori Algorithm was ranked in the top 10 of Data Mining algorithms with scientific influence by the IEEE International Conference on Data Mining (ICDM) in 2006. According to the algorithm, if an item set is frequent, then all its subsets will be frequent as well. It is due to the following property of support that this principle is valid. This Apriori Principle is referred to as the anti-monotone property of the support. It means that the support for a set of items never exceeds the support for its subsets:

Equation 1- Anti-monotone Property of the support.

$$\forall X, Y:(X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Regarding PM application on retail, it is known that companies have long been interested in analyzing customer behavior, especially the actions they take within the stores. The collection of movement data and the ability to track customers have, however, limited this type of analysis until recently. Recent features have now made it possible to track shoppers' activities within the store. As far as the literature reveals, PM techniques have not been widely applied or explored when it comes to analyzing customer behavior in retail. Yet, among the few cases studied*, llhoe Hwang and Young Jae Jang's* work analyzed each customer path in a physical store, using a smartphone equipped with WiFi-based positioning technology. By using PM, it was possible for the authors to verify a change in customer behavior before and after a change in the display configuration of some mannequins.

5. Design And Development

5.1 Data Understanding

Process mining is based on the analysis of event logs. Essentially, these are records of activities that took place during the execution of a particular process for a particular user at a specific time. Thus, it consists of three categorical elements: the Case ID, Timestamp, and Activity. *(Aalst, 2016).* With this, each event is associated with a specific Case ID that identifies a given process instance. Moreover, each process can contain multiple Cases, such as numerous Visit Id, as in our case. In addition, each event reflects an Activity, which represents a particular step in the process. In some Cases, certain activities may occur more than once, while in others, some may not take place at all. On the other hand, Timestamps are used to indicate when an event occurred. In the next section, we will look at data curation step. A total of 999, 367 rows and 14 columns were originally contained in the event dataset.

The data collected included information from the 1st of September of 2022 to the 29th day of the month. The columns provided originally can be found in Tables 1 and 2 in the appendix.

5.2 Data Curation & Feature Engineering

An event log with missing, incorrect, or duplicated values could result in an unstructured model, that is difficult to interpret and does not reflect the actual behavior of the business process. Thus, after understanding the extracted data, we had to treat, filter, and process it in order to load it into the PM software later on. To conduct this analysis, we used the datasets received from Events, Movements and Planograms. Throughout this preprocessing step, we have extensively treated existing variables and added new ones. Nevertheless, this section will only focus on the most critical changes in three areas: customers, events, and visits.

We began our data curation process by analyzing customer information, focusing particularly on the column Client. According to the initial data, this column contained approximately 17% of missing values, corresponding to 168,989 records. Consequently, due to the inability to distinguish and analyze unidentified customers, we decided to discard these rows, as they were irrelevant to our study. Moreover, to obtain more information regarding JM's customers, we also created a new variable, Customer Country, which included an identifier for the user's origin country. This variable was created based on the column Customer Country Code, which did not contain null values and had a total of 48 unique values.

To manage events information and simplify our analysis, we created a new variable, Event Category *(Table 5 in the appendix)*, in which events belonging to the same category were grouped together. In this phase, we also linked the Events and Movements' dataset, aiming to gather additional information about the products that customers purchased, including their area, division, family, category, and subcategory. To achieve this, the two datasets were merged, through the SAP codes of the products. Nevertheless, after carefully linking these datasets, we realized that SAP code could not be extracted from the Events dataset when products were added using the tap method. Due to this, we had to create a dictionary of the existing products and correspondingly SAP codes, which were added based on the scan method, to fill in the empty cases. Even though we were able to resolve most of our issues using this technique, some of them remained. For instance, there were some products that were just added based on the tap method. In addition, there were also some products that had more than one SAP code, while others required corrections to the identifier. Therefore, we have attempted to distinguish and solve these issues, always verifying and updating them through the Movements dataset. We were unable to extract event information related to different types of "coffee machine products". For this reason, to properly analyze the data, we have always had to work with the group of observations that comprise all "coffee machine products". It is also worth mentioning that there was a change in the way pastas and custom salads are registered on the last two days of the month. Since then, the base is now scanned first, followed by another scan for each additional ingredient. Due to this, and upon JM's suggestion, we decided to group all pastas and all salads, regardless of the number of ingredients into two different groups: "massa personalizada" and "salada personalizada".

Furthermore, we created a new variable, Store Zone. By using the Planogram dataset, we created a dictionary that associated each SAP code with a particular Store Zone. Consequently, we added this new column to the Events dataset. However, while creating the dictionary, we discovered that some products were missing from the Planogram data. Specifically, 505 products, corresponding to approximately 30% of September's purchases, were missing. For these cases, we were unable to identify the Store Zone in which the products were located. Due to the high percentage of products without this information, we could not conduct further analysis about this topic in the exploratory analysis. Additionally, the following variables were created: Weekday, Entry Datetime, Exit Datetime, Time of the Day, among others. An explanation of the variables, as well as possible values, is provided in *Table 4 in the appendix.*

After some initial processing, we were still left with 2,301 duplicated rows, meaning the same person did the same thing at the exact same moment. Considering this to be unrealistic, we decided to drop them. Hereafter, we started forming a logic for analyzing distinct customer journeys. To begin with, we excluded events that were unrelated to store entrances and exits, products being added to shopping baskets, as well as payments, since we considered these to be the only meaningful events. In addition, we decided to discard denied entries and exits, as they were irrelevant to the investigation of customer journeys. To facilitate an in-depth analysis later, in the PM software, we created additional columns to analyze and obtain information about the different items in the customers baskets. Through the addition of these new columns, it was possible to analyze the different customer journeys, in accordance with the desired level of detail *(Table 6 in the appendix).*

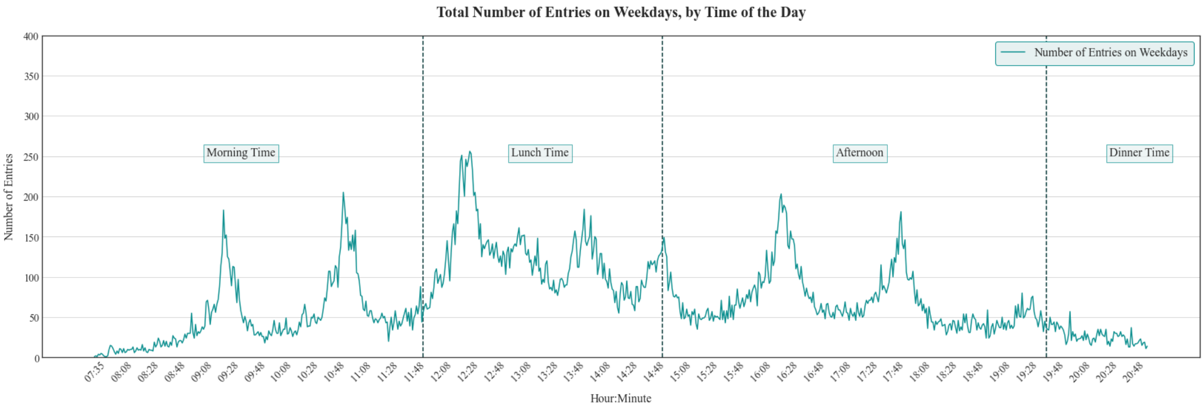
Finally, and with regard to the visit information, it became necessary to assign a Visit Id to each customer visit to the store. For this purpose, we created a function that changed the visit identifier whenever a new "Check-in" or a new customer was detected. Besides that, we also defined that a correct visit should always begin with a check-in and always end with a check-out, without any observations outside of these parameters. Upon assigning an ID to each visit, we realized that there was an imbalance between the number of check-ins and check-outs, since some entries were not associated with an exit (and vice versa). To solve this issue, we deleted these visits, as they did not comprise a complete customer journey and could harm our analysis. Additionally, there were some duplicated events. For instance, nine customers checked out twice, at the same time, however, through different gates. In this case, the problem was solved by eliminating one of the check-outs. Moreover, some visits had receipts that were outside the check-in/check-out range, resulting in the wrong visit ID being assigned. Considering that our goal was not to assess customers' payment compliance, but rather their in-store experience, we decided to discard all event logs relating to receipts. Finally, it should be noted that some visits were unusually long (e.g., twelve hours). Therefore, we decided to treat these extremely large values as outliers and discarded them accordingly. Due to the average duration of 5.56 minutes, 30 minutes was considered as a threshold to properly deal with outliers. As a final step, we also decided to drop the cases that were not within the normal operating hours and working days of Nova & Go store. By the end of data curation, we had 334,174 events with no duplicated rows. These activities characterized 57,881 visits made by a total of 6,295 unique customers *(Table 3 and 4 in the appendix).*

5.3 Exploratory Data Analysis

Having completed data curation, we proceeded by performing an exploratory data analysis (EDA), which identified potential insights derived from the data, beyond the formal modeling task. The majority of the customers were primarily from Portugal, followed by Germany, Italy, and France *(Fig 2. in the appendix).* As reported by Life at Nova SBE on November 7th, these results coincide with the top nationalities of the faculty's students *(Table 7 in the appendix).* During this exploratory analysis, one of our main objectives was to determine how entries flowed during the days of the month, weekdays, and even throughout the day.

Considering the average number of entries per day *(Fig 3 in the appendix),* we identified that Tuesday had the highest number, with an average of 3,717 entries. Besides that, Friday and Saturday were the only days of the week below the general entry average, which was 2,226 entries. We also realized that the number of entries tends to decrease throughout the week, after Tuesday. When analyzing the entries throughout the day, on weekdays, we realized that the greatest peak was observed during lunch time *(Fig 5 in the appendix).* Moreover, during the other periods of the day (morning and afternoon), the peaks of entries occurred during class breaks. Over the weekend, entries were much more evenly distributed, with no clearly noticeable peaks *(Fig 6. in the appendix).*

Fig 5- Number of Total Entries on Weekdays, by Time of the Day



As a result of our analysis of the flow of entries over the month, we were also able to conclude that all Saturdays were relative minimums *(Fig 7 in the appendix).* The same may be due to the fact that classes are not held on Saturdays. Consequently, it was already expected that there would be a significantly lower number of entries on those days. In addition, it can also be seen that the highest number of visits occurred on September 27, with 3,988 visits.

During the exploratory analysis, we also pretended to monitor the duration of each visit. As shown in *Fig 8* and *Table 9 in the appendix,* visits on average lasted five minutes. Moreover, based on our analysis of the Time of the Day variable *(Table 8* and *9 in the appendix)*, we were able to determine that the longest visits occurred during the lunch time (5.54 minutes), and similarly, during dinner time/evening (5.52 minutes). This could be associated with freshly prepared meals preparation. Additionally, following the suggestion of Jerónimo Martins, we analyzed the waiting time in the queue for portions of pasta and salad (*Fig 10* and *11 in the appendix*). As stated by the employee responsible for this store’s area, customers typically scan the QR code of pasta and salad when they begin to order. Therefore, we calculate the queue duration as the difference between the customer's entry time and the moment at which the customer scans the QR code for the pasta or salad.

Considering that the queue for pasta and salad is the same, we assumed that the two meals should be grouped together when calculating the average queue duration. Based on our analysis, the average duration was 8.07 minutes, with the peak occurring between 14:00-14:30.

In addition, it was critical to explore the main itemsets sold in the Nova & Go store. According to *Fig 16 in the appendix,* "água" and "garrafa de água ECO" were often associated. Similarly, we noted that one of the most popular itemsets was "croissant brioche misto," "coffee machine products", and "sumo laranja natural", which is an already available menu. In terms of meals solutions, the most popular itemset was pizza and oven service, which is not a product but rather a service. Additionally, it has been observed that there was a high portion of people purchasing “frango assado" and "arroz branco", along with drinks (“coca cola”, “água”, etc.). Furthermore, we found that “coffee machine products” were often accompanied by "merenda mista", "merenda chocolate", and "banana".

5.4 Model

Throughout this section, we will explore a Process Mining tool. We will also develop a Data Mining one, as a complement. We will begin by exploring several resources available within the software we have selected, which was Celonis. Consequently, we will provide an overview of the work undertaken, as well as the results and conclusions drawn from the various tools used.

5.4.1 Process Mining – Celonis

Several Process Mining tools are available in the industry today, including Apromore, ProM, Celonis, myInvenio, and Disco. For our thesis analysis, we selected Celonis, a B2B Process Mining tool that analyzes and improves existing processes *(Celonis, 2019).* Essentially, Celonis was designed to document and maintain business processes, as well as monitor and analyze its performance and compliance *(Badakhshan et al., 2020).* As one of the market's leading tools in the Process Mining industry, the software has several world-renowned clients, such as Vodafone, Siemens, Cisco, and Uber.

Celonis has an intuitive and engaging interface, allowing users to select the number of activities and connections (paths) that are relevant to their analysis. A significant feature is the ability to change the process map, depending on the intended analysis. For instance, we can analyze the customers' journeys based on the frequency of cases, the frequency of activities, and the processing time (median, mean, and trimmed mean). Furthermore, Celonis offers the possibility of integrating real-time data with relational database management systems (RDBMS) such as SAP, Salesforce, and Oracle. It should be noted that Jerónimo Martins already works with SAP, one of the biggest companies in the area of software development. Therefore, this could be considered a major advantage, as JM could then extend this analysis to its constantly changing real-time data.

The first step in using Celonis was to import the event logs that were provided by JM, and, in turn, previously extracted from SAP. Due to the maximum number of rows allowed in Celonis, we have chosen to import only the data from 13/09/2022 to 29/09/2022. To proceed with our analysis, three attributes had to be selected *(Celonis Guide 1 in the Appendix).*

Firstly, the Case Id was chosen. Throughout this analysis, the Case ID was always the Visit Id. In addition, we had to identify an "Activity" which describes the action captured by the event. Regarding this attribute and having created several columns with different levels of detail, we selected different activities according to the desired type of customer journey *(Table 6 in the appendix).* Lastly, we had to identify a Timestamp column, which indicates the date and time of the event. This attribute was reflected in the column Created On, for customer journeys. Moreover, other columns previously created from Created On, such as Time of the Day and Weekday were also added, to conduct other types of analysis in Celonis. For instance, it enabled us to examine whether behavior changed on different days of the week or throughout the day.

After identifying these attributes, it became possible to perform different types of analysis. Celonis offers a wide range of benefits at this point, including process overview, process explorer, variant explorer, compliance checks, and others.

a) Process Overview

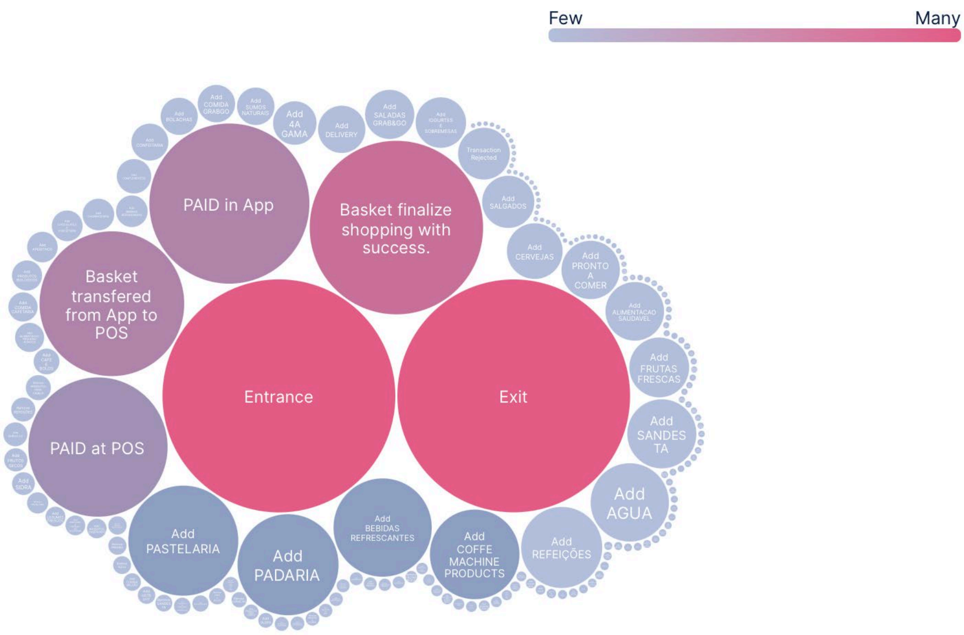
We began our analysis in Celonis with the Process Overview tool. It provided some general information about the process itself, namely the number of cases created per day, the number of events/activities within those cases, and the average throughput time per case *(Fig 17 in the appendix).* In addition, it presented data concerning the "Happy Path", which, according to Celonis definition, encompasses the sequence from the most frequent initial activity to the most frequent final one.

Using this tool, we were able to conclude that approximately 10.66% of the cases follow this so called "Happy Path", which ultimately turns out to be an "Unhappy Path". This path represents customers entering the store and leaving without purchasing anything *(Fig 18 in the appendix).* Given the multiplicity of paths that can occur within the store, this result was expected. Nevertheless, JM should continue to focus its efforts on understanding the reasons for these customers' non-purchases.

The "Activity Chart" was another relevant type of analysis in the process overview, in which each circle represented a specific activity during a visit to the Nova & Go store. According to these types of charts, the size and tone of each circle indicate the frequency of events, so the larger and darker circles represent more frequent ones, whereas the lighter and smaller circles indicate less recurrent events. In order to gain a deeper understanding of the various types of events and the distribution of each product, area, division, family, category, subcategory, and store zone, different types of "Activity Chart" were developed *(Fig 19 to 25 in the appendix).*

Taking *Fig 24* as an example, we were able to analyze the distinct weights assigned to the September events. Based on this, we concluded that there was no difference in the number of entries and exits, as the circles were of the same size. In addition, it was clear that customers preferred to pay via the app, rather than at the payment tower. Apart from this, it was also evident that the most frequently requested product families were "pastelaria", "padaria" and "bebidas refrigerantes".

Fig 24- Activity Chart, based on Product Family



b) Process Explorer

A second type of analysis was also performed at Celonis called "Process Explorer". In order to get a better understanding of the logic behind this visualization, we have created a Celonis Guide in the appendix. We will begin our analysis with *Fig 26* and *Fig 27,* where each arrow represents a link between two consecutive events in the store. Besides that, the numbers next to the arrows indicate both the frequency with which these two events were connected *(Fig 26),* or the average time between them *(Fig 27).*

At an operational level, we first identified several payment errors through the app. As a result, Jerónimo Martins should consider implementing some changes to prevent these situations, since payments through the app are the preferred method and typically cause a delay of two minutes when leaving the store. In addition, we were able to determine which product divisions were most commonly requested ("padaria/pastelaria", "take", "refrigerantes", “coffee machine products", etc.) and which, when added to the basket, typically extend the visit, such as “restauração".

Fig 26- Process Overview (Division of products, case frequency)

[97.3% of activities and 75.4% of connections]

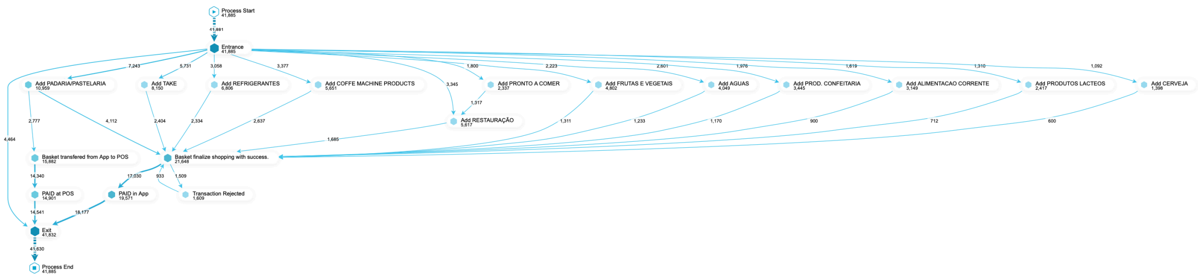
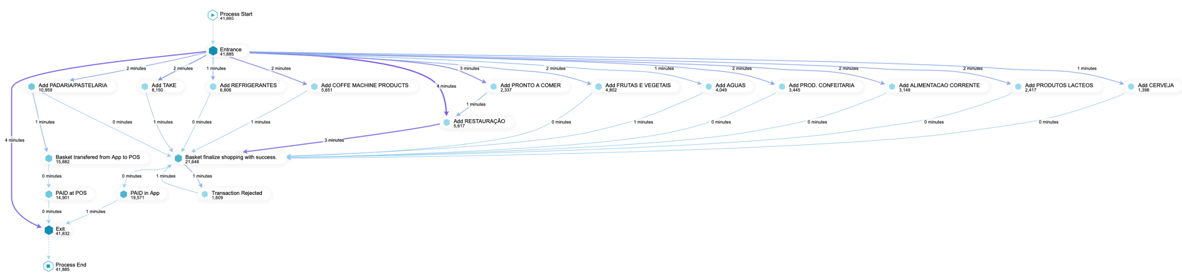


Fig 27- Process Overview (Division of products, KPI- throughput Time (AVG) in min)

[97.3% of activities and 75.4% of connections]



As a part of this section, we also sought to examine different customer journeys throughout different times of the day, as we believe customer behaviors differs. When performing the process explorer analysis of the events held in the morning *(Fig 36 to 39 in the appendix),* we noticed that the products added to the basket were mostly coffee machine, pastry and bakery products, as well as fruit and orange juices. These results make sense, as many students walk to the store in the morning to purchase their breakfasts. Moreover, we found that these visits were the fastest when compared to other times of the day, which is also in accordance with the results of *Table 9 in the appendix*. When performing the same analysis for lunch time *(Fig 40* *to 43 in the appendix),* we noticed that there were more entries in this condition and that these visits were also longer. Moreover, at lunch time, the main types of items contained in the users' baskets were primarily meals, such as chicken, pasta, pizza, and drinks (water and coke). Further, it was possible to also conclude that pizza and pasta took longer to scan at lunch time. This longer period may be justified by the fact that both meals must be prepared, which means people need to wait in line before they can scan the QR code.

Alternatively, in the afternoon *(Fig 44 to 47 in the appendix),* users' baskets and duration of visits were more similar to those in the morning, without any particular significant differences in behavior. At evening, which corresponds to dinner time *(Fig 48 to 51 in the appendix),* most of the products added to the basket were meals, including pizzas and pastas, as well as some number of beers purchased. These visits were also longer than those in the afternoon and again more similar to the duration at lunch time. Again, this can be attributed to the fact that more meal solutions were purchased at dinner time, which required meal preparation. Even so, this process appears to be faster at dinner than at lunch. In the case of pasta, people took on average 11 minutes to add it to their baskets at lunch, and only 4 minutes at dinner.

Having reviewed general customer journeys, we decided to narrow down our research to some of the most popular products sold in this store. A key feature of Celonis is the ability to filter the customer journey map, in order to include only journeys in which a specific event occurred. Consequently, we decided to investigate further the cases in which "massa personalizada", "serviço forno", and "coffee machine products" have been added to baskets, since they were among the most popular products sold during September.

To begin with, we examined specific customer journeys in which customers "Add massa personalizada." For the clients who actually decided to purchase "massa personalizada," they took an average of 8 minutes to scan the QR code *(Fig 52 and 53 in the appendix),* which, according to what was formulated earlier, corresponds to the time required to wait in the queue and reach the appropriate area in the store. Within 7/8 minutes of scanning the QR code of the pasta, customers who wished to continue with their purchases tended to add drinks (water, coke, and other carbonated soft drinks).

When customers "Add Serviço Forno" *(Fig 54 and 55 in the appendix),* they normally pre-scan different types of pizza QR-codes, typically within 3 minutes of entering the store. After that, it took an average of 4 minutes for customers who decided to continue their purchase to wait for the pizza to be cooked, move to another area, and ultimately purchase a drink. Based on similar conclusions, and since both pizza and pasta preparation services are located within the same area in the store, JM may consider creating “menus” with these two types of products, such as “Lunch Menu” and “Dinner Menu”. A further suggestion would be to place drinks near the pasta and pizza area so that customers would be reminded of their preferences for drinks when ordering their lunch.

