NOVA IMS Information Management School

MEGI

Mestrado em Estatística e Gestão de Informação

Master Program in Statistics and Information Management

Determine the potential and the extent to which geographic socio-demographic data impacts retail performance revenue and consumer behavior and determine how much discounts impact revenue.

Pedro Guilherme Ribeiro Veiga

Dissertation presented as partial requirement for the Master's Degree in Statistics and Information Management

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

DETERMINE THE POTENTIAL AND THE EXTENT TO WHICH GEOGRAPHIC SOCIO-DEMOGRAPHIC DATA IMPACTS RETAIL PERFORMANCE REVENUE AND CONSUMER BEHAVIOR

DETERMINE HOW MUCH DISCOUNTS IMPACT REVENUE.

by

Pedro Guilherme Ribeiro Veiga

Dissertation presented as partial requirement for the Master's Degree in Information Management, with a specialization in Knowledge Management

Advisor: Prof. Doutor Rui Gonçalves

July 2021

FOR MY BELOVED PARENTS AND THOSE WHO SUPPORTED ME IN MAKING THIS THESIS A REALITY, A HUGE THANK YOU

ACKNOWLEDGEMENTS

FOR THOSE WHO MADE THIS THESIS POSSIBLE AND A REALITY:

PROFESSOR DOUTOR RUI GONÇALVES

AND

FERNANDO RODRIGUES

AND

JOSÉ ALBUQUERQUE

AND

JOÃO GOMES PORTO

AND

SALWA KELIDAR

ABSTRACT

The objective of this thesis is to describe and analyze the sales and revenue of a food distribution company, Tasty and Sweet, by product or product type and relate it to geographic socio-demographic data provided by National Bureau of Statistics (INE). Tasty and Sweet operates and covers the entire Portuguese national territory, distributing and selling their products to any retailer company who is willing to resell them. The goal of this work is to develop an analytical model that allows the tracking of all the sales and revenue, by group item or by item, relating them, eventually, to the social and demographic characterization of a specific geography so that patterns may be (or not) identified. Another goal of the thesis is to determine the extent to which sales promotions have an impact on retailer sales. In order to achieve these objectives several methods of data analysis will be developed, supported and backed up by software from SAS Institute (Sas Guide and Sas Miner).

It was possible to come to the following conclusions: firstly there are relevant socio-demographic variables that impact, or are more related, with de retailer's revenue like: Indicators of enterprises by municipality N. º/km2 - enterprise density, 2016; Territorial structure by municipality - weight of resident population, 2011; Distribution of declared gross income less individual tax income paid of tax households by municipality (less 5k), 2016 are in fact more impacting; Secondly revenue volume is in fact impacted by discounts and promotions.

KEYWORDS

Data Mining; Retail; Socio Demographic Variables

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

In today's modern society where technology is at a mature stage – ubiquitously present and used to support almost everything we do – most human interactions are recorded and stored in a database format. Moreover, the possibilities of analyzing and relating it with other sources of information are also at an advanced stage allowing us to develop data models with a high level of complexity. In addition, in the last couple of years, we have witnessed an exponential growth in the overall volume of information stored, to the extent that, every 20 months, this volume doubles in all companies. Indeed, Data Science has continued to make rapid advances, particularly on the frontiers of machine learning and deep learning. Today, organizations have high volumes of raw data, which combined with powerful and sophisticated analytical tools enable us to create and generate insights based on that very data that can improve operational performance and create new market opportunities.

Supported by data transformed into information, the decision-making process now is supported by relevant facts that comes from evidence, experiment and more accurate forecast rather then made based on gut or entrepreneurial instinct. Data contributes and is an extremely important fact of production, (Mc Kinsey 2016)". Data empowers and improves decision-making processes, so that the correct and accurate decision is made at the right time and based on real data facts.

Having huge volumes of historical data, companies that operate in the following industries: utilities, energy and insurance, just to name a few, and supported with staff with the right skills and capabilities, like Data Mining techniques, they have a huge strategical advantage that is being able to turn collected data into to producing information regularly regarding their business activities. These companies are driven to knowledge discovery from databases as a core activity (Dakic et al all 2017).

Data Mining has become extremely relevant area that support companies that are data driven and that recognize the importance and relevance of using data, it supports several industries such as: utilities, banking, insurance, retail just to name a few. The knowledge produced from the data mining process allows to, for example: increase revenue, fraud detection, categorize risk, increase customer knowledge, etc. Data mining can simple be described as the result of intersection between computer science and statistics domains. One of the possible outputs from the datamining, is that enables us to discover and extract knowledge from the available data and relate different datasets, supported by principles, tasks, and several techniques (ex: association rules, decision trees, probability networks, neural networks, classification, cluster analysis, prediction) that are made available to facilitate the decision-making process (Marques 2013).