With respect to the “coffee machine products” *(Fig 56 and 57 in the appendix),* as stated previously, it was not possible to differentiate them in the event logs. In spite of this, we tried to find complementary products, which in this case would be those products often purchased along with coffee machine products. It has been observed that these products are usually purchased with existing menu items ("croissant brioche" and "sumo natural"), with other pastry and bakery products, or even with fruit (more specifically bananas). It would therefore be appropriate to expand the existing menus to include other pastries and bakery items. As expected, visits during which customers purchased coffee machine products were concluded significantly faster than visits during which customers ordered pasta or pizza.

c) Dashboards

In response to Jerónimo Martins' request, we developed some dashboards for prototyping previously acquired information. With the help of these dashboards, it will be possible to identify and visualize essential information from the real-time data collected from the process.

This will enable the store manager to access the necessary information more quickly without the need to examine raw SAP data. The first dashboard is a general dashboard *(Fig 58 in the appendix)* that provides information regarding visits throughout the month. Using this dashboard, the store manager will be able to obtain the number of visits, the total number of events held in the store, the average duration of each visit, as well as the number of distinct brands sold. Furthermore, we have developed another dashboard *(Fig 59 in the appendix),* that provides store managers with information pertaining to the cases in which customers purchase pasta and salad. As a result of this dashboard, the store manager will be able to better monitor these events, which are often responsible for delaying the process.

In addition, we have created two additional dashboards *(Fig 60 and 61 in the appendix)* to provide more detailed information about products' divisions, areas, families, categories, and subcategories. In summary, in the first dashboard, the store manager will be able to monitor sales by day of the month, week, and time of the day, as well as the average duration of visits during which the chosen group of products was purchased. An example of this dashboard is shown in *Fig 60 in the appendix,* where the filter is applied to products that belong to the "padaria/pastelaria" division. Similarly, to the previous dashboard, in the second one *(Fig 61 in the appendix)* the store manager will be able to gain a deeper understanding of the selected product group by reviewing the top products, as well as the number of products purchased and the average duration of visits by customers who purchased these products.

5.4.2 Data Mining – Apriori Algorithm

As previously mentioned, Jerónimo Martins is committed to providing the highest level of service to its clients. It is therefore imperative to identify which products are of interest to customers and whether there is any complementary relationship between them. To discover these rules, we implemented the Apriori algorithm, whose procedure can be found in more detail in the Apriori Algorithm Guide in the appendix. In order to implement the Apriori

Algorithm, the dataset had to be represented as a long list where each transaction was included. Upon creating the transaction list in the format required by the Apriori Algorithm, it was necessary to specify the parameters of the algorithm to limit the number of rules generated by the model and select only the most useful ones.

The support of a rule is simply the number of transactions that contain antecedents and consequents. Support can be either absolute or relative *(Agrawal et al., 1993).* The former, absolute support of an itemset, represents the number of transactions where this itemset is found in the database in question, while the relative support is calculated by dividing the absolute support by the total number of transactions. Another measure used in discovering association rules is the confidence in the set of frequent items. Confidence measures the conditional probability of the consequent occurring given that the antecedent occurred. Lastly, the lift represents one of the most widely used methods for determining the degree of dependence between an antecedent and its consequence. The greater the lift value, the more interesting the rule, since it indicates a greater degree of interdependence between the items.

According to the *Table 11 in the appendix*, high support values were tested at an early stage. However, when these values were used, no association rules were generated. Therefore, this value was reduced until some association rules were identified. It should be noted that although we used low support values at the end, the minimum lift was never less than one, since we were trying to establish rules where products were positively correlated. By applying this algorithm to the shopping basket analysis, a set of association rules was identified, along with the associated support, confidence, and lift parameters.

Based on the defined parameters, we were able to identify 44 association rules. We then examined the top 25 rules in terms of confidence, which can be found in *Table 12 in the appendix.* To visualize these rules, we created a parallel coordinate graph. We can thus see the rules presented in the form of Antecedent (X) → Consequent (Y) by looking at these graphs

(Fig 62 in the appendix). In line with expectations, some of the top rules relate to the addition of pizzas and the subsequent addition of oven service. Moreover, we were also able to find an association rule between "croissant com chocolate" and "coffee machine products", where "croissant com chocolate" is the antecedent and the latter is the consequent.

Nevertheless, to obtain an accurate visualization of the quality and strength of these rules, we decided to create a heatmap to examine whether a product is likely to be added to the basket if another product has already been added *(Fig 63 in the appendix).* We found, for instance, that people who have added “perna frango assado” have a 25% probability of adding “arroz branco” afterwards. Additionally, 78% of those who have added “croissant brioche Misto” later on are likely to add a “sumo laranja natural”.

5.4.3 Comparison between Data Mining and Process Mining results

Data mining and Process Mining techniques are applied to analyze the relationships between products. Celonis was used to understand, interpret, and produce results by implementing the Process Mining methodology. By using the Apriori Algorithm, we were able to examine various associations between products, defined as association rules, as well as their support, confidence, and lift. We were able to obtain very complementary outcomes by using these two tools. It is essential to note that both techniques fall under the umbrella of business intelligence and provide the information needed for data-driven business decisions. Nevertheless, there are also some differences between these two techniques that should be considered. As part of the Data Mining process, static and arbitrary data were analyzed in order to predict patterns and investigate correlations between the data to gain new insights. Process Mining, on the other hand, analyzes real-time and dynamic data in order to determine the causes of problems and examine exceptions. As a result, it is primarily concerned with how information is generated, and how it contributes to a whole process. Although both approaches are extremely insightful, and even though they are not mutually exclusive, we believe that store managers would prefer to access information through Celonis in a more visually appealing manner.

6. Challenges And Limitations

Several conclusions and suggestions were derived from the previous sections. Furthermore, we were able to identify bottlenecks in the process, such as meal preparation, as well as flaws, such as payment failures. Even so, it is important to recognize the main limitations of this study so that possible improvements can be made in the future. This thesis presents the results of Data Curation, Exploratory Analysis, and Data Mining on the basis of a one-month events analysis. When using the Process Mining tool, however, we were not able to analyze the entire month, despite having an extended student license. Rather, only 250,000 lines were analyzed. It is also important to point out that in this thesis, some assumptions were made. To calculate the waiting time in the queue for pasta and salads in the Exploratory Analysis, we had to make an assumption that people only scan the QR code when they begin ordering, which may not always be the case. As another limitation, we did not have access to the event logs related to the specific coffee machine products, which caused us to often compare particular products with this group of products. For instance, "coffee machine products" includes a wide range of products, such as: "espresso", "duplo espresso", and "cappuccino", among others. One of the project's most challenging and critical steps was preparing the logs for the proper structure. First of all, we are aware that we had to eliminate several visits that, despite being legitimate, valid, and accurate, did not have an associated customer and could not be differentiated. In a similar manner, we had to eliminate events that did not occur between the moment the user entered and the moment they exited, in order to ensure that each visit was complete. Consequently, we are aware that in the process of elimination, we may have deleted events that in fact occurred, but were not well defined chronologically. Besides that, to identify the products, we had to make many assumptions based on the movement's dataset. For instance, we had to correct the instances where different SAP codes corresponded to the same product. Similarly, we changed the SAP codes for products with slightly different names (which we assumed would be due to spelling errors). Finally, we identified that another limitation was the inability to assign store zones to all products.

7. Recommendations For Future Steps

As stated in the previous section, the data used to support the approach, models and systems implemented consisted of data from just 15 days. JM should, therefore, seek to deepen the process analysis with increased granularity using more complete event logs, encompassing a longer time span. As part of the second phase, it would be interesting to test this solution using real-time data collection and processing. The direct association between Celonis and SAP would enable Jerónimo Martins employees to continue mining in the cloud. This would therefore be a very valuable opportunity to extend this analysis to Jerónimo Martins' ever-changing data. In this way, Celonis would be able to distill JM data into graphical representations in real time, facilitating better process monitoring. Additionally, it would be interesting to complement our analysis by accessing sales data, specifically product prices. In this way, we could analyze the financial impact of different customer journeys, and therefore understand the opportunity cost of each journey.

8. Conclusion

In *Hopping's (2000)* view, the history of retail industry reflects the progress of technology. In fact, retailers have been able to streamline and accelerate their processes over the years by implementing technology, from ATM card payments to RFID systems.

Nova & Go store exemplifies this increased ability to modernize and improve the shopping experience. Among other things, this store can gather various types of data, from sales, in-store product placement, inventory levels or even the cooking process of takeaway meals. The adoption of tools that can then synthesize, standardize, analyze, and visualize all these data collected is, therefore, essential to generate value.

For the purpose of this thesis, Jerónimo Martins identified a challenge affecting different aspects of the store. Consequently, Process Mining and later Data Mining were used to gain insight into the customer journey. By using a Process Mining tool, Celonis, the most common routes and behaviors inside the store, as well as system errors and processes bottlenecks were identified. In light of the results obtained, some recommendations were provided. For instance, changes in the placement of products on the store's shelves and potential new product menus were discussed. Later, a Data Mining tool, the Apriori Algorithm, was developed, to detect and evaluate consistent association rules and niches among products. While some of these rules were somewhat predictable, the algorithm was still able to stimulate the development of deeper knowledge about the understanding of the business.

9. References

Aalst, Wil M.P. van der, Arya Adriansyah, Ana Karla Alves de Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, et al. 2011. "Process Mining Manifesto." SpringerLink. Springer, Berlin, Heidelberg. January 1, 2011. https://www.researchgate.net/publication/221585990_Process_Mining_Manifesto.

Aalst, Wil M.P. van der. 2016. *Process Mining: Data Science in Action*. Berlin: Springer.

Aalst, Wil M.P. van der, and S. Dustdar. 2012. "Process Mining Put into Context." Undefined. January 2012. https://www.semanticscholar.org/paper/Process-Mining-Put-into-Context-Aalst-Dustdar/68fe733f2945297ff96509af8af257b95db49d11.

Aalst, Wil M.P. van der. 2011. *Process Mining Discovery, Conformance and Enhancement of Business Processes*. Berlin: Springer Berlin.

Aalst, Wil M.P. van der. 2013. "'Mine Your Own Business': Using Process Mining to Turn Big Data into Real Value." 2013. https://www.researchgate.net/publication/283581576_Mine_your_own_business_Using_process_mining_to_turn_big_data_into_real_value. .

Agrawal, R., and R. Srikant. 1994. "Fast Algorithms for Mining Association Rules in Large Databases: Semantic Scholar." Undefined. January 1, 1994. https://www.semanticscholar.org/paper/Fast-Algorithms-for-Mining-Association-Rules-in-Agrawal-Srikant/5288d14f6a3937df5e10109d4e23d79b7ddf080f.

Agrawal, Rakesh, Tomasz Imielinski, and Arun Swami. 1993. "Mining Association Rules Between Sets of Items in Large Databases." 1993. https://www.researchgate.net/publication/200043124_Mining_Association_Rules_Between_Sets_of_Items_in_Large_Databases_SIGMOD_Conference.

Alzoubi, Haitham M., and Neama A Khafajy. 2015. "The Impact of Business Process Management on Business Performance." 2015. https://www.researchgate.net/publication/292138180_THE_IMPACT_OF_BUSINESS_PROCESS_MANAGEMENT_ON_BUSINESS_PERFORMANCE_SUPERIORITY.

Ashton, Kevin. 2009. "That 'Internet of Things' Thing." 2009. https://www.itrco.jp/libraries/RFIDjournal-That%20Internet%20of%20Things%20Thing.pdf.

Badakhshan, Peyman, Jerome Geyer-Klingeberg, Muhammad El-Halaby, Thomas Lutzeyer, and Gabriela Vianna Lembo Affonseca. 2020. "Celonis Process Repository: A Bridge between Business Process Management and Process Mining: Semantic Scholar." Undefined. January 2020. https://www.semanticscholar.org/paper/Celonis-Process-Repository%3A-A-Bridge-between-and-Badakhshan-Geyer%E2%80%90Klingeberg/a27601336aebef26b7b9632475cd75e5d63cf387.

Berisha-Shaqiri, Aferdita. 2014. "Management Information System and Decision-Making." 2014.https://www.researchgate.net/publication/287205806_Management_Information_System_and_Decision-Making.

Bernard, Gaël, and Periklis Andritsos. 2017. "A Process Mining Based Model for Customer Journey Mapping." 2017. https://www.researchgate.net/publication/316521600_A_Process_Mining_Based_Model_for_Customer_Journey_Mapping.

Bok, B .G.J. 2016. "Innovating the Retail Industry; an IOT Approach." 2016. https://essay.utwente.nl/69982/1/ten%20Bok_BA_BMS.pdf.

Celonis. 2022. "What Is Process Mining?: Celonis: Process Mining Software." Celonis. 2022. https://www.celonis.com/process-mining/what-is-process-mining/.

Chen, Min, Shiwen Mao, Yin Zhang, and Leung Victor C M. 2014. *Big Data: Related Technologies, Challenges and Future Prospects*. Cham: Springer International Publishing.

Chen, Xu, and Tao Wang. 2017. "Combining Active Learning and Semi-Supervised Learning by Using Selective Label Spreading: Semantic Scholar." Undefined. January 1, 2017. https://www.semanticscholar.org/paper/Combining-Active-Learning-and-Semi-Supervised-by-Chen-Wang/5fb5226137f9ddc8455358beaaadc5e39a57adc1.

Deloitte. 2011. "Data as a Strategic Asset." Deloitte United States. 2011. https://www2.deloitte.com/us/en/pages/consulting/articles/data-strategic-asset.html.

Fayyad, Usama M. 1996. *Advances in Knowledge Discovery and Data Mining*. Menlo Park (Calif.): AAA/MIT Press.

Fayyad, Usama, Gregory Piatetsky-Shapiro, and Padhraic Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases." AI Magazine. 1996. https://ojs.aaai.org//index.php/aimagazine/article/view/1230.

Følstad, Asbjørn, and Knut Kvale. 2018. "Customer Journeys: A Systematic Literature Review." SINTEF Open. February 8, 2018. https://sintef.brage.unit.no/sintef-xmlui/handle/11250/2483763.

Gazzola, Patrizia, Daniele Grechi, Ilaria Martinelli, and Roberta Pezzetti. 2022. "The Innovation of the Cashierless Store." MDPI. Multidisciplinary Digital Publishing Institute. February 11, 2022. https://www.mdpi.com/2071-1050/14/4/2034/htm.

Han, Jiawei, Micheline Kamber, and Jian Pei. 2012. "Data Mining: Concepts and Techniques." Catalogue Home. Amsterdam; London: Morgan Kaufmann, c2012. 2012. http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf.

Har, Loh Li, Umi Kartini Rashid, Seah Choon Sen, Loh Yin Xia, and Lee Te Chuan. 2022. "Revolution of Retail Industry: From Perspective of Retail 1.0 to 4.0." Procedia Computer Science. Elsevier. March 8, 2022. https://www.sciencedirect.com/science/article/pii/S1877050922003714?fr=RR-2&ref=pdf_download&rr=777a9fa62f08866f.

Hevner, Alan R., Salvatore T. March, Jinsoo Park, and Sudha Ram. 2004. "Design Science in Information Systems Research." University of Arizona. Management Information Systems Research Center. 2004. https://experts.arizona.edu/en/publications/design-science-in-information-systems-research.

Hopping, Dan. 2000. "Technology in Retail: Semantic Scholar." Undefined. January 1, 2000. https://www.sciencedirect.com/science/article/pii/S0160791X99000421?via=ihub.

Hwang, Illhoe, and Young Jang. 2017. "Process Mining to Discover Shoppers' Pathways at a Fashion Retail Store Using a WiFi-Base Indoor Positioning System." 2017. https://ieeexplore.ieee.org/abstract/document/7926395.

Kindström, Daniel, and Christian Kowalkowski. 2016. "Service Driven Business Model Innovation: Organizing the Shift from a Product-Based to a Service-Centric Business Model." SSRN. March 15, 2016. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2747504.

Lemon, Katherine N., and Peter C. Verhoef. 2016. "Understanding Customer Experience throughout the Customer Journey." 2016. https://pure.rug.nl/ws/files/81733365/Understanding_Customer_Experience_Throughout_the_Customer_Journey.pdf.

Marquez, Joe, Annie Downey, and Ryan Clement. 2015. "Walking a Mile in the User's Shoes: Customer Journey Mapping as a Method to Understanding the User Experience." 2015. https://www.researchgate.net/publication/288006156_Walking_a_Mile_in_the_User's_Shoes_Customer_Journey_Mapping_as_a_Method_to_Understanding_the_User_Experience.

Mckinsey. 2018. "Achieving Business Impact with Data - Mckinsey & Company." 2018. https://www.mckinsey.com/~/media/mckinsey/business%20functions/mckinsey%20analytics/our%20insights/achieving%20business%20impact%20with%20data/achieving-business-impact-with-data_final.ashx.

Patrício, Lia, João Cunha, Larry Constantine, and Raymond P. Fisk. 2011. "Multilevel Service Design: From Customer Value Constellation to Service Experience Blueprint." 2011. https://www.researchgate.net/publication/235818398_Multilevel_Service_Design_From_Customer_Value_Constellation_to_Service_Experience_Blueprint.

Perkins, Ben, and Céline Fenech. 2014. "The Growing Power of Consumers - Deloitte Us." 2014. https://www2.deloitte.com/content/dam/Deloitte/uk/Documents/consumer-business/consumer-review-8-the-growing-power-of-consumers.pdf.

Raeder, Troy, and Nitesh V. Chawla. 2010. "Market Basket Analysis with Networks - Social Network Analysis and Mining." SpringerLink. Springer Vienna. August 28, 2010. https://link.springer.com/article/10.1007/s13278-010-0003-7.

Reinartz, Werner, and V. Kumar. 2000. "On the Profitability of Long-Life Customers in a Noncontractual Setting: An Empirical Investigation and Implications for Marketing: Semantic Scholar." Undefined. October 2000. https://www.researchgate.net/publication/229819388_On_the_Profitability_of_Long-Life_Customers_in_a_Noncontractual_Setting_An_Empirical_Investigation_and_Implications_for_Marketing. Ross, David Frederick. 2015. *Distribution Planning and Control: Managing in the Era of Supply Chain Management*. Berlin: Springer.

Salgueiro, Maria. 2019. "Pingo Doce & Go: Experimentámos o Novo Conceito De Supermercado Sem Filas." NiT. 2019. https://www.nit.pt/comida/gourmet-e-vinhos/ja-fomos-ao-novo-pingo-doce-sem-filas-ficamos-impressionados.

Tiwari, A., C.J. Turner, and B. Majeed. 2008. "A Review of Business Process Mining: State-of-the-Art and Future Trends." Business Process Management Journal. Emerald Group Publishing Limited. February 8, 2008. https://www.emerald.com/insight/content/doi/10.1108/14637150810849373/full/html.

Çelik, U., and E. Akçetin. 2018. "[PDF] Process Mining Tools Comparison: Semantic Scholar." Undefined. January 2018. https://www.semanticscholar.org/paper/Process-Mining-Tools-Comparison-%C3%87elik-Ak%C3%A7etin/54fd2100653165a797fa86c0125b61176b753812.

10. Appendix

Appendix Formulas

Equation 1- Anti-monotone Property of the support

$$\forall X, Y:(X \subseteq Y) \Rightarrow s(X) \geq s(Y)$$

Appendix Tables

Table 1- Information of the Dataset before Data Curation

Nº Observations	802921
Nº Columns	14
Time Interval	01/09/2022 to 30/09/2022

Table 2- Columns Description and Characteristics before Data Curation

Variable	*Eg. Observation*	*NºUnique Values*	*Null Values*	*Null Values (%)- rounded*
Created on	2022-09-30T23:59:50.000000000	474831	0	0%
Created by	System	1	64404	6%
Client	c0d519db-853b-400c-b6a0-dcf3e890a40a	6379	168989	17%
Customer Country Code	371.	48	0	0%
Event Type	CheckInRequestAllowed	47	0	0%
Event	Check in allowed	47	0	0%
Desc Event	Out Customer X Gate E 2022-09-02 11:49:09	394607	0	0%
Session UUID	-	-	452367	100%
Basket ID	1039555	67249	182511	22%
Basket GUID	5da05635-8264-4e79-9e4a-3c1f92d1692e	67235	183548	22.1%

Table 3- Nº Observations, Columns and Duplicated Rows After Data Curation

Nº Observations	**334174**
Nº Columns	53
Nº Duplicate rows	0
Time Interval	01/09/2022 to 29/09/2022

Table 4- Columns Description and Characteristics after Data Curation

VARIABLE	EXISTING/CREATED	EG. OBSERVATION	NºUNIQUE VALUES	NULL VALUES	NULL VALUES (%)
Created by	DELETED				
Client	EXISTING	c0d519db-853b-400c-b6a0-dcf3e890a40a	6295	0	0%
Customer Country Code	EXISTING	+371	48	0	0%
Created on	EXISTING	2022-09-30T23:59:50.000000000	259600	0	0%
Event type	EXISTING	CheckInRequestAllowed	16	0	0%
Event	EXISTING	Check in allowed	16	0	0%
Desc Evento	EXISTING	Out Customer X Gate E 2022-09-02 11:49:09	264874	0	0%
Session UUID		DELETED			
Basket ID		DELETED			
Basket GUID		DELETED			
Customer Country	CREATED	Portugal	48	0	0%
Event Category	CREATED	Check In/Out	3	0	0%
Event Desc 2	CREATED	TABLETE M&M'S BLOCK CHOCO 165G	59631	0	0%
Id Article	MOV_STOCK	5.280000e+02	1517	186146	68%
Desc Article	MOV_STOCK	ACTIVIA LQ CER DANONE 155G	1619	219465	66%
Sub Categ	MOV_STOCK	6053 - IOGURTE BIFIDUS LIQ	558	227762	68%
Categ	MOV_STOCK	4183 - IOGURTE SAUDE	239	227762	68%
Desc brand	MOV_STOCK	ACTIVIA	269	233272	70%
Code Sub Categ	MOV_STOCK	5363	558	227762	68%
Sub Categ Name	MOV_STOCK	APERITIVOS FRITOS	558	227762	68%
Code Categ	MOV_STOCK	4183	239	227762	68%
Categ Name	MOV_STOCK	PRONTO A COMER	239	227762	68%
Code Family	MOV_STOCK	3049	93	227762	68%
Family Name	MOV_STOCK	IOGURTES E SOBREMESAS	93	227762	68%
Code Division	MOV_STOCK	2018	29	227762	68%
Division Name	MOV_STOCK	PRODUTOS LACTEOS	29	227762	68%
Code Area	MOV_STOCK	1003	11	227762	68%
Area Name	MOV_STOCK	PERECÍVEIS NÃO ESPECIALIZADOS	11	227762	68%
Our Product Category	CREATED	PADARIA E PASTELARIA	20	227762	68%
Store Zone	PLANOGRAMA	CONGELADOS LAB STORE	45	219465	66%
Store Zone Journey	CREATED	Add Product from 'CONGELADOS LAB STORE	96	0	0%
Product Journey	CREATED	Add BOL RECH PD LARANJA 300G	2968	0	0%
Area Journey	CREATED	Add PERECÍVEIS NÃO ESPECIALIZADOS	31	0	0%
Division Journey	CREATED	Add PRODUTOS LACTEOS	66	0	0%
Family Journey	CREATED	Add IOGURTES E SOBREMESAS	183	0	0%
Category Journey	CREATED	Add IOGURTE SAUDE	415	0	0%
Subcategory Journey	CREATED	Add IOGURTE BIFIDUS LIQ	898	0	0%
Our_Product Category Journey	CREATED	Add LACTEOS	50	0	0%
Event Day	CREATED	06/09/2022	25	0	0%
Event Time	CREATED	17:39:47	45682	0	0%
Event Minutes	CREATED	39	60	0	0%
Weekday	CREATED	Tuesday	6	0	0%
Time of the Day	CREATED	Afternoon	5	0	0%
Hour Intervals	CREATED	[17, 18)	15	0	0%
Visit Id	CREATED	76173	57881	0	0%
Entry Datetime	CREATED	17:39:47	29450	0	0%
Exit Datetime	CREATED	17:45:04	29690	0	0%
Duration	CREATED	0 days 00:05:17	1688	0	0%
Duration Minutes	CREATED	5.28	1688	0	0%
Duration Seconds	CREATED	317	1688	0	0%
Duration Bins	CREATED	[4, 5)	30	0	0%
Adding Removing	CREATED	Add	3	0	0%
Number Prod	CREATED	1	3	0	0%

Table 5- Event Category Variable- Unique Values and Characteristics

EVENT CATEGORY	EVENT OBSERVATIONS
APP SETTINGS/PROFILE	Client register
	Device registered for first time'
	New Payment Method
	Alert registration in new device
	Client Enable NFC
	Poupa Mais card added
	Client Edit Profile
	Client Disable NFC
	Staff See Customer Profile
	Payment Method removed
	Customer PIN Disabled
	Customer PIN Saved
	Customer PIN Enabled
	Staff Edit Client Profile
PAYMENT	Client receipt received from POS
	Payment successful at POS
	Basket sent to POS for self check out
	Payment successful in App
	Basket Finalize Shopping
	Basket Unable To Pay
	POS returned basket
	Basket in POS timed out
	Basket paid in BO
ADD/REMOVE PRODUCTS	Coffee Machine Products Added To Basket
	Client Add Product Scan
	Client Add Product Manually
	Client Remove Product Manually
	Staff Remove Product
	Staff Add Product
	Client Add Product HCE
	Client Remove Product HCE
CHECK IN/OUT	Check in allowed
	Check out allowed
	Check out denied
	Check in denied
GO 24/7	BuyBye basket snapshot created
	BuyBye Basket Unable To Pay
BASKET AUDIT	Basket Audit Pending
	Basket Calculate Risk
	Basket Audit Done Notify
	Basket Set Risk Feedback
	Basket Audit OK
	Basket Audit Created Notify
	Basket Audit Created
	Basket Audit NotOK
	Basket Audit Skipped
	Basket Calculate Risk Error
BASKET STATES	Basket no shopping
	Finalize Shopping With Error
	Basket in debt
	Basket snapshot created
	Basket Abandoned At Staff
OTHERS	Basket Requires Strong Authentication
	Coffee Machine Associated To User
	Callout tag
	Pingo Doce And Go NOVA Error
	Item added to basket queue
	Export Customer Data
	New Complaint
	Phone call
	Basket valuation
	Basket Set As Error

Table 6- Auxiliar columns for Customer Journey

	VISIT_ID	STORE_ZONE_JOURNEY	PRODUCT_JOURNEY	AREA_JOURNEY	DIVISION_JOURNEY	FAMILY_JOURNEY	CATEGORY_JOURNEY	SUBCATEGORY_JOURNEY	OUR_PRODUCT_CATEGORY_JOURNEY
21	8	Entrance	Entrance	Entrance	Entrance	Entrance	Entrance	Entrance	Entrance
22	8	Add Product from 'BATATAS FRITAS LAB STORE'	Add SUNBITES SAL MARINHO 95G	Add MERCEARIA + PET FOOD	Add PROD. CONFEITARIA	Add APERITIVOS	Add APERITIVOS EMBALADOS	Add APERITIVOS FRITOS	Add APERITIVOS E FRUTOS SECOS
23	8	Basket transfered from App to POS	Basket transfered from App to POS	Basket transfered from App to POS	Basket transfered from App to POS	Basket transfered from App to POS	Basket transfered from App to POS	Basket transfered from App to POS	Basket transfered from App to POS
24	8	PAID at POS	PAID at POS	PAID at POS	PAID at POS	PAID at POS	PAID at POS	PAID at POS	PAID at POS
25	8	Exit	Exit	Exit	Exit	Exit	Exit	Exit	Exit

Table 7- Top Nationalities- Information provided by Life at NOVA SBE on 7th of November 2022

NATIONALITY	% STUDENTS
PORTUGAL	50%
GERMANY	21%
ITALY	7%
FRANCE	2%
BRAZIL	2%
BELGIUM	1%
AUSTRIA	1%
SPAIN	1%
OTHERS	14%

Table 8- Time of the day- Schedules

PERIODS OF DAY	INTERVAL OF HOURS
MORNING	7:30 - 12:00
LUNCHTIME	12:00- 15:00
AFTERNOON	15:00- 19:30
DINNERTIME	19:30- 21:00

Table 9- Number of Visits and Visit Duration Throughout Weekday and Time of the Day

TIME DISTRIBUTION	*OBSERVATIONS*	*Nº VISITS*	*AVERAGE DURATION*
Day of Week	**Monday** (4 Mondays in September)	11602	04.59 min
	Tuesday (4 Tuesdays in September)	14185	4.49 min
	Wednesday (4 Wednesdays in September)	13586	04.51 min
	Thursday (5 Thursdays in September)	14316	4.56 min
	Friday (5 Fridays in September)	10905	4.53 min
	Saturday (4 Saturdays in September)	1833	6.04 min
Time of The Day	**Morning** (7:30 - 12:00)	14428	4.07 min
	Lunch Time (12:00- 15:00)	26223	5.49 min
	Afternoon (15:00- 19:30)	22539	4.26 min
	Evening (19:30- 21:00)	3236	5.52 min

Table 10- Items with the Highest Support

support	itemsets
0.147320	(COFFE MACHINE PRODUCTS)
0.050190	(MERENDA MISTA 95 G)
0.038070	(BANANA UNIDADE)
0.037389	(SERVIÇO FORNO)
0.034355	(MASSA PERSONALIZADA)
0.026554	(PÃO DE QUEIJO UN)
0.024764	(FOLHADO DE SALSICHA 100 G)
0.024414	(SUMO LARANJA 330 ML)
0.020873	(ÁGUA LUSO SPORT 75CL)
0.018131	(MERENDA DE CHOCOLATE 80 G)

Table 11- Choosing a Parameter for the Apriori Algorithm

Minimum Support	Lift	Number of Rules
0.006	1	0
0.005	1	8
0.004	1	18
0.003	1	18
0.002	1	38

Table 12- Top 25 Rules in Terms of Confidence

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
16	(PIZZA FRESCA P DOCE FIAMBRE QUEIJO 410G)	(SERVIÇO FORNO)	0.002568	0.037389	0.002509	0.977273	26.137723
32	(CROISSANT BRIOCHE MISTO, COFFE MACHINE PRODUCTS)	(SUMO LARANJA NATURAL 250 ML)	0.002140	0.003716	0.002043	0.954545	256.902665
18	(PIZZA FRESCA PDOCE PRESUNTO SERRANO 390G)	(SERVIÇO FORNO)	0.006167	0.037389	0.005855	0.949527	25.395643
20	(PIZZA FRESCA PINGO DOCE 4 QUEIJOS 390G)	(SERVIÇO FORNO)	0.005953	0.037389	0.005564	0.934641	24.997501
26	(PIZZA FRESCA PINGO DOCE PEPPERONI 410G)	(SERVIÇO FORNO)	0.006147	0.037389	0.005739	0.933544	24.968182
24	(PIZZA FRESCA PINGO DOCE FRANGO 400G)	(SERVIÇO FORNO)	0.004513	0.037389	0.004202	0.931034	24.901055
22	(PIZZA FRESCA PINGO DOCE CARBONARA 415G)	(SERVIÇO FORNO)	0.004182	0.037389	0.003871	0.925581	24.755209
28	(PIZZA FRS P DOCE QJ/FIAMB/COG/AZEIT 415G)	(SERVIÇO FORNO)	0.006167	0.037389	0.005564	0.902208	24.130079
9	(SUMO LARANJA NATURAL 250 ML)	(COFFE MACHINE PRODUCTS)	0.003716	0.147320	0.003229	0.869110	5.899458
33	(CROISSANT BRIOCHE MISTO, SUMO LARANJA NATURAL...	(COFFE MACHINE PRODUCTS)	0.002354	0.147320	0.002043	0.867769	5.890353
30	(PIZZA FRS S/GLUT S/LACT QJ/FIAMB PD 420G)	(SERVIÇO FORNO)	0.004144	0.037389	0.003579	0.863850	23.104161
14	(PIZZA CALZONE FRANGO PINGO DOCE 315G)	(SERVIÇO FORNO)	0.002665	0.037389	0.002257	0.846715	22.645890
10	(CROISSANT BRIOCHE MISTO)	(SUMO LARANJA NATURAL 250 ML)	0.003035	0.003716	0.002354	0.775641	208.753021
2	(CROISSANT BRIOCHE MISTO)	(COFFE MACHINE PRODUCTS)	0.003035	0.147320	0.002140	0.705128	4.786361
35	(CROISSANT BRIOCHE MISTO)	(COFFE MACHINE PRODUCTS, SUMO LARANJA NATURAL ...	0.003035	0.003229	0.002043	0.673077	208.430839
11	(SUMO LARANJA NATURAL 250 ML)	(CROISSANT BRIOCHE MISTO)	0.003716	0.003035	0.002354	0.633508	208.753021
34	(COFFE MACHINE PRODUCTS, SUMO LARANJA NATURAL ...	(CROISSANT BRIOCHE MISTO)	0.003229	0.003035	0.002043	0.632530	208.430839
13	(GARRAFÃO REUTILIZÁVEL ECO 1,5LT)	(ÁGUA ECO 1,5LT)	0.004727	0.017508	0.002821	0.596708	34.081962
37	(SUMO LARANJA NATURAL 250 ML)	(CROISSANT BRIOCHE MISTO, COFFE MACHINE PRODUCTS)	0.003716	0.002140	0.002043	0.549738	256.902665
1	(PERNA FRANGO ASSADO)	(ARROZ BRANCO)	0.008287	0.009785	0.002043	0.246479	25.189357
0	(ARROZ BRANCO)	(PERNA FRANGO ASSADO)	0.009785	0.008287	0.002043	0.208748	25.189357
7	(MERENDA DE CHOCOLATE 80 G)	(COFFE MACHINE PRODUCTS)	0.018131	0.147320	0.003307	0.182403	1.238142
12	(ÁGUA ECO 1,5LT)	(GARRAFÃO REUTILIZÁVEL ECO 1,5LT)	0.017508	0.004727	0.002821	0.161111	34.081962
5	(CROISSANT COM CHOCOLATE 100 G)	(COFFE MACHINE PRODUCTS)	0.015854	0.147320	0.002529	0.159509	1.082737
19	(SERVIÇO FORNO)	(PIZZA FRESCA PDOCE PRESUNTO SERRANO 390G)	0.037389	0.006167	0.005855	0.156608	25.395643

Table 13 – Top 10 of Most Prepared Food Recipes, in Total Number of Batches

	Food Recipe	Total Number of Batches	Frequency
0	MISTO PASTELARIA NOVO	1041	0.27
1	PAO	929	0.24
2	PAO DE QUEIJO PD	423	0.11
3	SOPA	234	0.06
4	FRANGO	183	0.05
5	PASTEL DE NATA	175	0.05
6	BURGUER VEGGIE	152	0.04
7	BURGUER CONGELADO	149	0.04
8	SALGADOS P.D	130	0.03
9	CROQUETES P.D	104	0.03

Appendix Figures

Fig 1– Data Science Research

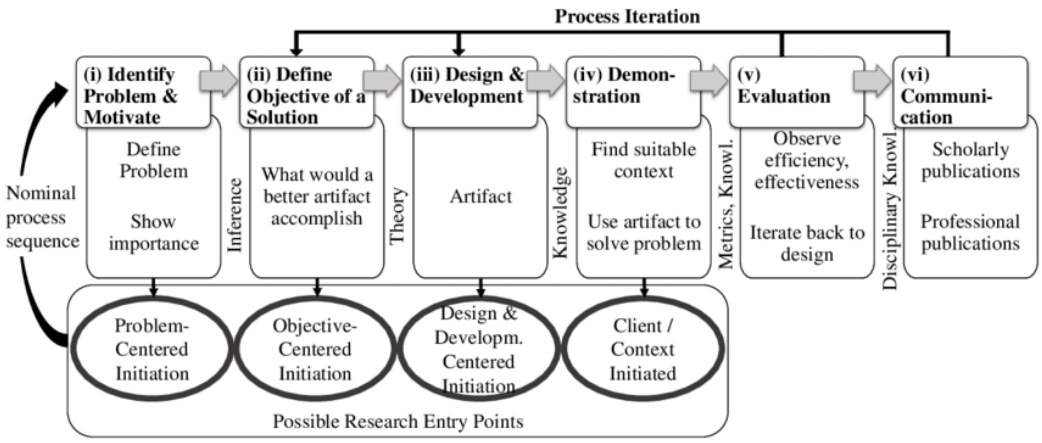


Fig 2– Top 10 Customer's Nationalities, in percentage (%)

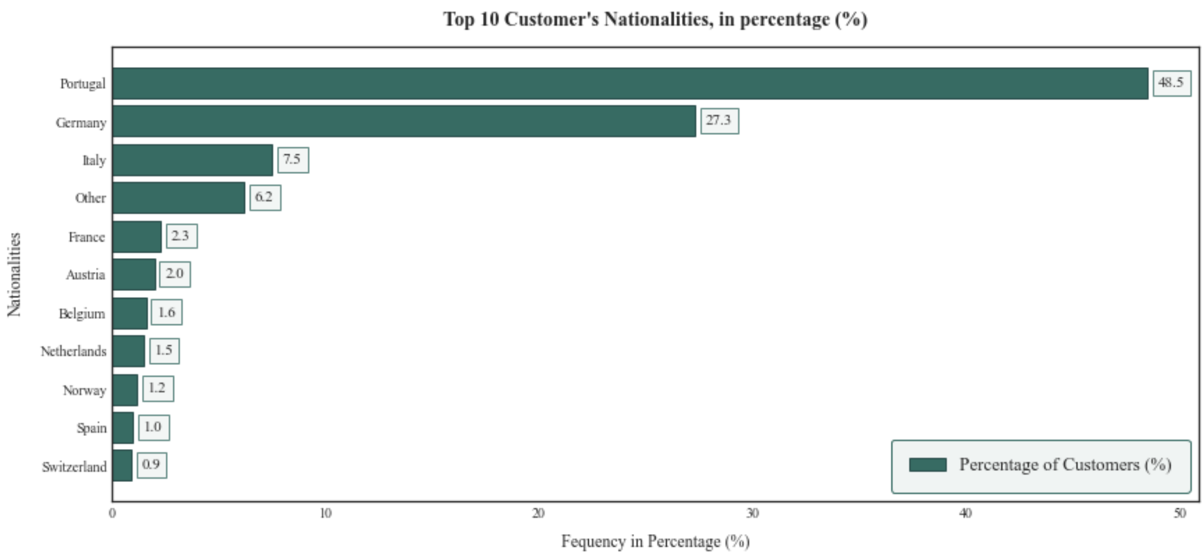


Fig 3– Average Number of Entries, by Weekday

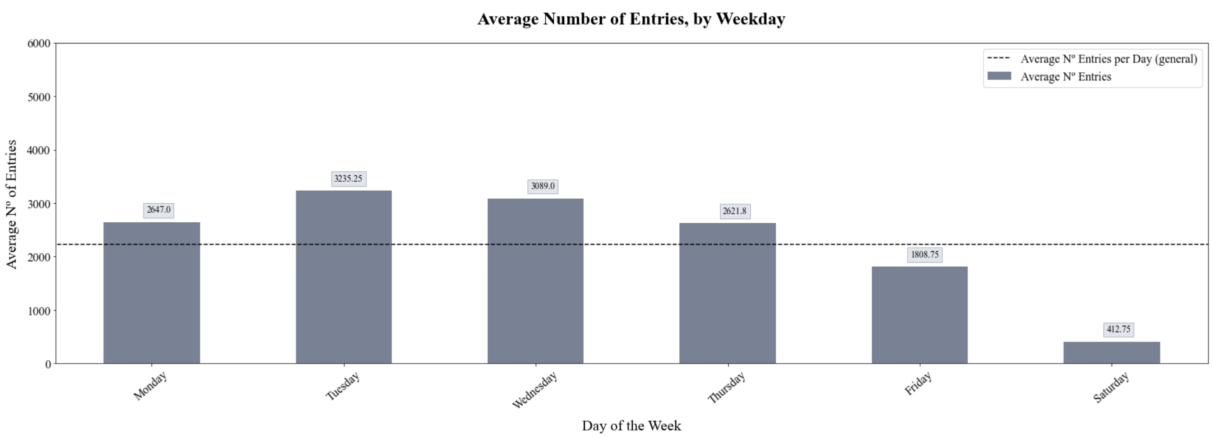
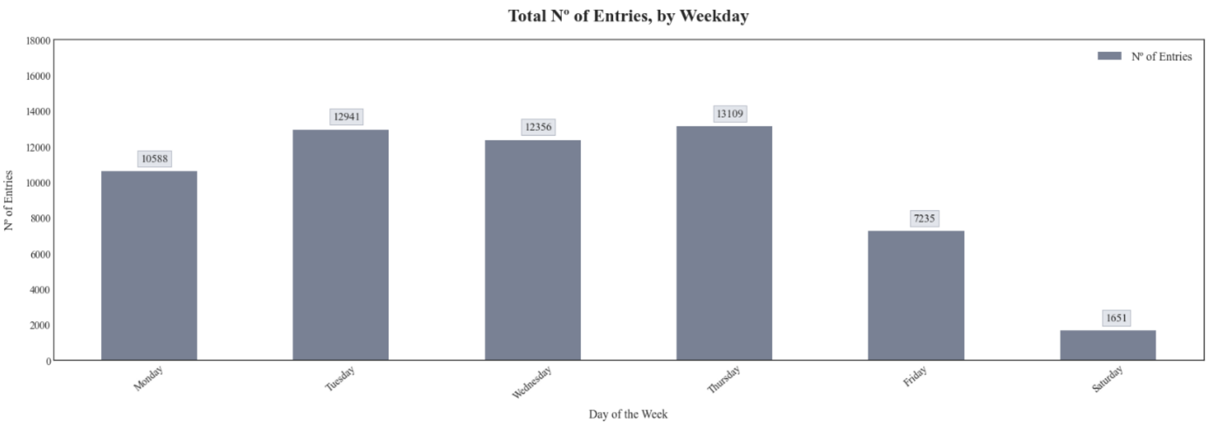


Fig 4– Total Nº of Entries, by Weekday (absolute terms)



In Fig 3. it was possible to identify that Tuesday was the day with the most entries per day (in relative terms). However, in absolute terms, Thursday was the day with the highest number of entries on average. The same happened because September had 5 Thursdays and only 4 Tuesdays.

Fig 5– Total Number of Entries on Weekdays, by Time of the Day

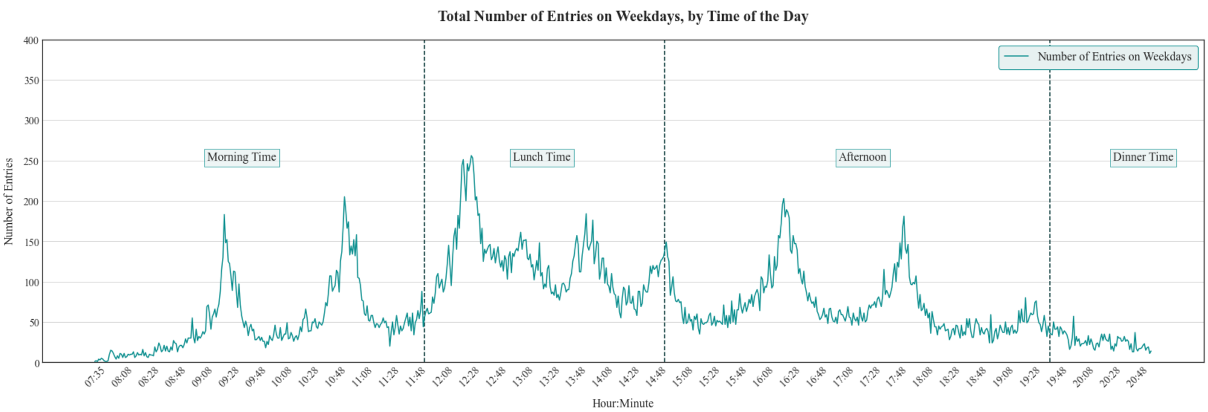


Fig 6– Total Number of Entires on Saturdays, by Time of the Day

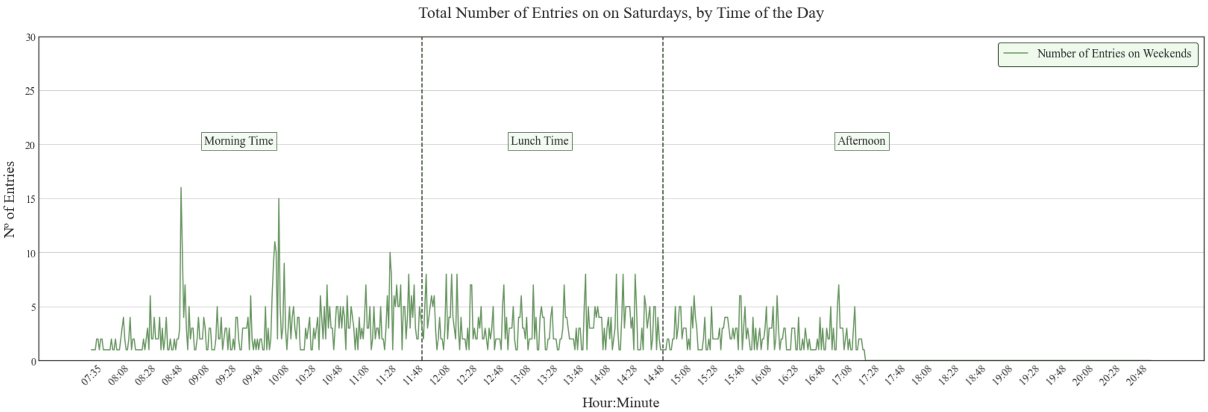


Fig 7– Total Number of Entries by day, on September 20222

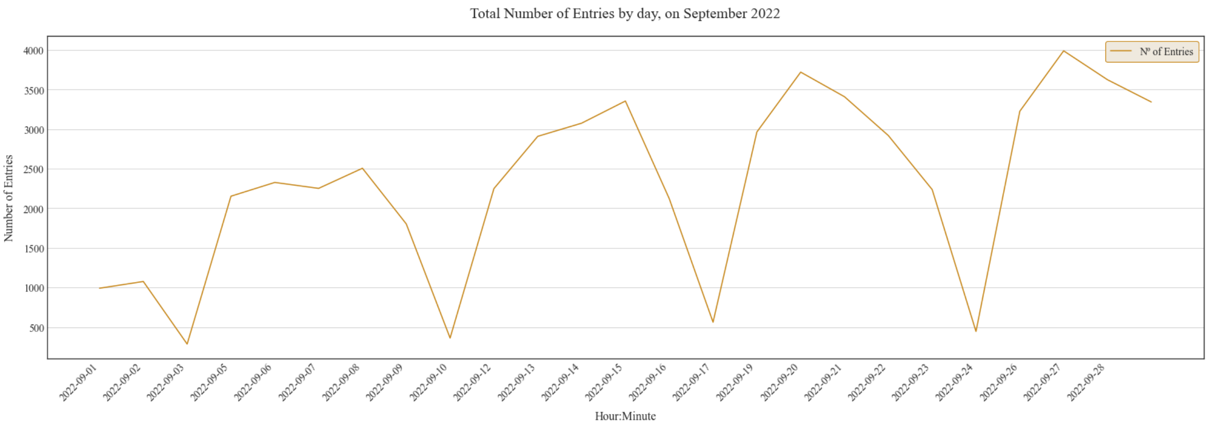


Fig 8– Nº of Visits by Duration

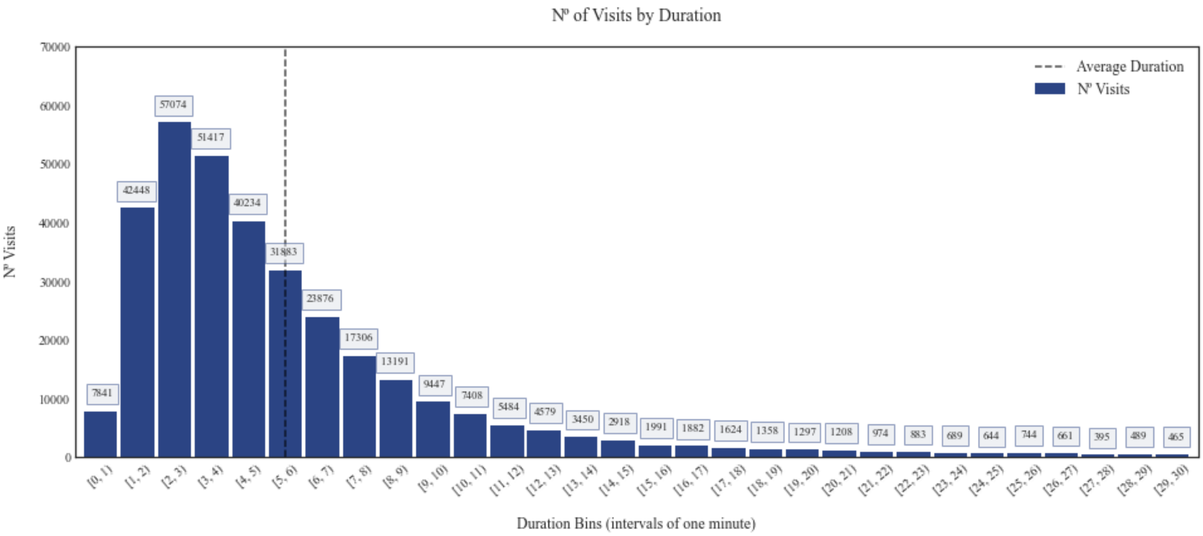


Fig 9– Average Duration by Time of the day (30 min Bins)