Taking the above comments on the relevance of the use of Data Mining to analyze and relate different and various data sources, sales and socio-demographic data, this field of knowledge is of high relevance to the company Tasty and Sweet with potential significant outcomes as it will enable the company to understand which and how much different socio-demographic variables impact the company's revenue and performance. Regarding simple business analysis we will be able to determine the time frame when sales increase if discounts impact on sales volume and what kind of products are sold by geography. Data Mining will support key decision-making processes regarding sales, revenue expectations and stock distribution optimization – all supported by a data model driven by data facts that can be used in Portugal but can also be applied to any geographic location.

This thesis has several sections and is expected to develop an analytical model that allows the quantification of which variables from the dataset impact the target variable or revenue and to understand if promotions are related with revenue. To this end, the thesis is structured in the following way: firstly, the gathering of several data sources (Tasty and Sweet, CTT, INE) and their transformation into one; secondly, cleaning the dataset to one that has meaningful, non-redundant and relevant

variables; thirdly, analyzing the details from each chosen variable used by the model and analyzing them individually; and, finally, identifying which are the more impacting independent variable on the target variable. The last section is the conclusion which contains the major findings from the thesis, supported by the developed best analytic model.

2. LITERATURE REVIEW

This master's thesis aims to understand if and in what ways socio demographic data impacts the revenue of a retail company. To that end, it is necessary to review and reflect on a series of concepts and processes. For this purpose, a dataset from the stock management company will be accessed with sales and all financial movements for a specific time frame, from ordering to selling by geographic location. In this dissertation, several concepts, considered the relevant and impacting ones, will be discussed, and addressed. These include stock management, revenue analysis, socio-demographics, and psychographics (National Statistical Bureau) and regression analysis. These will be discussed in the pages below.

2.1 STOCK MANAGEMENT

Stock management and the process for managing it is key for any retail company. Indeed, one of the main problems for a retailer relates to how to manage stock properly, due to its high financial impact on company results. Product search and demand both have a huge relevance, we can consider as one of the most important and relevant to consider in stock management, predictions in this specific area should not be consider. Any disruption in the supply chain and in stock management, may result in significant negative impacts and therefore consequences for any retailer, as it stops revenue streams and that is crucial in a profit-oriented organization, also the huge competition that any retailer faces presently is compatible is stock disruption. In this regard, Vu considers that, an efficient inventory control enables that the businesses runs profitably and without selling disruption. Because retailers face an extremely competitive reality the profits are becoming narrow to the point that neither excessive in-stock nor depleted out-of- stock is negotiable. Vu pointed out from the research provided a comprehensive evaluation of substantial inventory management models which are widely used by retailers throughout history. It also considers that Big Data Analytics (BDA) have a huge impact on inventory control, the inventory control reflects that the central point for inventory optimization and management is minimising the sum of costs associated with holding it, stockless leads to sales lost, costs of back orders and salvage costs. Presently any retailer has an inventory management model ranging from Economic order quantity (EOQ) model, Order-up-to model, last-in-first-out (LIFO) system, multi-item retail inventory system to decision support system (DSS) (Vu 2018)".[1]

[1] For a comprehensive discussion of these issues refer to, inter alia, Gwynne Richards, *Warehouse Management: A Complete Guide to Improving Efficiency and Minimizing Costs in the Modern Warehouse*; see also David J. Piasecki, *Inventory Accuracy: People, Processes, & Technology*.

2.2 GEOGRAPHY

"Geomarketing is the cornerstone of successful businesses in this age of digitization and quickly changing markets (Rosu 2013)."

Specific data was added to the project with the purpose of "transforming" retailer point of sales into geographic data, so it could be related to other data sources – acting as a "bridge" between the retailer data and other external datasets. Any territorial widespread retailer must have a geolocation view on how the business develops, so it can identify good and bad performance areas so that actions are taken to improve the performance in the worst locations. Indeed, as noted by O'Kelly spatial interaction models consider the attraction of a specific facility is determined by the impact in terms of attraction to it and simultaneously by the consumer distance to alternative facilities. Huff (1962, 1963, 1964) and Lakshmanan and Hansen (1965) improved spatial models by developing specialized 'retail' variants, for example Huff added and introduced an extremely practical approach, where he related the attraction to a facility as the amount of floor space, opposite to using the population in the surrounding area as main input of previous used models. With this development the model was able to consider several variables (e.g., number of functions, parking capacity, etc.), so that attractiveness is to be treated as an independent variable that could be estimated. Other important operational development was that Huff could fit the exponent for distance in trip-making behaviour, or in other words the distance that the consumer will take to get to the store choice. Finally, another upgrade was the introduction a balancing term that constrained the sum of individual or zonal travel or sales to fit within an overall travel or sales limit (2009).