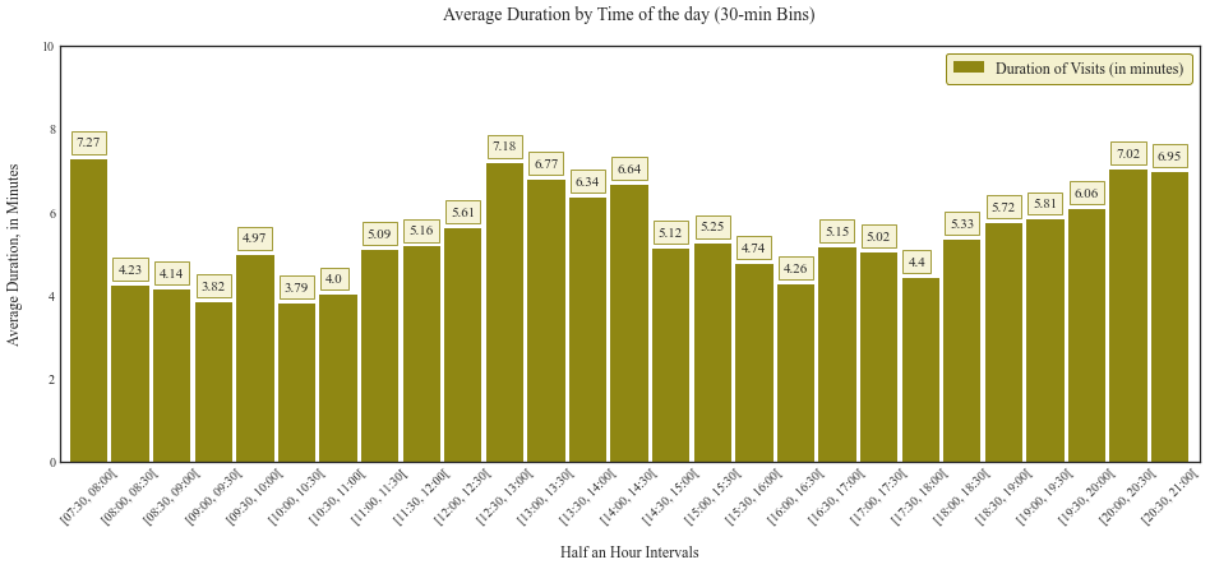


Fig 10– Nº of Orders by Queue Duration for Massa and Salada Personalizada

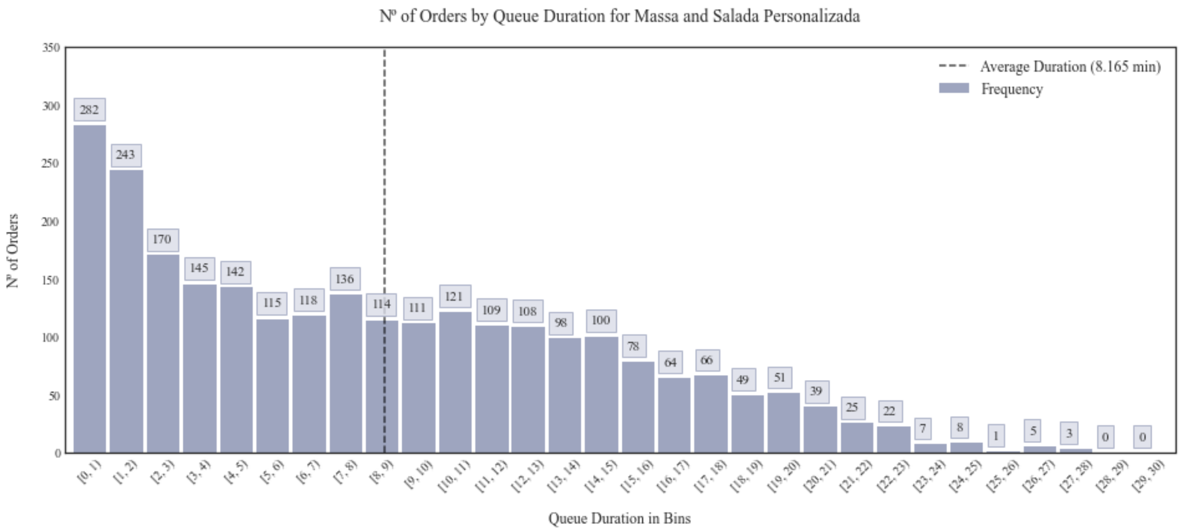


Fig 11– Average Queue Duration for Ordering Massa and Salada Personalizada, by Time of the Day (30- min Bins)

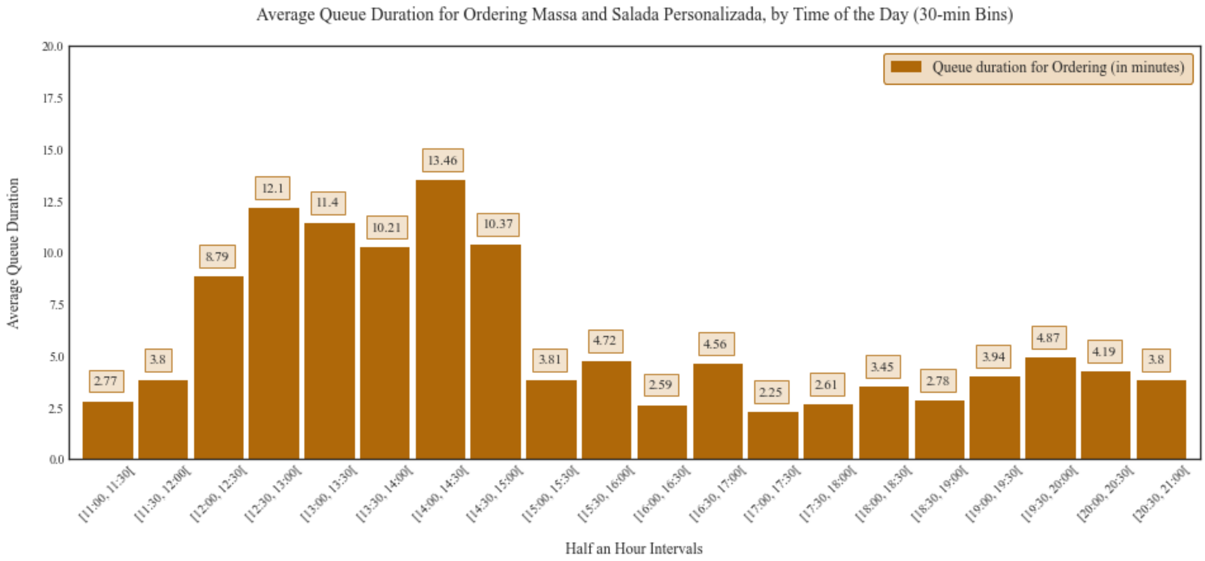


Fig 12– Total Number of Visits on Weekdays, by Time of the Day

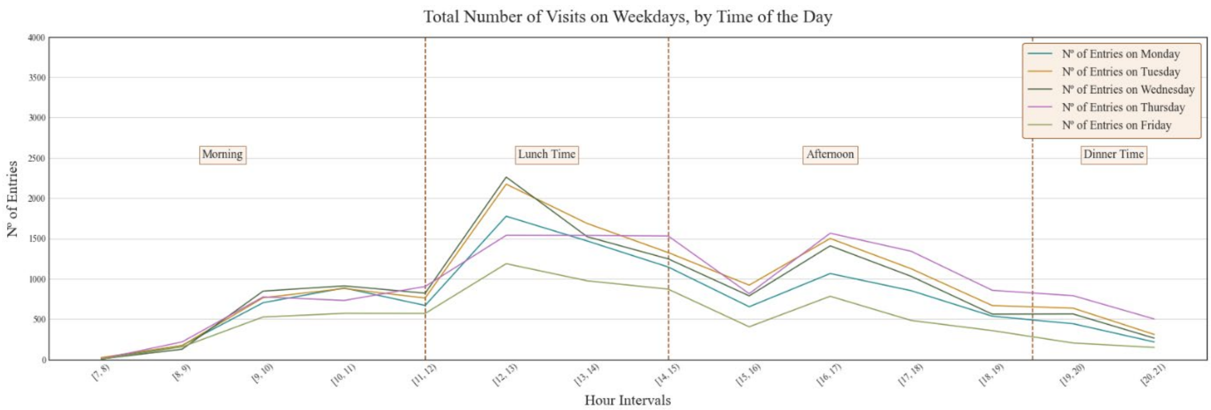


Fig 13– Total Number of Payments per Hour, by Payment Method

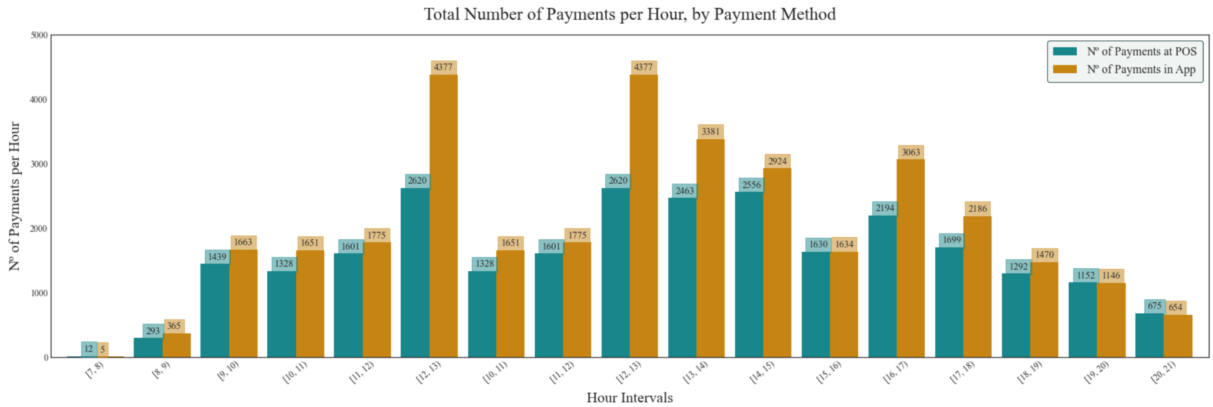


Fig 14– Number of Payments per Hour

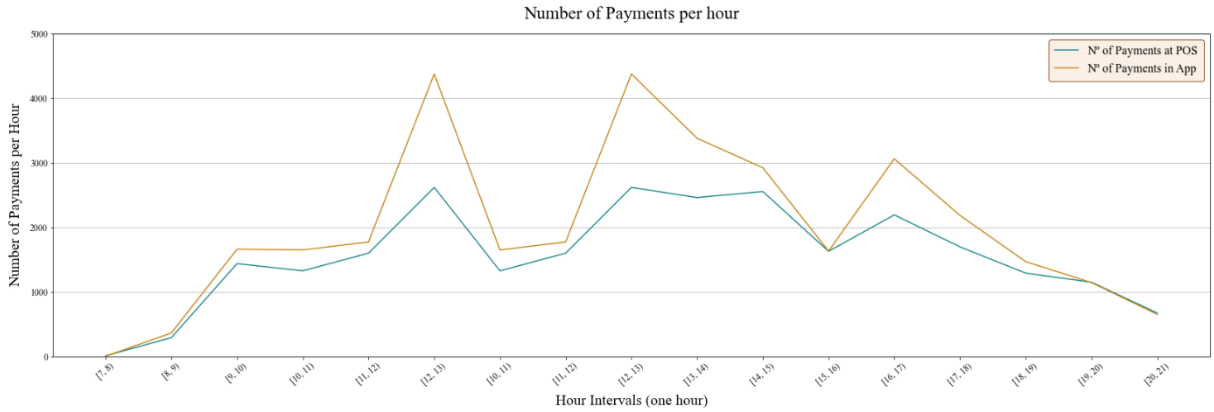


Fig 15– Number of Payments by day

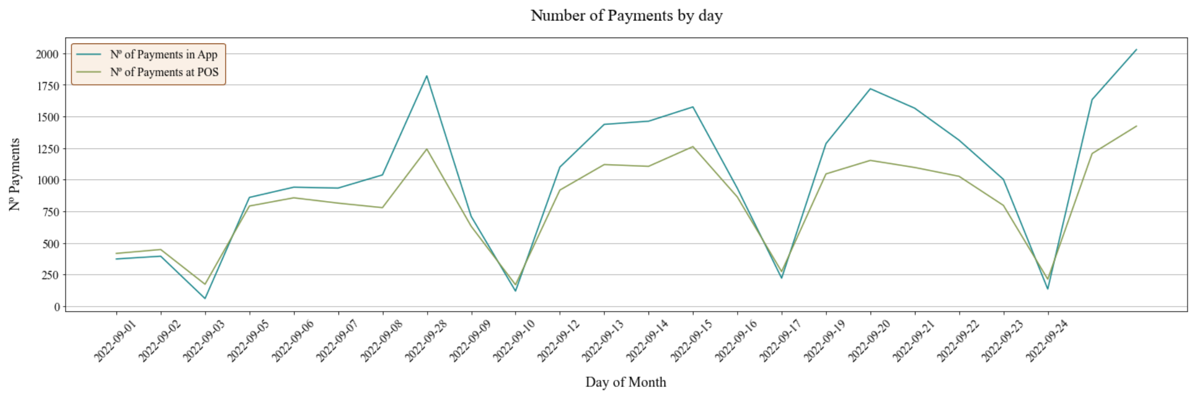


Fig 16– Top 25 Different Itemsets

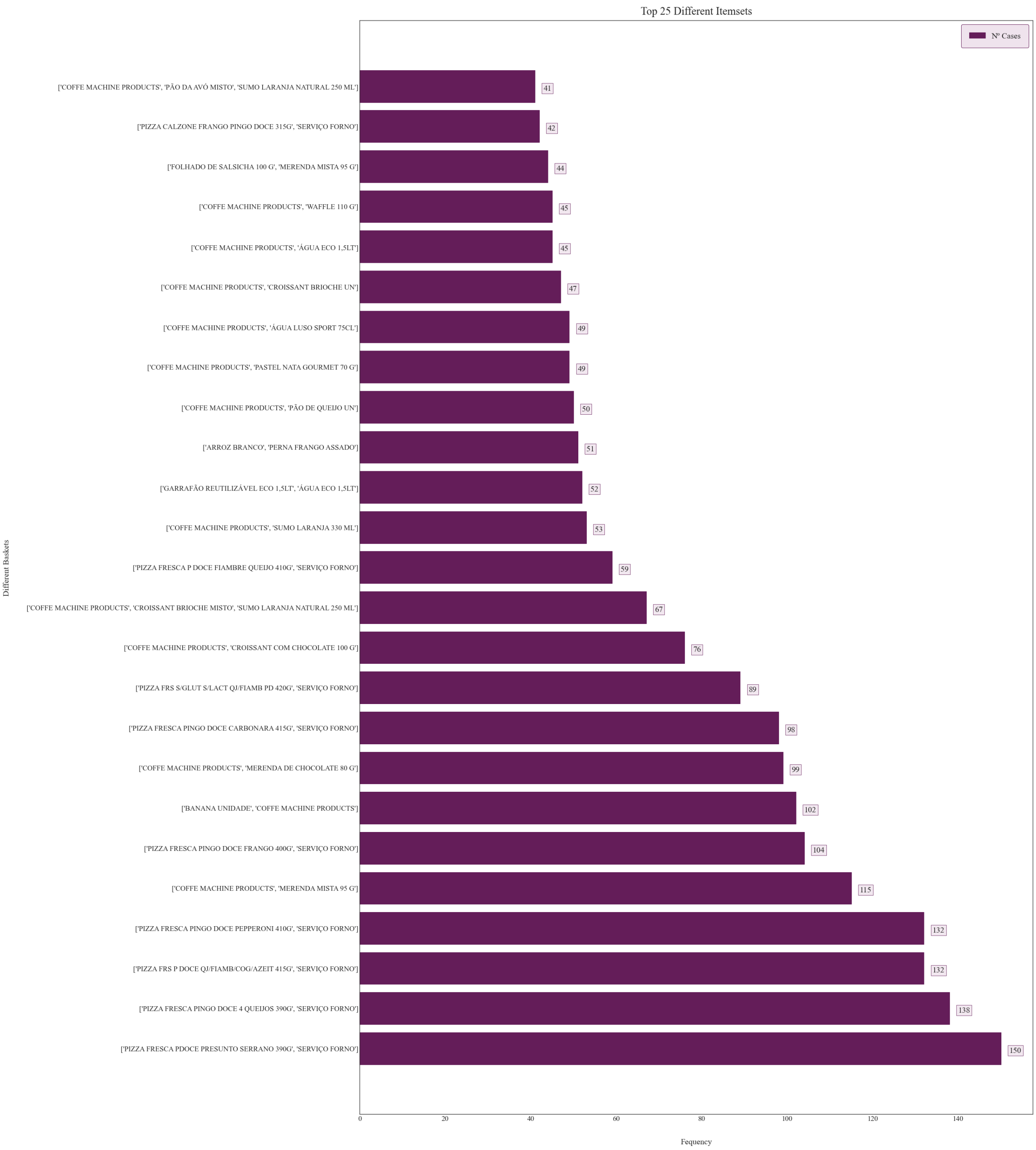
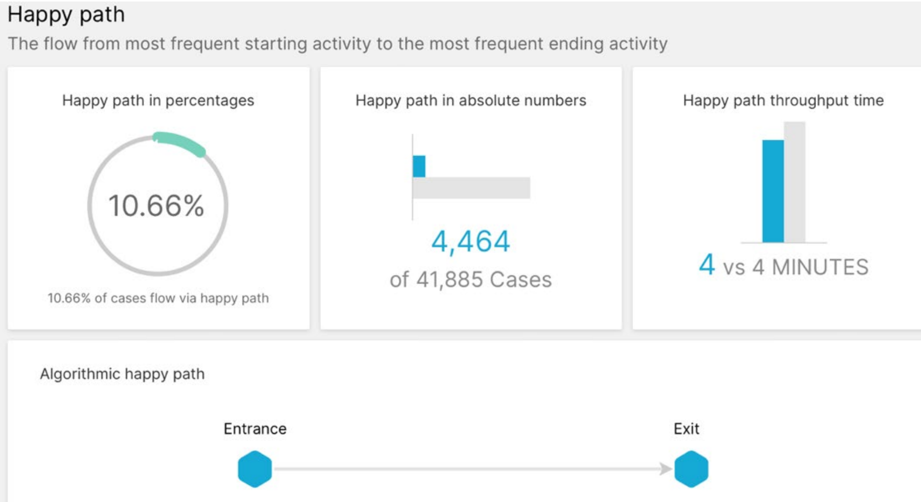


Fig 17– Process Overview (key statistics about the process)

Fig 18– Happy Path, which ultimately turns out to be an Unhappy Path



ACTIVITIES FIGURES

Fig 19– Activity Chart, based on the Product Subcategory

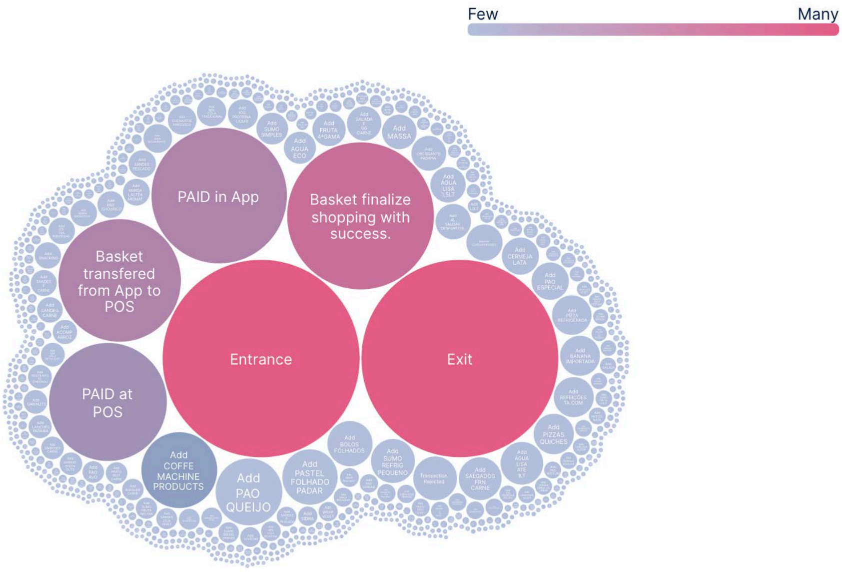


Fig 20– Activity Chart, based on Product Category

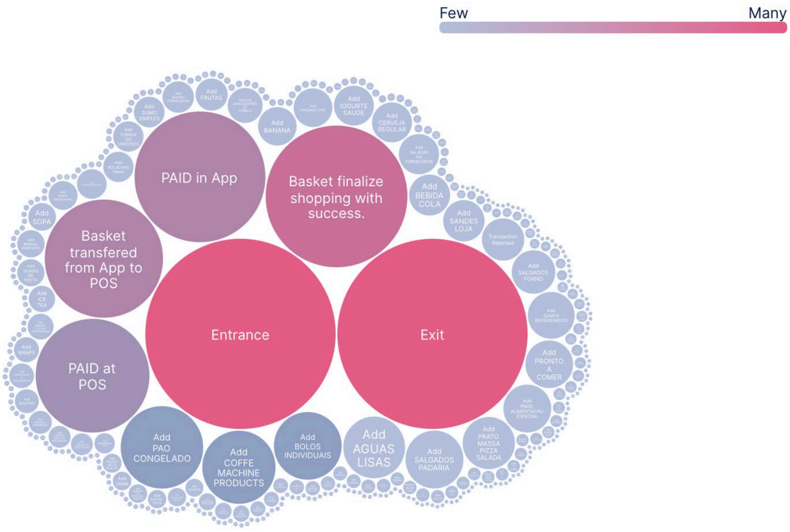


Fig 21– Activity Chart, based on Product Area

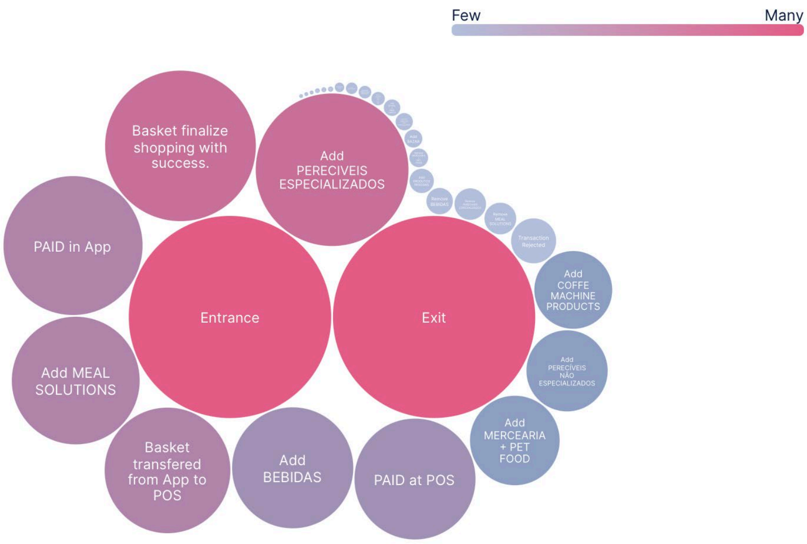


Fig 22– Activity Chart, based on Product Division

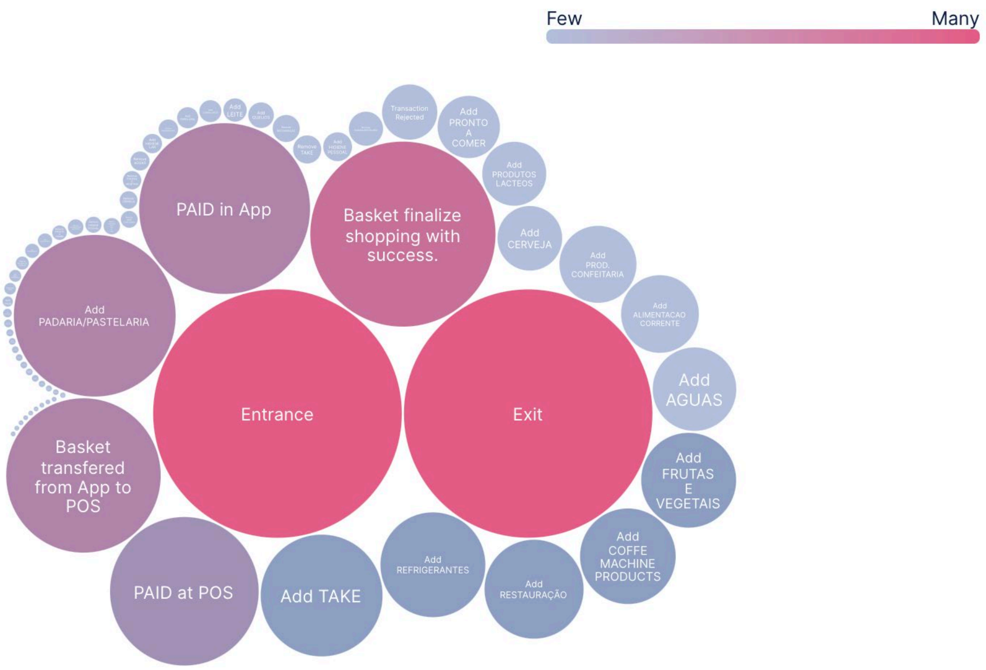


Fig 23– Activity Chart, based on Product Store Zone

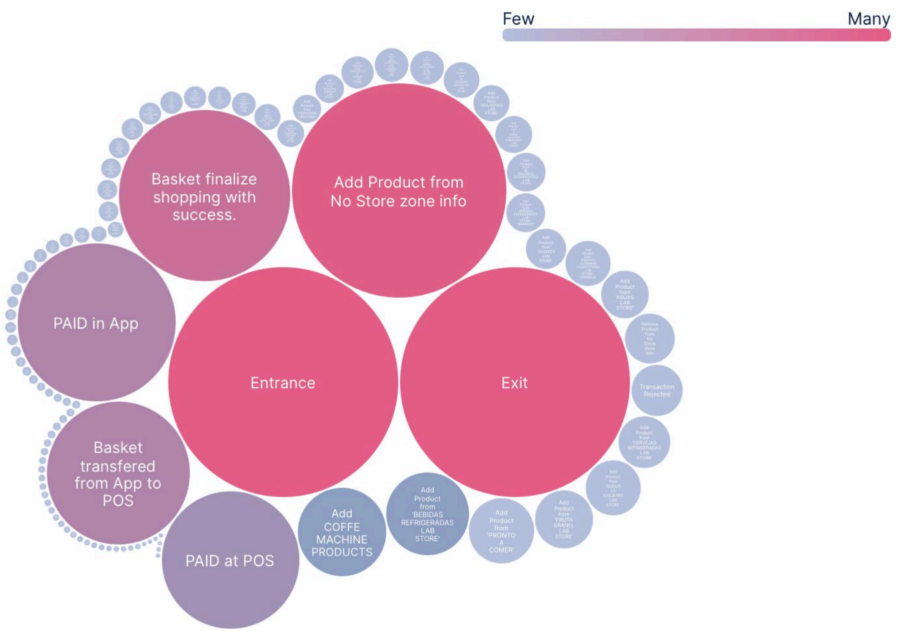


Fig 24– Activity Chart, based on Product Family

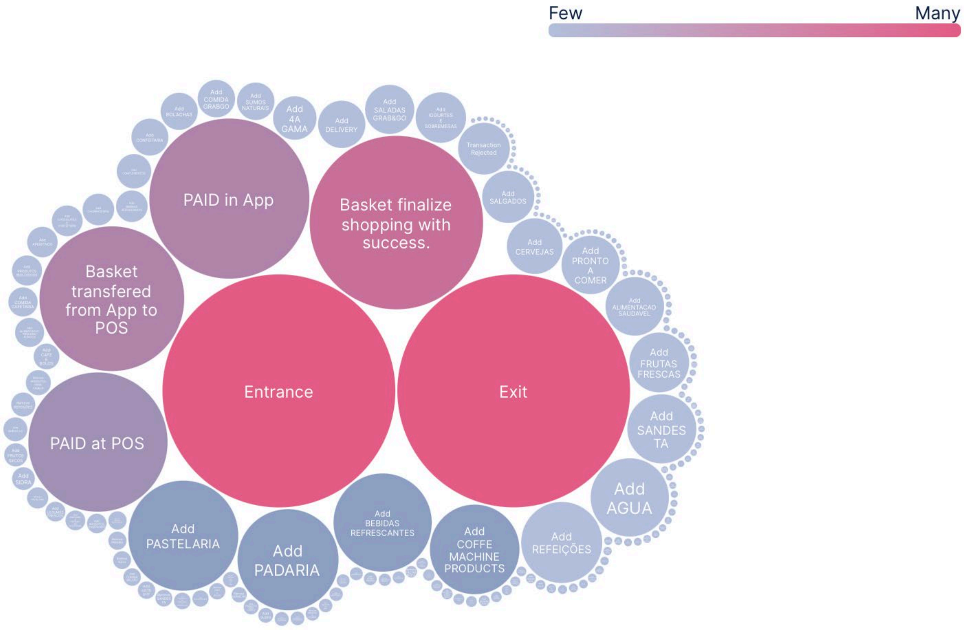
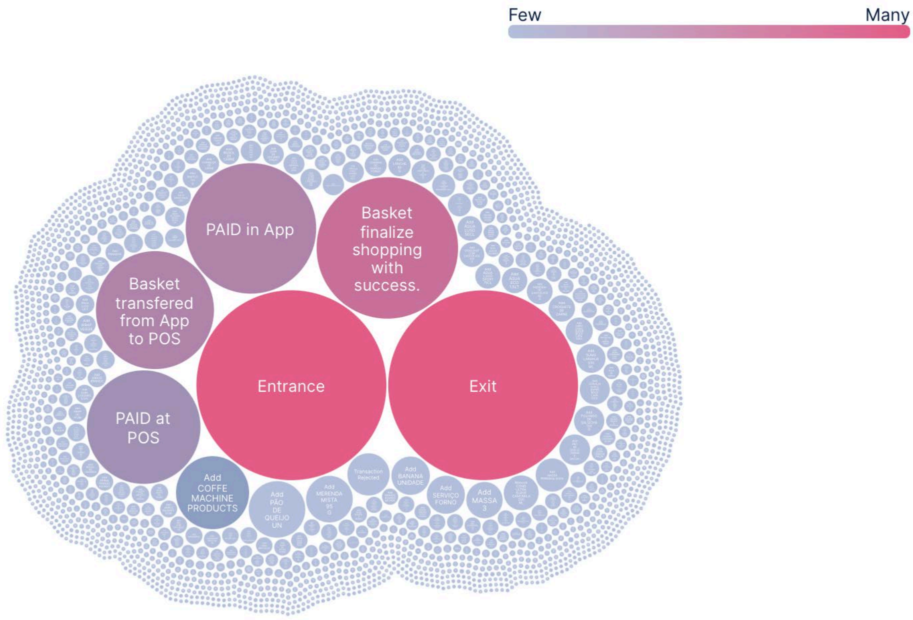


Fig 25– Activity Chart, based on Product



EXPLORER FIGURES

It is important to note that even though the customer journeys will be visualized primarily by the specific product and its division, we will always contextualize our conclusions with other customer journeys developed based on the specific product, family, category, and store zone (*Fig 26* to *Fig 35*). Through it, we will be able to gain a deeper understanding of the customer journey.

Fig 26– Process Overview (Division of products, case frequency)
[97.3% of activities and 75.4% of connections]

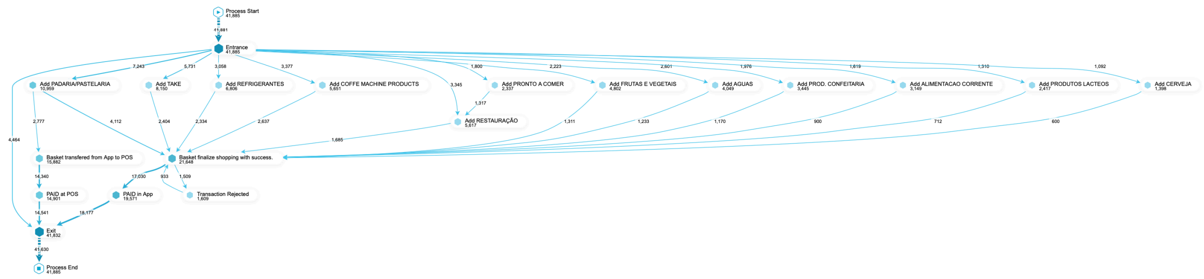
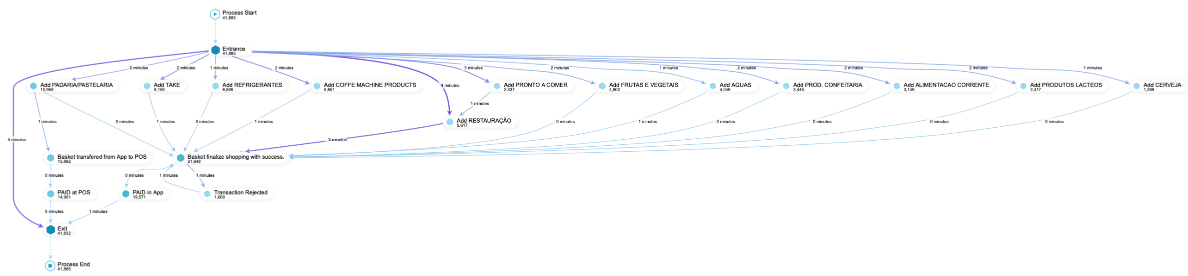


Fig 27– Process Overview (Division of products, KPI- throughput Time (AVG) in min)
[97.3% of activities and 75.4% of connections]



<u>Product:</u>

Fig 28– Process Overview (Products' names, General, case frequency)
[74.3% of activities and 61.4% of connections]

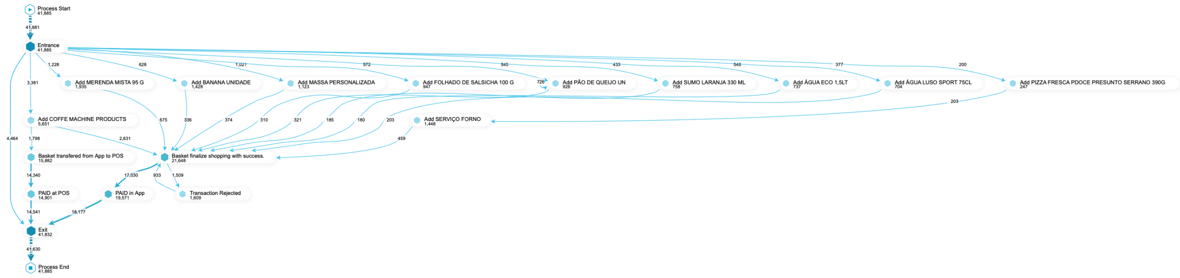
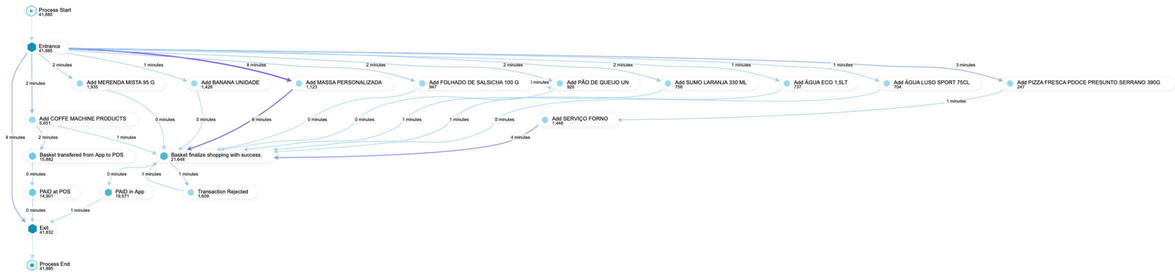


Fig 29– Process Overview (Products' names, General, throughput Time (AVG) in min)
[74.3% of activities and 61.4% of connections]



<u>Family:</u>

Fig 30– Process Overview (Family, General, case frequency)
[83.6% of activities and 65.1% of connections]

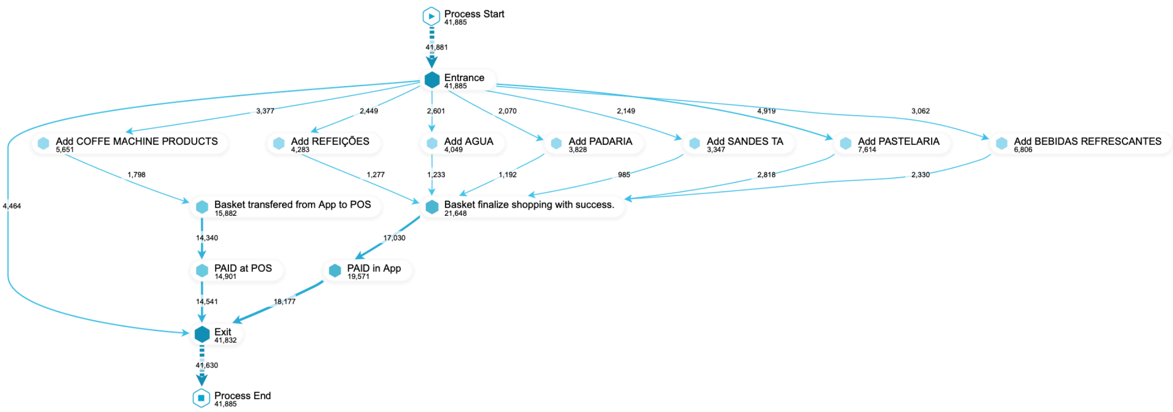
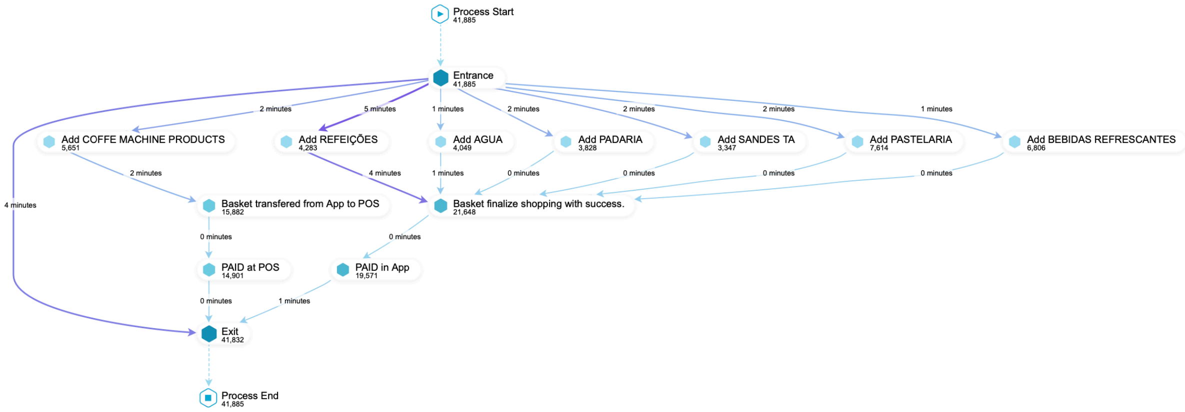


Fig 31– Process Overview (Family, General, throughput Time (AVG) in min)
[83.6% of activities and 65.1% of connections]



Category:

Fig 32– Process Overview (Category, General, case frequency)
[83.7% of activities and 66.9% of connections]

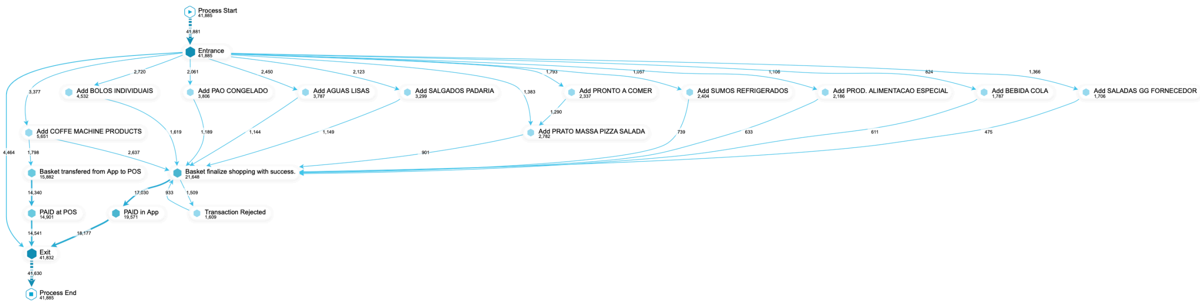
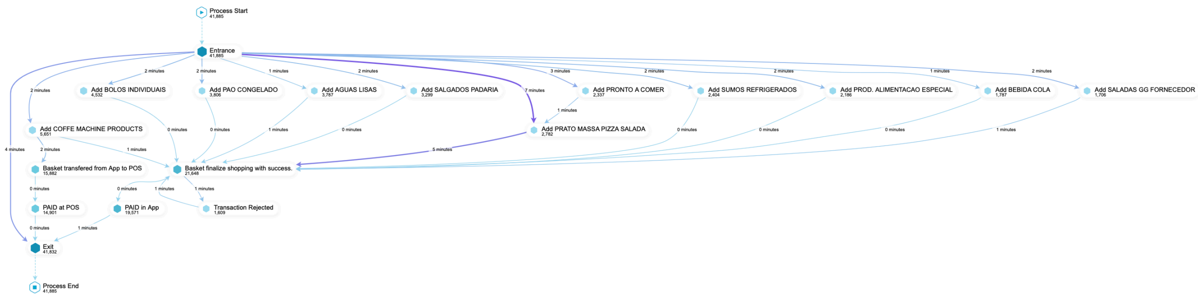


Fig 33– Process Overview (Category, General, throughput Time (AVG) in min)
[83.7% of activities and 66.9% of connections]



Store Zone:

Fig 34– Process Overview (Store zone, General, case frequency)
[91.1% of activities and 70.9% of connections]

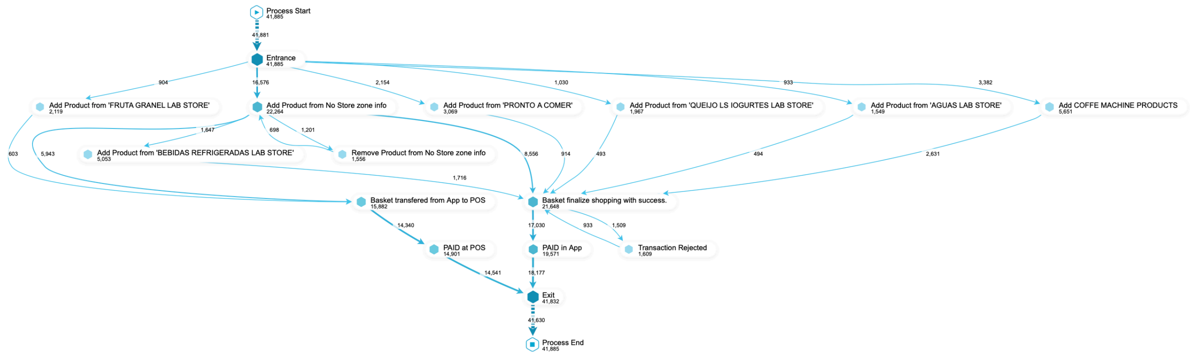
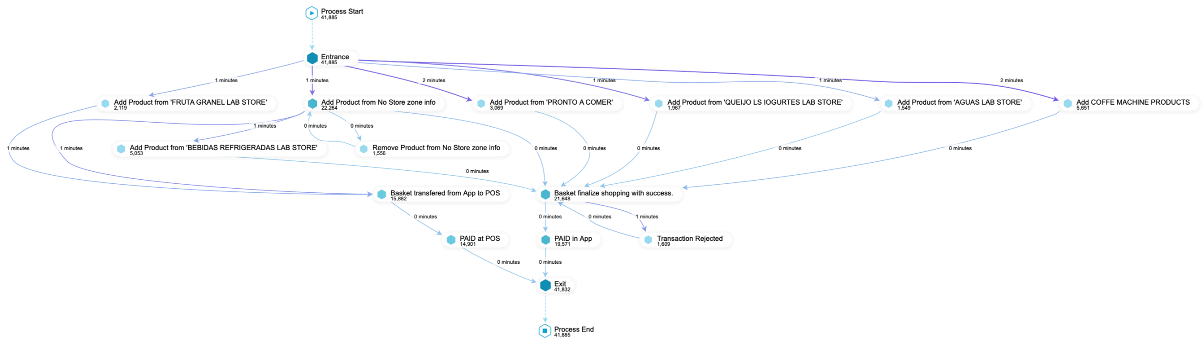


Fig 35– Process Overview (Store zone, General, throughput Time (AVG) in min)
[91.1% of activities and 70.9% of connections]



MORNING

Division:

Fig 36– Process Overview (Division of products,Morning, case frequency)
[93.1% of activities and 71.1% of connections]