This led to the rise of a new discipline, as noted by Ziliani (2000), Geomarketing is a discipline that is the result of the crossroad between marketing and geography, their relations, and the intersection between them that consider the space characteristics in a retail network`s marketing. Geomarketing is oriented and focused on micro marketing operationalized by local adaptation being based on geographical bases. The micro marketing consists of adapting the characteristics of the local market to the specificity of the local network marketing. It represents a new defensive and offensive tool for the networks (2000).[2]

Indeed, Geomarketing is crucial for retail footprint optimization, geographic models can be built and supported by different variables such as: retailer internal data, national bureau of statistics data, competition footprint and from the combination of all of them have a final output to define how suitable or define what must be done regarding the existing footprint. In this regard, Katia Campo provides an interesting review of some of the pivotal contributions to be considered, the store localization was always considered, Volle (2006), Geomarketing add and is a huge innovation because from the use of specific analytic geographic software in combination of several dataset enables to look at geographic knowledge and not only information Bessen (1993), Cliquet (2006), Ziliani (2000). The power of analysing geographic data is huge and covers a large scope of areas such as: merchandising and assortment, Volle (2006), prices, Gonzales-Benito (2004), Montgomery (1997), promotions, service levels (e.g., store hours) can be accurately defined and supported by geographical bases. Nevertheless, retailers must always keep in mind and maintain the network identity cohesion.

Combined with mobile geolocation data, Geomarketing allows us to understand the population flow during the day, and in that way, we can analyse population dynamics and have a clear view where people are this can also be considering a privacy violation, but all depends on the data that is really

[2] See also G. Cliquet and Jérôme Baray's "Location-Based Marketing: Geomarketing and Geolocation".

collected. From the Geomarketing store footprint analysis, it is possible to identify geographic areas with business potential and having simultaneously a sense lack of retail store coverage. The geographic identification of this location, support the process from retailer expansion assisted by an analytic process rather than a blind try/error approach that will lead to bad decisions. Location is core for retail success, is the most costly and long-term marketing-mix decision. A poor location will affect and impact the retailer performance for many years unlike bad pricing or promotions decision. Being close to the consumers may also impact the retailer, due to being exposed to other retailers and intense competition. This phenomenon of stores locating near one another is called agglomeration. Different store types commonly co-locate in shopping's and malls (inter-type agglomeration). According to Fox (2007), the location of the same type of stores together (intra-type agglomeration), such as restaurants, hotels, jewellers, furniture stores, and automobile dealerships can be driven by retailers' need to be near consumers and at the same time it can also be intrinsically beneficial for retailers.

Miller, Reardon, and McCorkle (1999) suggested that net gains/losses from agglomeration depend on the balance of two countervailing forces. The first force (symbiosis) captures the incremental attractiveness of stores located close together compared to the attractiveness of those same stores individually. This incremental attractiveness reflects a reduction in consumers' costs of searching among stores and multi-purpose shopping. In effect, an agglomeration of stores becomes a shopping destination (idem, ibidem). The second force (Darwinism), a process of natural selection, reflects competition for consumer purchases among stores that sell similar products (even if they sell different products, stores compete for consumers' disposable income)- (idem, ibidem). The balance of these two forces can result in either a positive, neutral, or negative effect of agglomeration on retailer performance (Fox 2007).

This new era where everybody owns a smartphone or tablet with geolocation capabilities, and if allowed, is possible to track any activity from a geo point of view. For example, any android handset "knows where we are" and even ask to rate the locations we went with use precision. This may bring privacy concerns but on the other hand everybody will know how any business was previously rated by previous customers sharing reviews is a trend that cannot be ignored by the retailer itself.

Nowadays with tablets and smartphones massification, retailers are aiming to target consumers from a geolocation perspective, any device with geolocation can track, keep records, and report user`s real-time location with a high accuracy level. As emphasised by Bateson (2019), geolocation offers the opportunity to transform the face of any retail business completely:

> 1. location-based marketing, ex. Beacon technology (hardware transmitters installed around the store that wirelessly communicate with mobile devices within a narrow area such as a specific department or aisle) to send customized offers, push notifications, coupons, and promotional content.
> 2. Micro-moments research has shown that 50 percent of customers who run a local search on their mobile phone end up visiting a store within a day, and that 18 percent of the searches end up in the customer purchasing within 24 hours, there is the need to ensure that their mobile content plan is streamlined for every stage of customers' journeys.
> 3. Geo-targeted mobile advertisements allow retailers to reach the customer based on geographic details, enabling the ideal placement for potential ads this way we make sure that the message gets through to the people we aimed at.
> 4. Better customer experience geolocation technologies leverage to improve the overall customer experience, understanding the target audience`s preferences can be used by retailers to come up with appropriate strategies for customer`s needs and desires, ex. Taco Bell using geolocation determines when a customer who placed an order is near the store and then ensures their order is ready as soon as they walk in (Bateson 2019)".