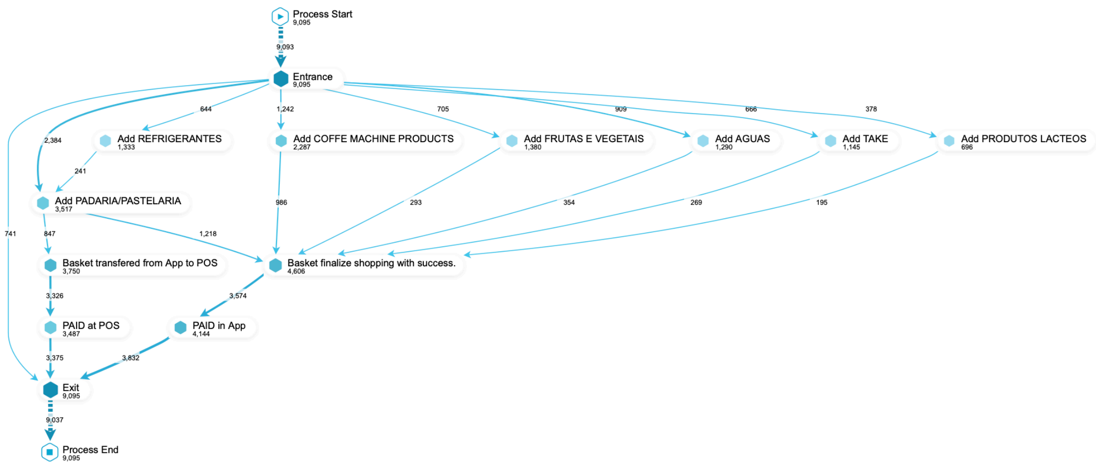
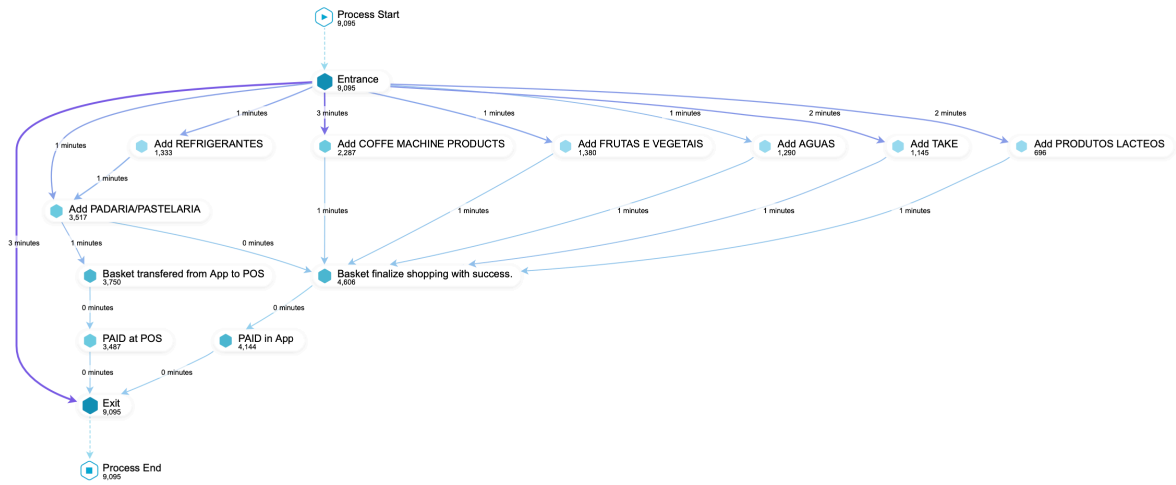


Fig 37– Process Overview (Division of products, Morning, throughput Time (AVG) in min)
[93.1% of activit and 71.1% of connections]



Product:

Fig 38– Process Overview (Products' names, Morning, case frequency)
[76.7% of activities and 63.3% of connections]

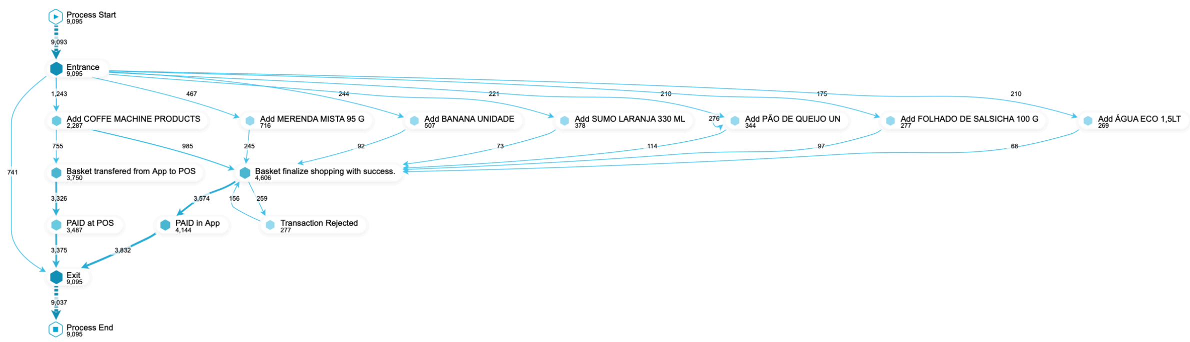
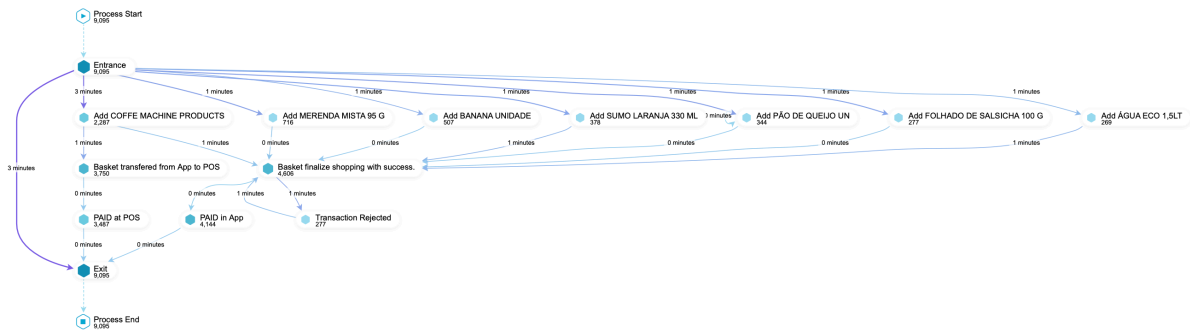


Fig 39– Process Overview (Products' names, Morning, throughput Time (AVG) in min)
[76.7% of activities and 63.3% of connections]



LUNCH TIME

<u>Division:</u>

Fig 40– Process Overview (Division of products, Lunch Time, case frequency)
[97.9 % of activities and 76.8% of connections]

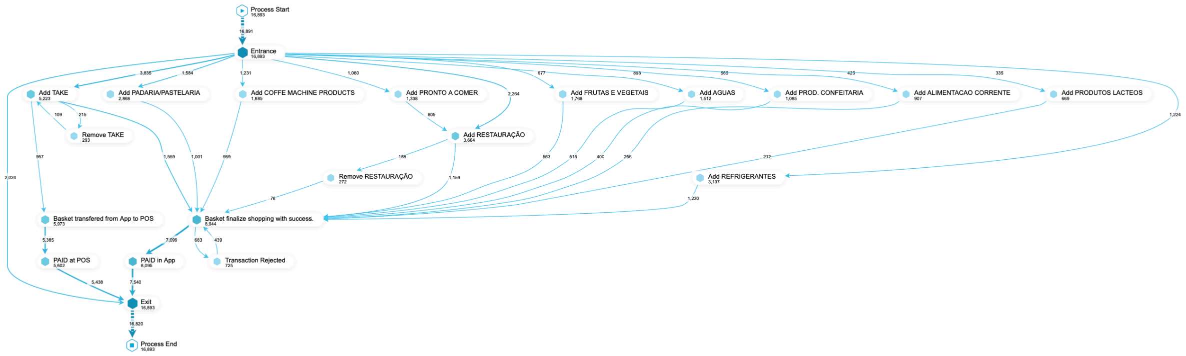
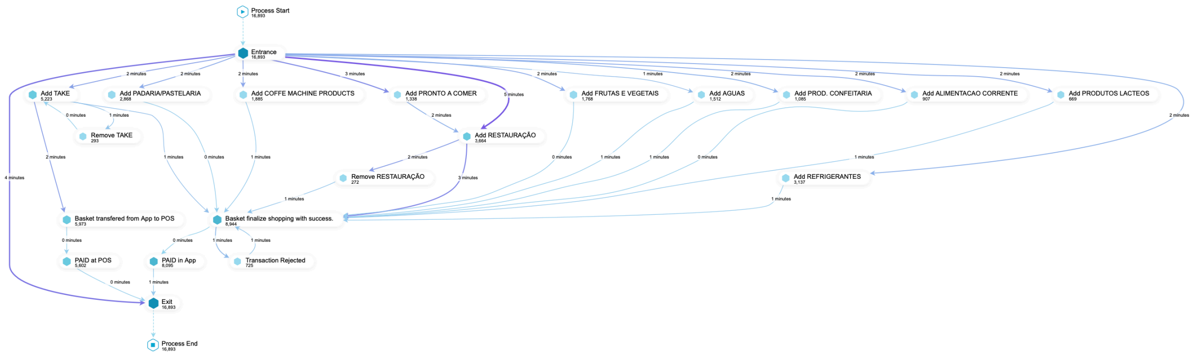


Fig 41– Process Overview (Division of products, Lunch Time, throughput Time (AVG) in min)
[97.9 % of activities and 76.8% of connections]



<u>Product:</u>

Fig 42– Process Overview (Products' names, Lunch Time, case frequency)
[73.4% of activities and 60.5% of connections]

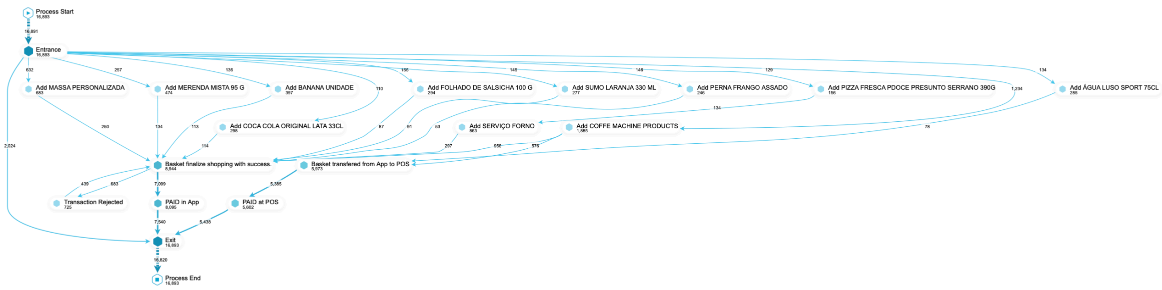
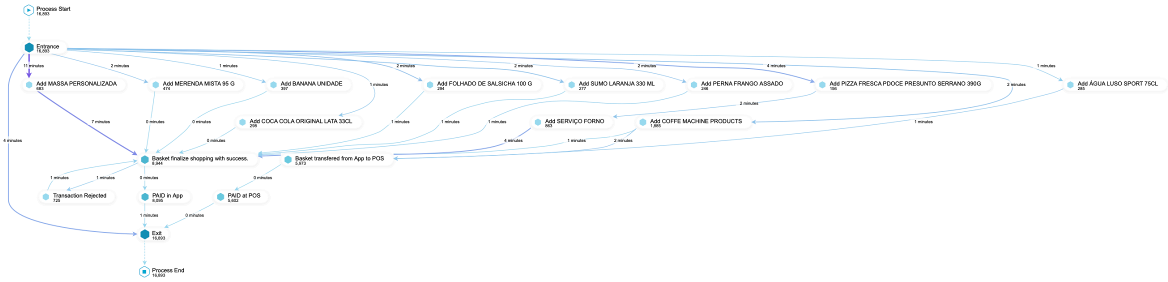


Fig 43– Process Overview (Products' names, Lunch Time, throughput Time (AVG) in min) *[73.4% of activities and 60.5% of connections]*



AFTERNOON
<u>Division:</u>

Fig 44– Process Overview (Division of products, Afternoon, case frequency) *[88.5% of activities and 69.7% of connections]*

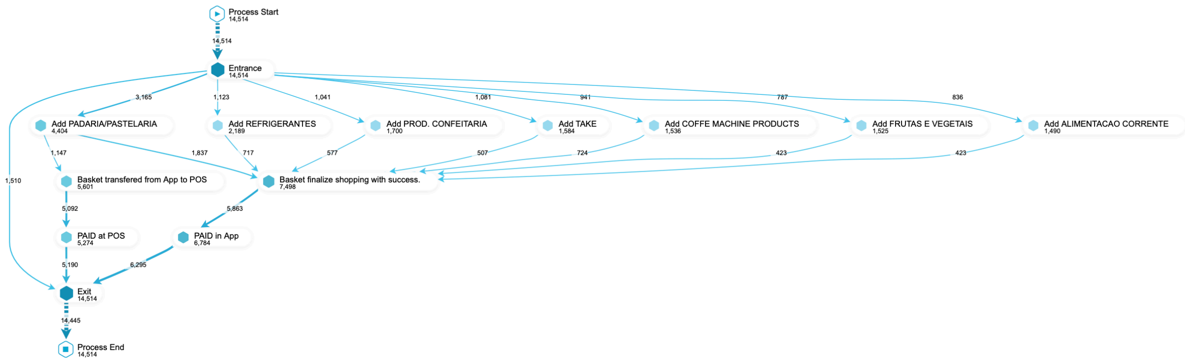
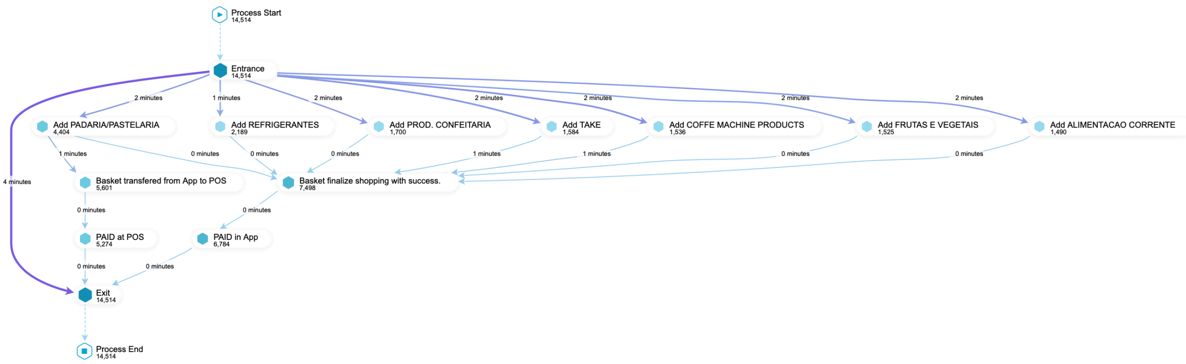


Fig 45– Process Overview (Division of products, Afternoon, throughput Time (AVG) in min) *[88.5% of activities and 69.7% of connections]*



Product:

Fig 46– Process Overview (Products' names, Afternoon, case frequency) *[72.7% of activities and 60.4% of connections]*

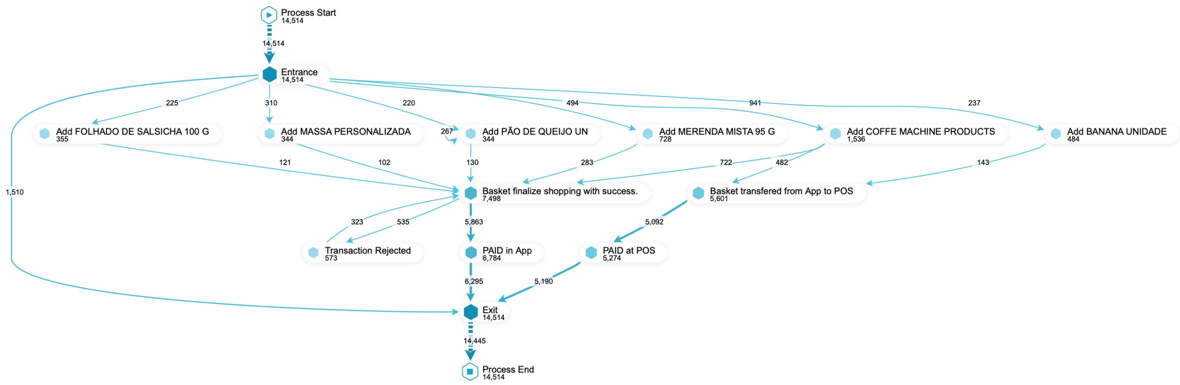
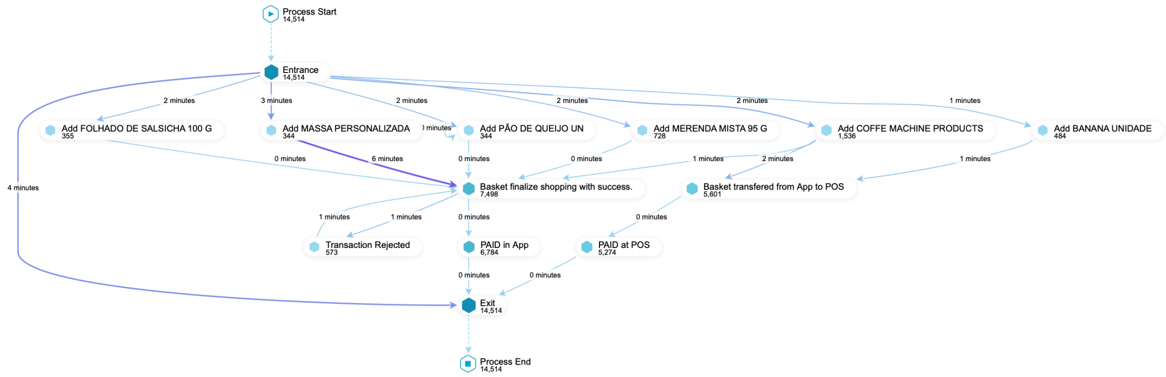


Fig 47– Process Overview (Products' names, Afternoon, throughput Time (AVG) in min) *[72. 7% of activities and 60.4% of connections]*



EVENING
<u>Division:</u>

Fig 48– Process Overview (Division of products, Evening, case frequency)
[94.1% of activities and 71% of connections]

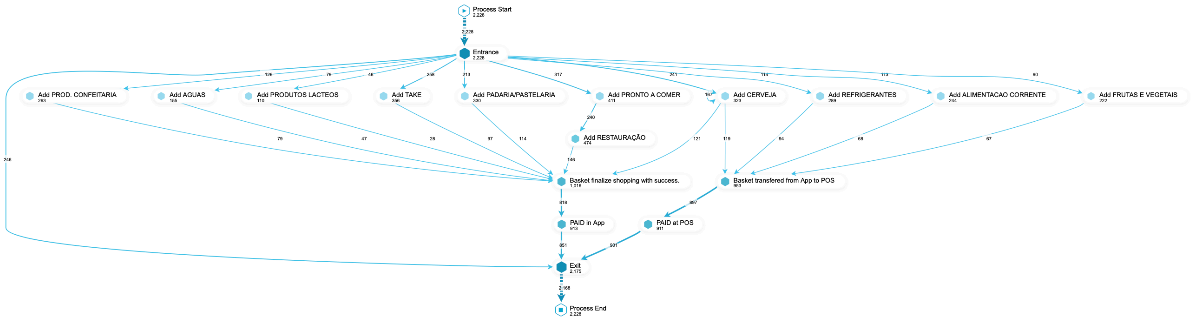
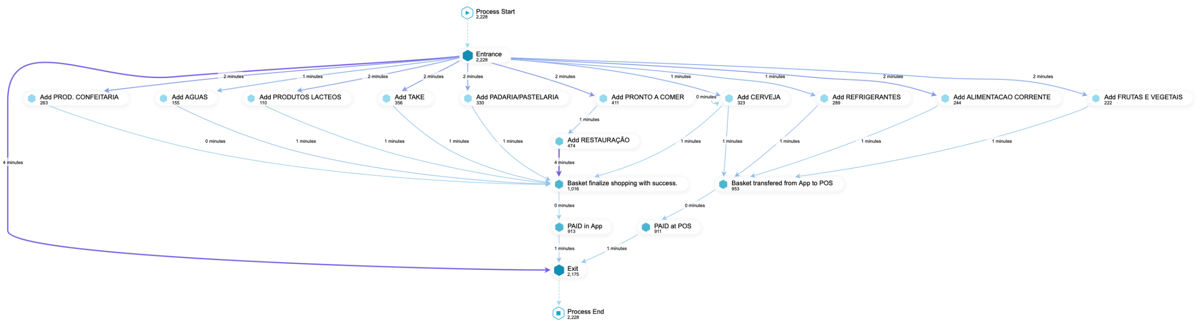


Fig 49– Process Overview (Division of products, Evening, throughput Time (AVG) in min)
[94.1% of activities and 71% of connections]



<u>Product:</u>

Fig 50– Process Overview (Products' names, Evening, case frequency)
[69.4% of activities and 57.1% of connections]

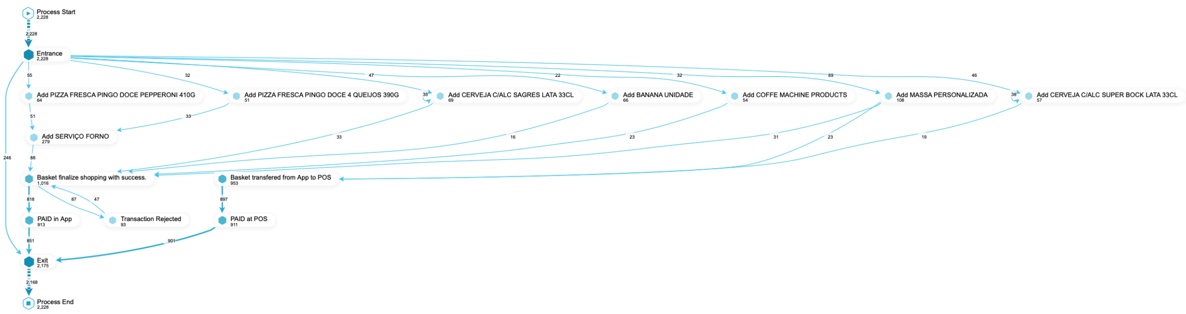


Fig 51– Process Overview (Products' names, Evening, throughput Time (AVG) in min)
[69.4% of activities and 57.1% of connections]

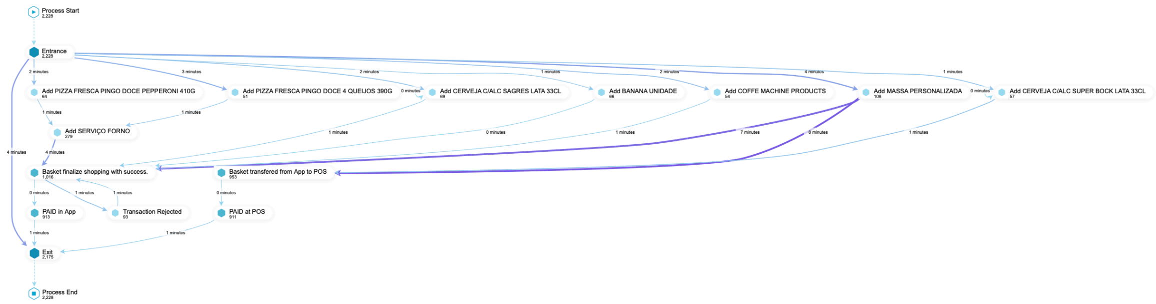


Fig 52– Process Overview (Products' names, cases with massa personalizada, case frequency)
[90.4% of activities and 80.6% of connections]

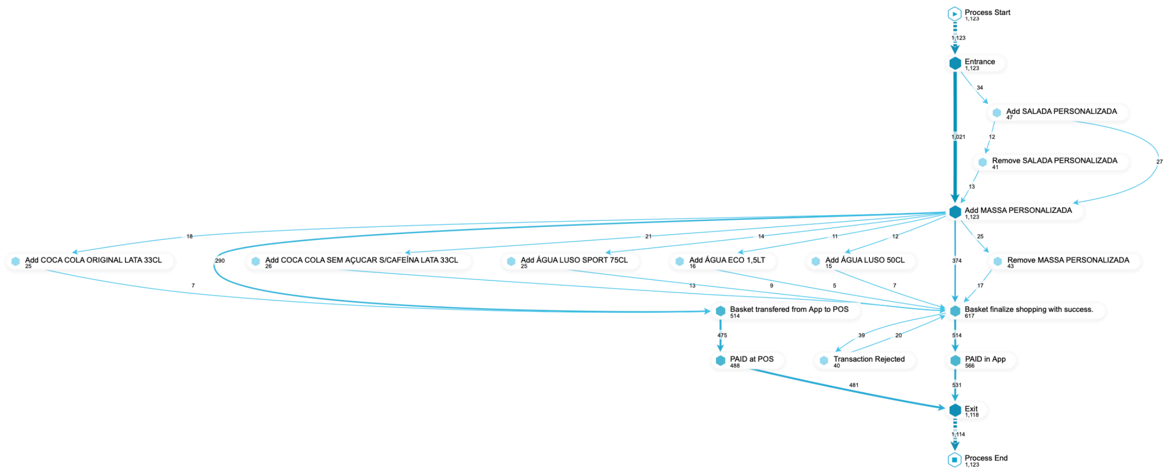


Fig 53– Process Overview (Products' names, cases with massa personalizada, throughput Time (AVG) in min)
[90.4% of activities and 80.6% of connections]

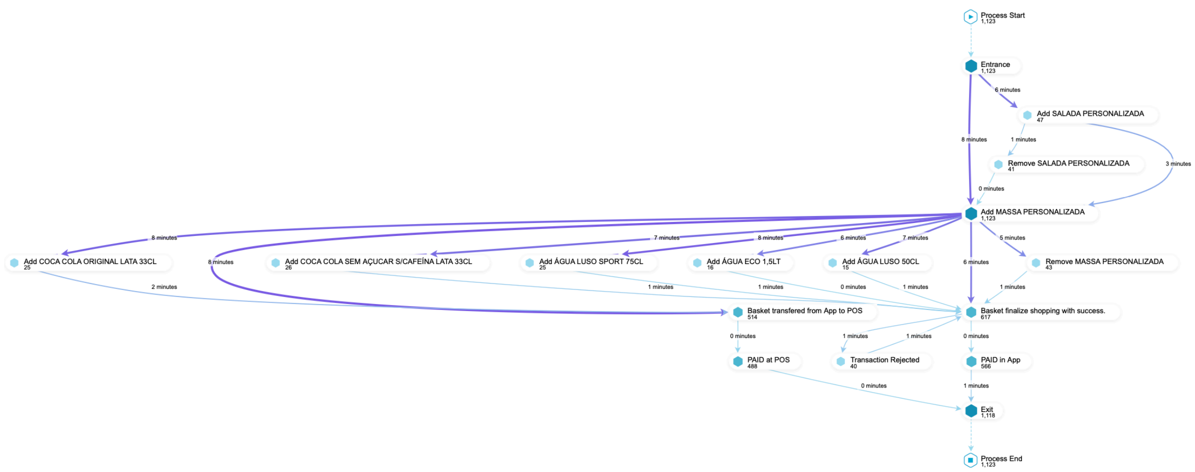


Fig 54– Process Overview (Products' names, cases with serviço forno, case frequency *[90.3% of activities and 77.1% of connections]*

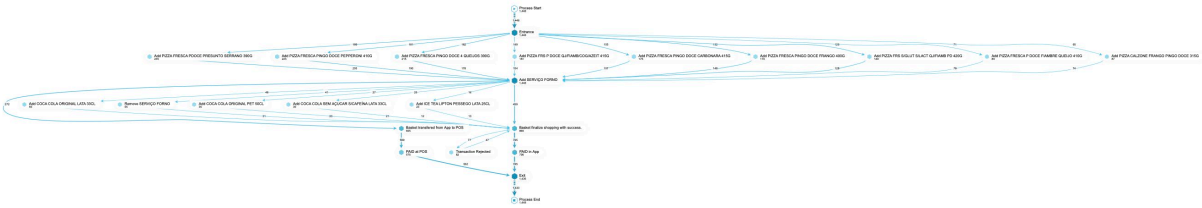


Fig 55– Process Overview (Products' names, cases with serviço, throughput Time (AVG) in min)
[90.3% of activities and 77.1% of connections]

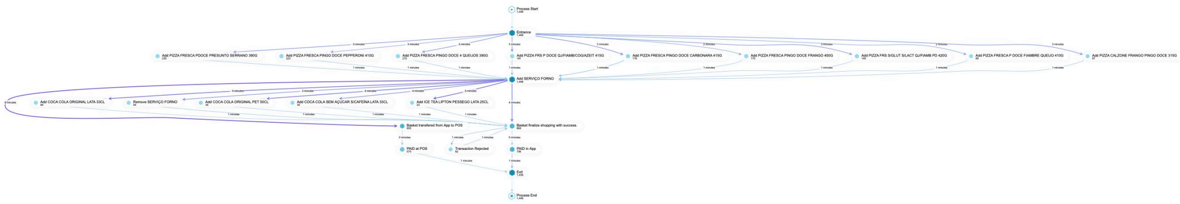


Fig 56– Process Overview (Products' names, cases with coffee machine products, case frequency)
[89% of activities and 77.3% of connections]

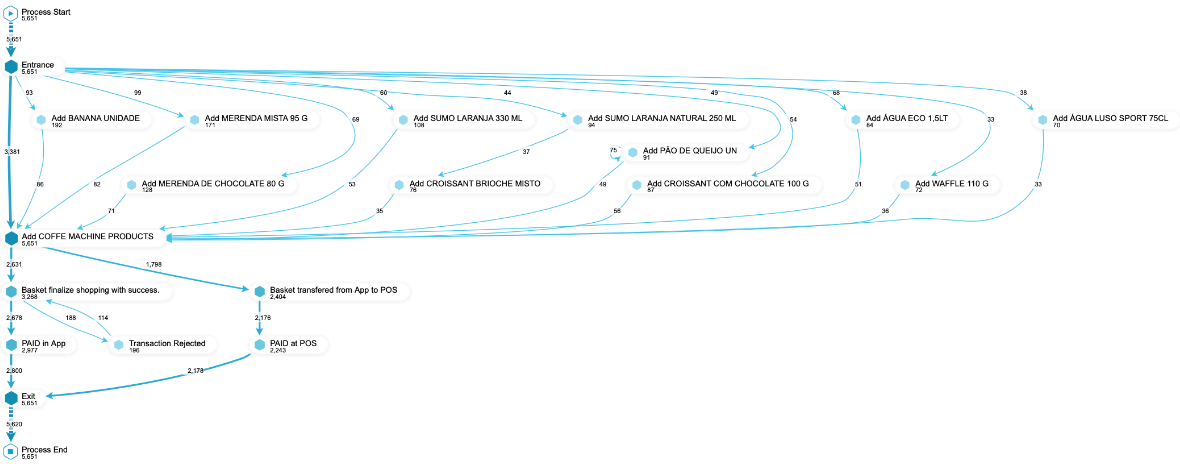
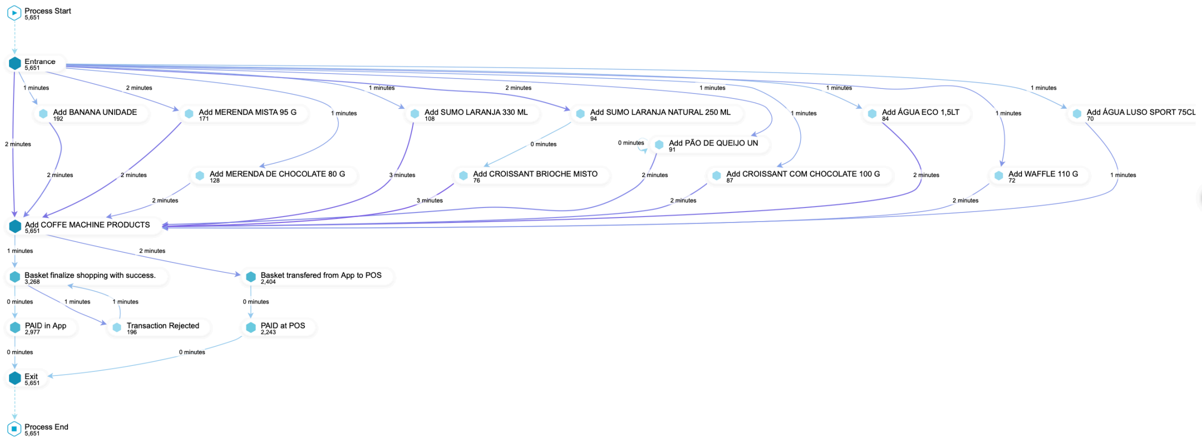


Fig 57– Process Overview (Products' names, cases with cofffe machine products, throughput Time (AVG) in min)
[89.2% of activities and 77.5% of connections]



DASHBOARDS FIGURES

Fig 58– General Dashboard

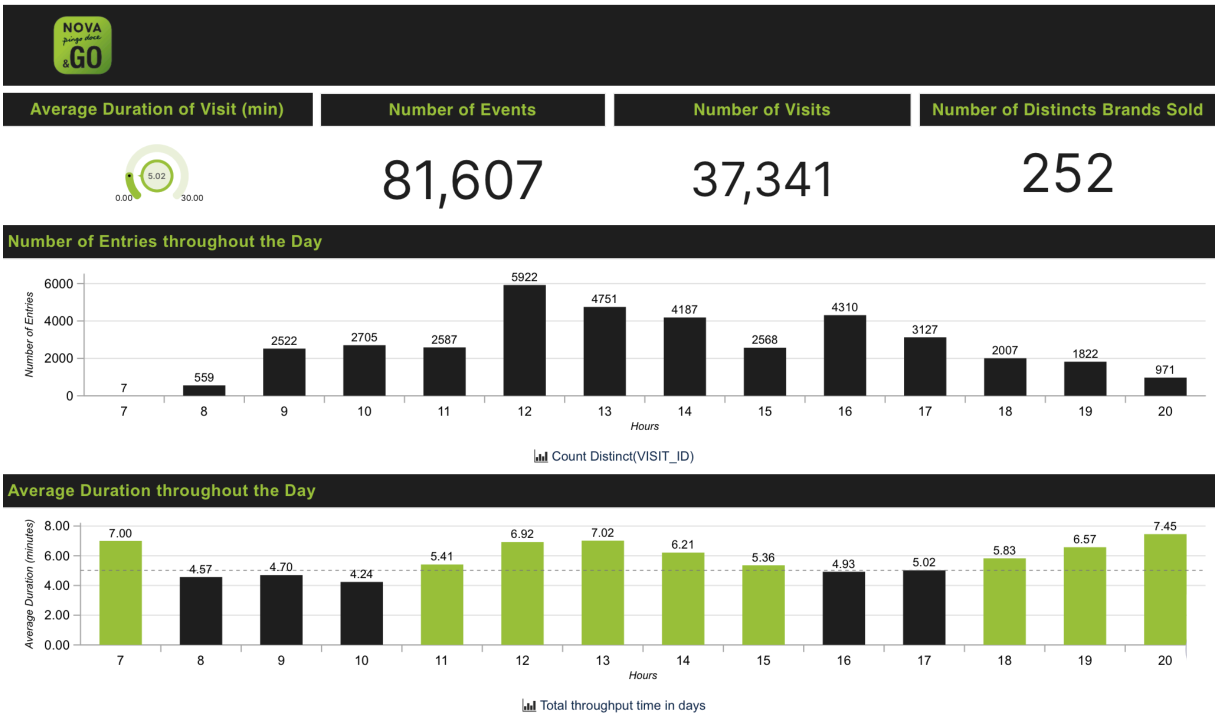
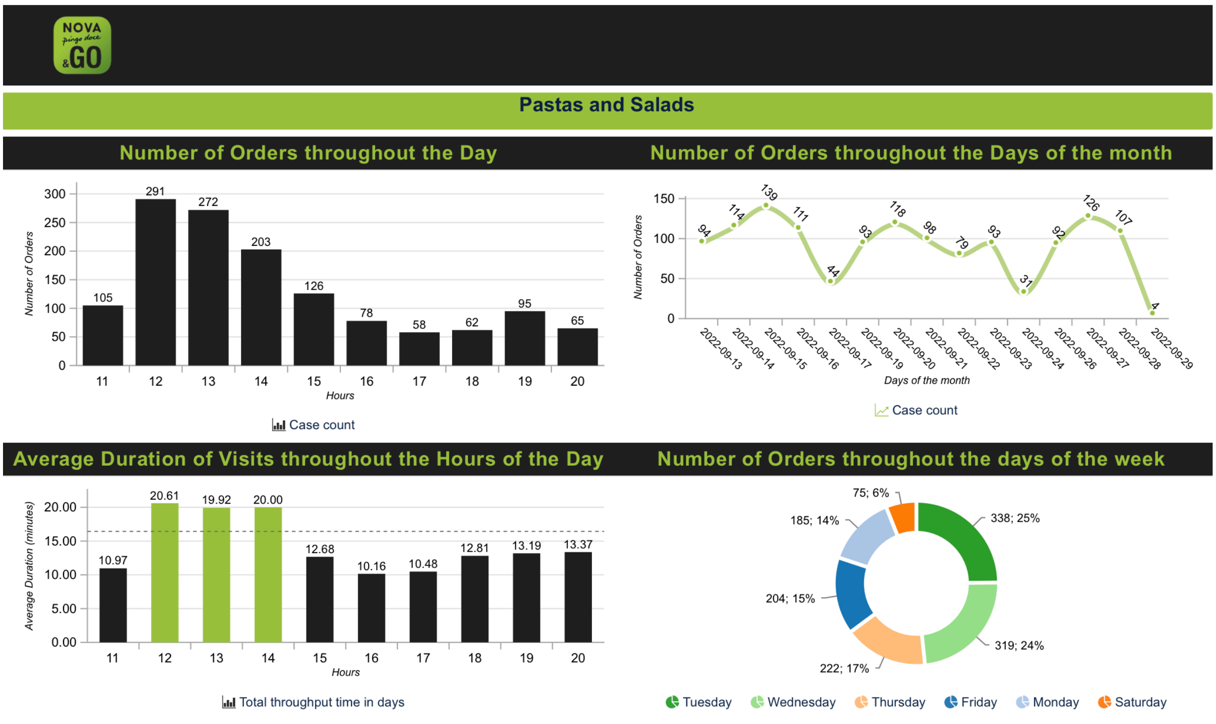


Fig 59– Pastas and Salads Dashboard



This visualization complements the information regarding the waiting time to order a pasta or a salad *(Fig 10 and 11 in the appendix)*. In this specific case, once again, we were able to identify that the durations of visits when the person asks for one of these two products were much higher than the average duration of a regular visit. Based on that, making a portion of pasta or salad can be considered a bottleneck in the process. From this specific dashboard, we can also conclude that the lunch hours, when people buy more pasta and salads, are also the hours when visits are more extended (between 12:00 and 14:00). This can be due to the size of the queue, which therefore requires a longer waiting time.

Fig 60– Group Dashboard (Division)- 1st Part

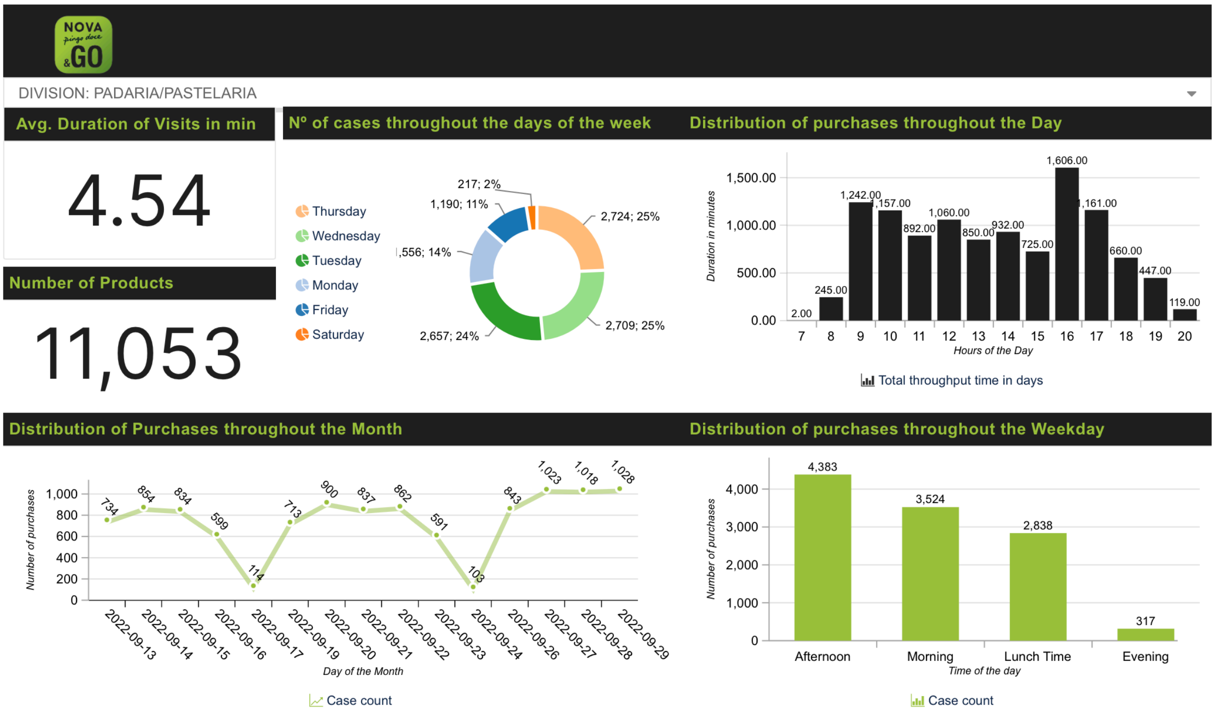


Fig 61– Group Dashboard (Division)- 2nd Part

DIVISION: AGUAS

Event_log_visits_test_profess...	Number of Products	Avg. Duration of Visits
ÁGUA ECO 1,5LT	737	4.95
ÁGUA LUSO SPORT 75CL	704	4.76
ÁGUA LUSO 50CL	591	4.50
ÁGUA MONCHIQUE 1,5LT	416	4.71
ÁGUA MONCHIQUE PET 100%...	390	4.11
ÁGUA ALCALINA PINGO DOC...	360	4.78
ÁGUA LUSO 1,5LT	317	4.94
GARRAFÃO REUTILIZÁVEL E...	169	6.32
ÁGUA EVIAN 75CL	123	4.71
ÁGUA C/GÁS SAN PELLEGRI...	110	5.30
AGUA PINGO DOCE 1,5LT	87	3.60
ÁGUA C/GÁS PEDRAS SALG...	66	5.62
ÁGUA ECO 6LT	53	6.00
ÁGUA ECO 3LT	48	6.27
LUSO DE FRUTA FRUTOS VE...	37	5.69
TORNEIRA DISPENSADORA E...	25	9.28
PLENO LIGHT CHÁ VERDE/LI...	23	4.50
GARRAFÃO REUTILIZÁVEL E...	22	6.33
ÁGUA C/GÁS PEDRAS LIMÃO...	16	2.94
GARRAFÃO REUTILIZÁVEL E...	10	5.95
ÁGUA C/GÁS PEDRAS SALG...	9	7.23
TISANA ROMÃ & CAMOMILA ...	6	5.67

Number of Distincts Brands

8

Number of Visits

4,972

Average Duration of Visit

Fig 62– Top 25 rules in terms of confidence- Parallel Coordinates Plot

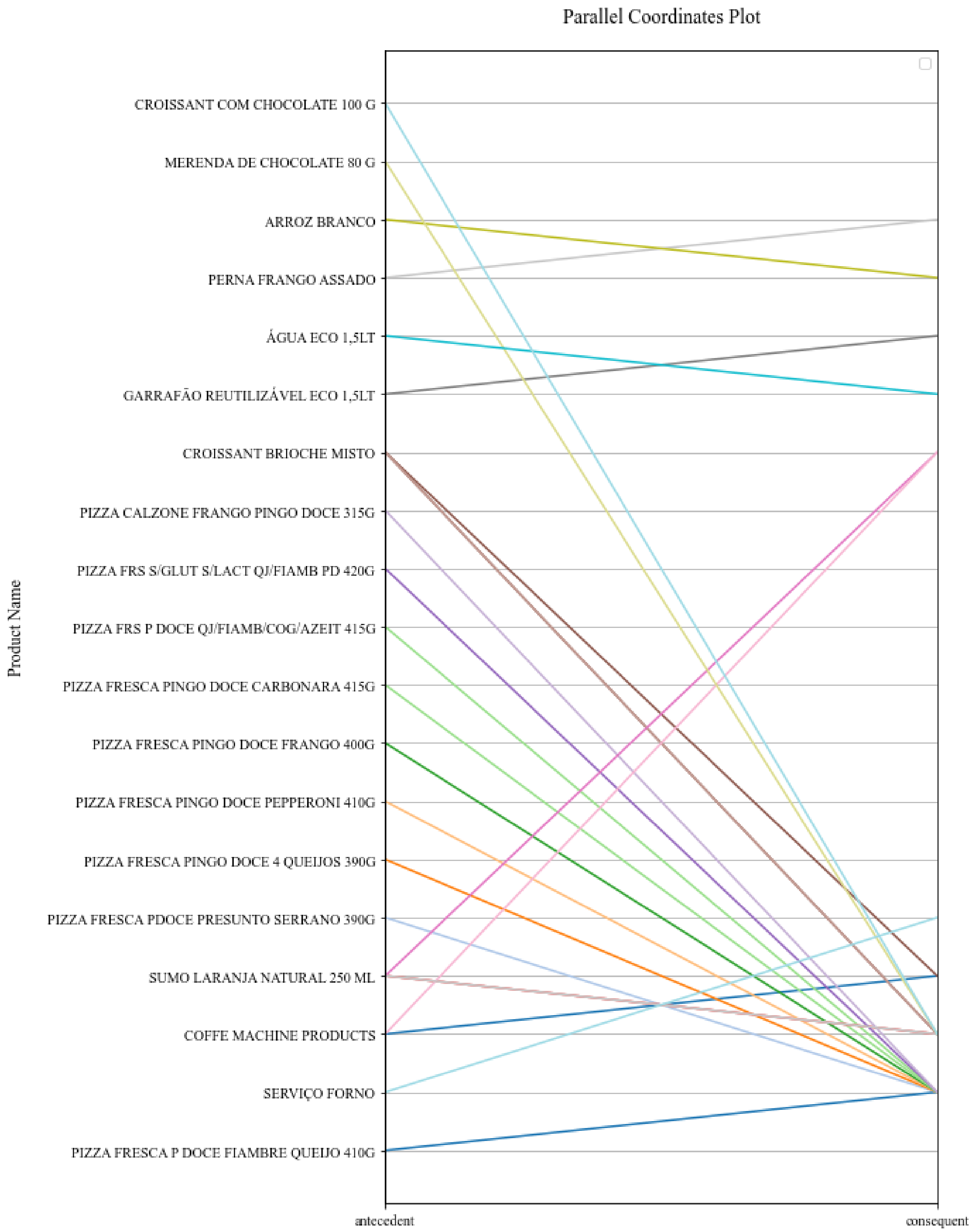


Fig 63– Top 25 rules in terms of confidence- Item Support Matrix

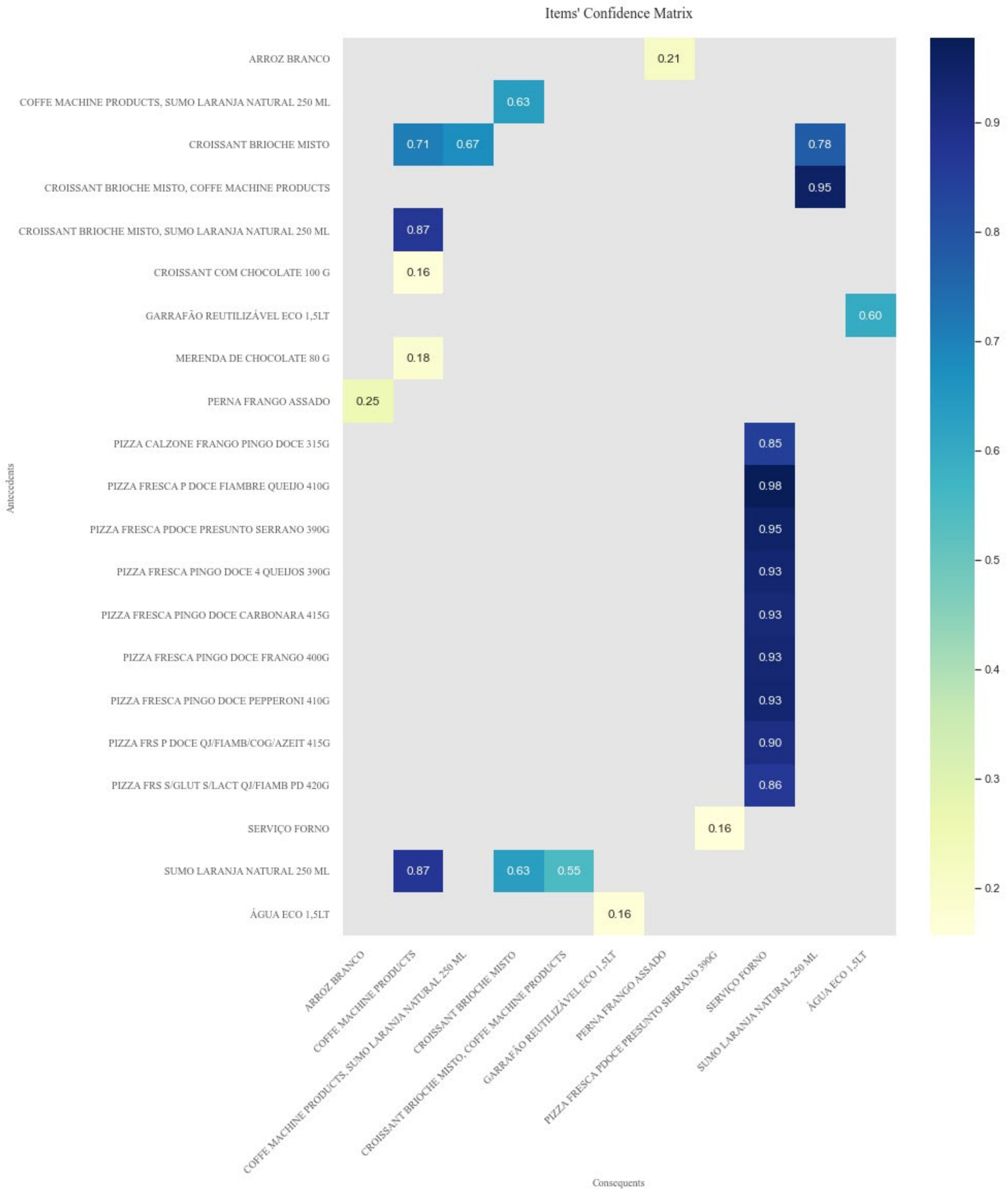


Fig 64– Process Overview (Cases that go to “Serviço forno”, case frequency)

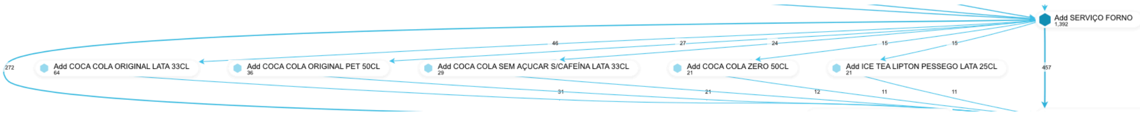


Fig 65– Top Rules (when antecedent is “Serviço Forno)

	antecedents	consequents	antecedent support	consequent support	support	confidence
38	(SERVIÇO FORNO)	(COCA COLA ORIGINAL LATA 33CL)	0.037389	0.013034	0.001751	0.046826
41	(SERVIÇO FORNO)	(COCA COLA ORIGINAL PET 50CL)	0.037389	0.007295	0.000856	0.022893
47	(SERVIÇO FORNO)	(COCA COLA SEM AÇUCAR S/CAFEÍNA LATA 33CL)	0.037389	0.008793	0.000681	0.018210
106	(SERVIÇO FORNO)	(ICED TEA MANGA PD 0,5LT)	0.037389	0.004786	0.000681	0.018210
105	(SERVIÇO FORNO)	(ICE TEA LIPTON PESSEGO LATA 25CL)	0.037389	0.005525	0.000642	0.017170
49	(SERVIÇO FORNO)	(COCA COLA ZERO 50CL)	0.037389	0.008093	0.000564	0.015088
108	(SERVIÇO FORNO)	(ICED TEA PESSEGO PD 0,5LT)	0.037389	0.005389	0.000545	0.014568

Appendix – Celonis Guide

<u>Guide for Importing and selecting variables</u>

The first step was to download the necessary files. Celonis automatically recognizes the columns that are present in the files and since the files have headers, these are used to identify the columns. Following the import of the data, the platform requests categorization of event logs based on Case ID, Timestamp, and Activity.

Fig 66– Logic for assigning "Case ID" in Celonis

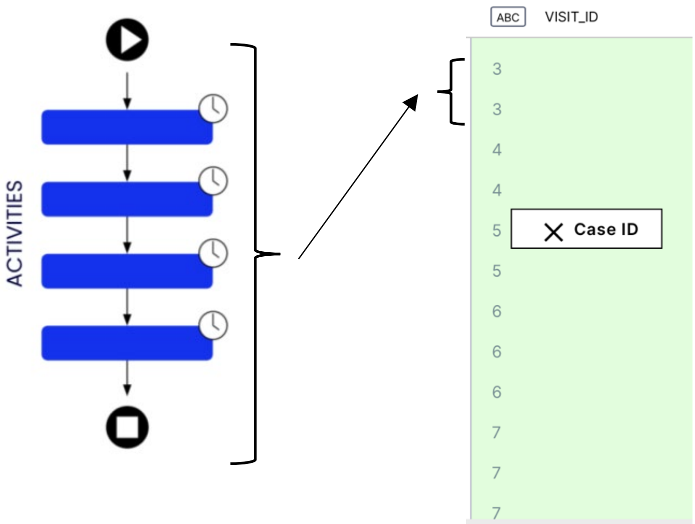
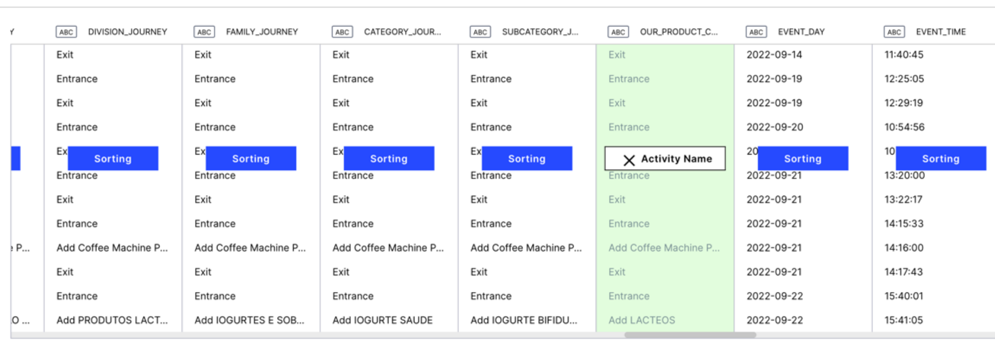
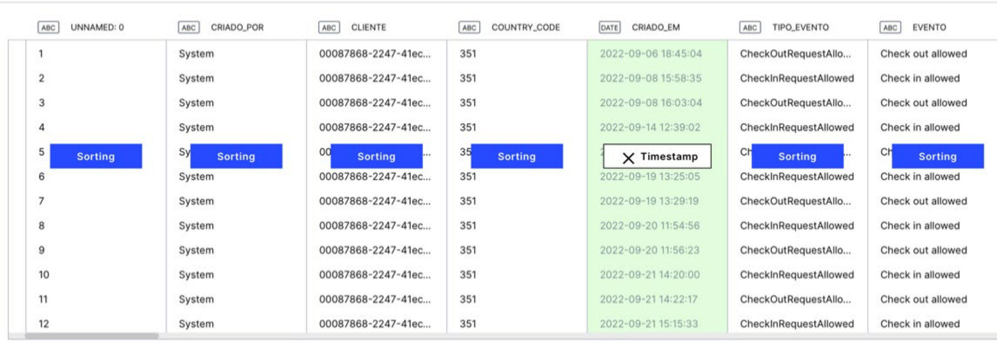


Fig 67– Assignment of the "Activity Name"



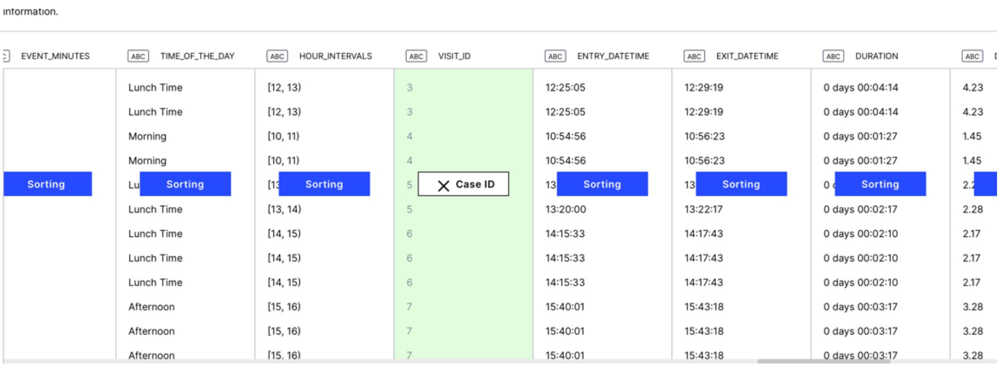
Y	DIVISION_JOURNEY	FAMILY_JOURNEY	CATEGORY_JOUR...	SUBCATEGORY_J...	OUR_PRODUCT_C...	EVENT_DAY	EVENT_TIME
	Exit	Exit	Exit	Exit	Exit	2022-09-14	11:40:45
	Entrance	Entrance	Entrance	Entrance	Entrance	2022-09-19	12:25:05
	Exit	Exit	Exit	Exit	Exit	2022-09-19	12:29:19
	Entrance	Entrance	Entrance	Entrance	Entrance	2022-09-20	10:54:56
	Ex Sorting	Ex Sorting	Ex Sorting	Ex Sorting	Activity Name	20 Sorting	10 Sorting
	Entrance	Entrance	Entrance	Entrance	Entrance	2022-09-21	13:20:00
	Exit	Exit	Exit	Exit	Exit	2022-09-21	13:22:17
	Entrance	Entrance	Entrance	Entrance	Entrance	2022-09-21	14:15:33
P...	Add Coffee Machine P...	Add Coffee Machine P...	Add Coffee Machine P...	Add Coffee Machine P...	Add Coffee Machine P...	2022-09-21	14:16:00
	Exit	Exit	Exit	Exit	Exit	2022-09-21	14:17:43
	Entrance	Entrance	Entrance	Entrance	Entrance	2022-09-22	15:40:01
O ...	Add PRODUTOS LACT...	Add IOGURTES E SOB...	Add IOGURTE SAUDE	Add IOGURTE BIFIDU...	Add LACTEOS	2022-09-22	15:41:05

Fig 68– Assignment of the "Timestamp"



UNNAMED: 0	CRIADO_POR	CLIENTE	COUNTRY_CODE	CRIADO_EM	TIPO_EVENTO	EVENTO
1	System	00087868-2247-41ec...	351	2022-09-06 18:45:04	CheckOutRequestAllo...	Check out allowed
2	System	00087868-2247-41ec...	351	2022-09-08 15:58:35	CheckInRequestAllowed	Check in allowed
3	System	00087868-2247-41ec...	351	2022-09-08 16:03:04	CheckOutRequestAllo...	Check out allowed
4	System	00087868-2247-41ec...	351	2022-09-14 12:39:02	CheckInRequestAllowed	Check in allowed
5 Sorting	Sy Sorting	00 Sorting	35 Sorting	Timestamp	Ch Sorting ...	Ch Sorting
6	System	00087868-2247-41ec...	351	2022-09-19 13:25:05	CheckInRequestAllowed	Check in allowed
7	System	00087868-2247-41ec...	351	2022-09-19 13:29:19	CheckOutRequestAllo...	Check out allowed
8	System	00087868-2247-41ec...	351	2022-09-20 11:54:56	CheckInRequestAllowed	Check in allowed
9	System	00087868-2247-41ec...	351	2022-09-20 11:56:23	CheckOutRequestAllo...	Check out allowed
10	System	00087868-2247-41ec...	351	2022-09-21 14:20:00	CheckInRequestAllowed	Check in allowed
11	System	00087868-2247-41ec...	351	2022-09-21 14:22:17	CheckOutRequestAllo...	Check out allowed
12	System	00087868-2247-41ec...	351	2022-09-21 15:15:33	CheckInRequestAllowed	Check in allowed

Fig 69– Assignment of the "Case ID"



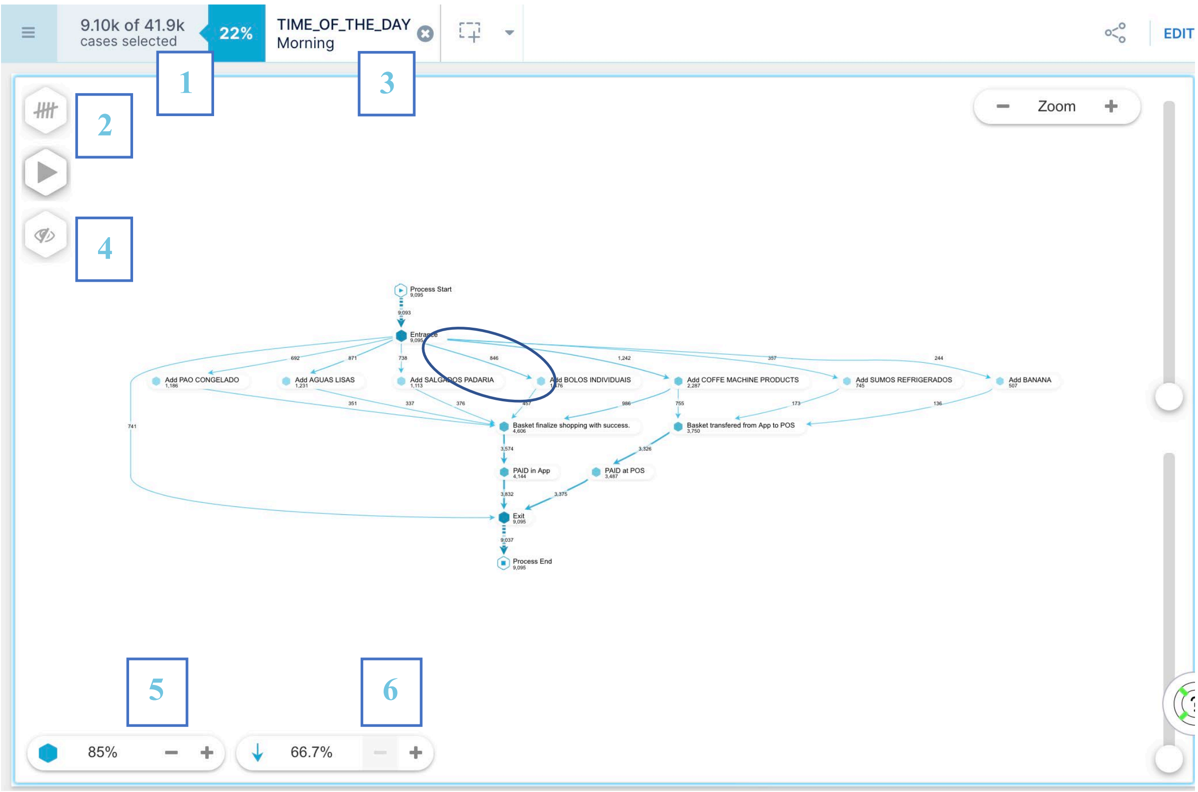
information.

EVENT_MINUTES	TIME_OF_THE_DAY	HOUR_INTERVALS	VISIT_ID	ENTRY_DATETIME	EXIT_DATETIME	DURATION	
	Lunch Time	[12, 13)	3	12:25:05	12:29:19	0 days 00:04:14	4.23
	Lunch Time	[12, 13)	3	12:25:05	12:29:19	0 days 00:04:14	4.23
	Morning	[10, 11)	4	10:54:56	10:56:23	0 days 00:01:27	1.45
	Morning	[10, 11)	4	10:54:56	10:56:23	0 days 00:01:27	1.45
Sorting	Lu Sorting	[13 Sorting	5 Case ID	13 Sorting	13 Sorting	0 Sorting	2.2
	Lunch Time	[13, 14)	5	13:20:00	13:22:17	0 days 00:02:17	2.28
	Lunch Time	[14, 15)	6	14:15:33	14:17:43	0 days 00:02:10	2.17
	Lunch Time	[14, 15)	6	14:15:33	14:17:43	0 days 00:02:10	2.17
	Lunch Time	[14, 15)	6	14:15:33	14:17:43	0 days 00:02:10	2.17
	Afternoon	[15, 16)	7	15:40:01	15:43:18	0 days 00:03:17	3.28
	Afternoon	[15, 16)	7	15:40:01	15:43:18	0 days 00:03:17	3.28
	Afternoon	[15, 16)	7	15:40:01	15:43:18	0 days 00:03:17	3.28

<u>Guide for Process Explorer Visualization</u>

In these graphs, all the arrows represent a link between two consecutive events. Next to each arrow is a number representing the value of the KPI chosen to analyze the link, in this case the frequency of the link. As an example, in the case below it can be concluded that there were 846 visits in which customers entered the store and added "bolos individuais". Nevertheless, it is important to emphasize that other KPIs can be selected, such as process time (AVG). As a result, we will explain them in greater detail later (in point 2). As shown in the figure below, this tool provides a variety of resources. Following that, we will discuss each one in greater detail.

Fig 70– Process Explorer logic in Celonis



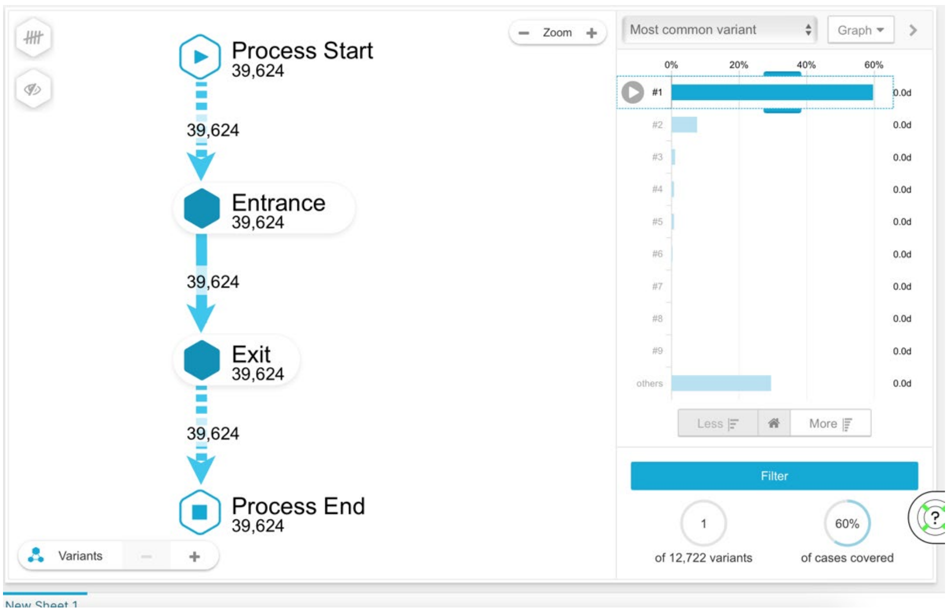
1. Presents the number of cases displayed in the process model. Adding a filter here will lead to a selection of the number of cases that will be displayed.
2. Allows the choice of KPIs that will be presented in the process model.
    - ***Options:*** 'Case frequency', 'Activity frequency', 'Processing time (median)', 'Processing time (AVG)' and 'Processing time (trimmed mean)'. → Throughout the thesis, we used two KPIs: **case frequency** and **process time (AVG).** When using the KPI case frequency, the blue lines indicate the number of times an antecedent activity was followed by a consequent activity. When selecting the KPI process time (average), the lines in blue represent the average duration from the moment people take the first action (antecedent) to the moment they take the second action (consequent).

3. Adds filters to the analysis.
    - ***Options:*** Attribute selection, Activity selection, Process flow selection, Through time selection, Rework Selection, Crop Selection
4. Allows the displaying or hiding of wanted/unwanted activities in the process model.
5. Configure how many activities are shown. In cases of 100% all activities are shown.
6. An increase in the number of connections leads to more edges being displayed in the Process Map (however, no new activities are added).

<u>Guide for Variant Explorer Analysis</u>

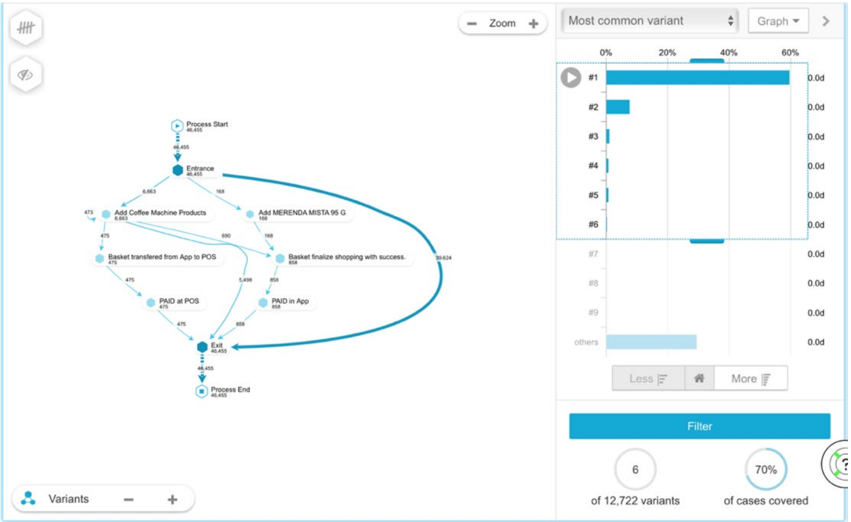
In most cases, the process takes place in different ways, which is why Celonis offers Variant Explorer analysis. In addition to showing the sequence of activities, Variant Explorer allows us to see how many cases (in this case, purchase orders) follow the same path. The figure below shows the most common variant, which is the path that happens the most.

Fig 71– Logic Variant Explorer in Celonis – Part 1



By increasing the number of variants shown, we could perceive the complexity and variability of paths within the same store. In the figure below, we can already see that not all cases go through the same activities.

Fig 72– Logic Variant Explorer in Celonis – Part 2



Selecting the total activities and connecting them would result in a spaghetti-like graph. A graph such as this can only be used to illustrate the complexity of the process, so it justifies the need to conduct a deeper analysis of the process.

Fig 73– Logic Variant Explorer in Celonis – Part 3