Sharing in the end of the day will benefit us all, we will know which retailers is worth to spend our money, we will have more knowledge sharing so any retailer will be more cautious regarding customer satisfaction. This will promote a atitude of extreme responsibility regarding how the business is run. A bad review from retailer, hotel, etc. is a huge red flag and eventually causing the business to terminate.

The discussion makes clear that there is a need to analyse retail from a geo location perspective; and that such analysis should include not only sales data, but also relate that data to other data sources (such as statistical data Central Bureaus of Statistics which cover themes such as population, population wages, ethnicity, land characterization, just to name a few). As will be further discussed below, we will add one extra layer of complexity, as we will be using the possibilities provided by tracking indoor retail activity using mobile devices and, if allowed, joining this data with Big Data sets with other sources of information like transaction history, internet content accessed, Facebook user interaction and demographics. The possibilities that come from these combinations are significant regarding analysis and insights creation (ex. Infer income level), so that individuals are targeted accordingly or geotargeted. In addition, location data is extremely meaningful from an individual perspective, allowing to add based on location, push notifications through its app if you are in a nearby store or offers ads based on geographic location.[3]

2.3 REVENUE ANALYSIS

Having a clear picture and deep understanding of revenue streams is crucial to any territorial widespread retailer, in this master`s thesis this component is the main business question. In the following sections we will try to understand how revenue is impacted supported by the analysis if the available dataset. In fact, Stubbs considers that "revenue analysis is a component of any retail data analysis" (2018). This author further notes that revenue analysis enables the ability to effectively track customer actions like their purchases and foot traffic in your store, enabling management to have a clear picture of what is happening on the ground floor. This is important because there are many moving parts in a retail store, from sales inventory to customer experience and everything in between that. Retail data analysis provides the powerful insights needed to make informed decisions, avoiding the guess component, to grow your revenue and profitability by 1. Knowing our customers, 2. Analyse trends to meet customer demands, 3. Learn the true costs (2018).

Revenue is the reason for a retailer existence, to meet what the market needs and aiming to increase it is the main objective. Supported by customer satisfaction and geographic coverage for locations with potential. Revenue depends on stock management, understanding product demand is crucial. Having a retail strategy focus on it, or a system that allows centralized stock management allowing and enabling the selling process. In addition, classical regression models can also be used. Mild emphasises this very point when he says that: items that are regularly replenished, available stock can influence demand via possible stock-outs or a merchandise density effect but is no constraint on price optimization. As sales are in integer units, count distributions like the Poisson are suitable candidates for describing demand. For regularly purchased items, sales are often sufficiently high for these distributions to be approximated by a normal distribution and classical regression models can be used to estimate the elasticities of the various demand-influencing factors based on past sales time series. Otherwise, special models for slow moving items must be employed for this task. From this basis, regular and promotional prices are thus optimized (2006).

[3] See also G. Cliquet and Jérôme Baray's "Location-Based Marketing: Geomarketing and Geolocation".

With regards to dynamic pricing, the work of Erol is pertinent. This author notes that dynamic pricing is structured and continuously updated to predict demand, supported by analysis outcome, and considering the costs of the product, corporate policies and conditions prevailing in the market, the goal is to maximize revenue streams and obtain a maximum revenue level. Recently many companies adopted several policies for uplift the profitability, for example: such as optimization activities, examination of the product demand structures and markets development of products addressing different consumption groups defined in the market and introducing of such products to the market. The technological advances enable to manage revenue, so that higher yields are generated supported by tools and techniques made available recently. There are several ongoing studies conducted by academicians and company managers, aiming to improving the output amount generated by this positive effect and bettering the system.

Today ERP systems have the possibility to automatically manage company stock, this way product disruption are harder to happen. Another possibility to understand business performance is having real time metric performance or key performance indicators so that the company management is aware of business performance. Examples are key KPI, are as follow: 1. Sales per Square foot or the average amount of revenue earned per square foot of selling space, determines how effectively we are using the retail space; 2. Retail Conversion Rate is the number of visitors who make a purchase at the store; 3. Net Profit Margin or the percentage of revenue makes per euro of sales considering all business costs: marketing, payroll, transportation (Stubbs 2018). Supported by the correct advanced analytics and commercial-performance solutions, enables the cycle of consumer-goods, manufactures and retailers to capture and sustain revenue streams and higher returns through better pricing, promotions, and assortment (McKinsey & Company 2013).

2.4 National Statistical Bureau (INE) – Socio- Demographics and Psychographic

Adding variables from the National Statistical Bureau (INE) to the dataset was extremely important, as one of the business questions to answer in this master thesis is regards the relationship between socio-demographic data and retailer revenue. In the following sections, the relationship is investigated by determining which independent variables are more impactful in explaining the dependent variable (revenue). These relationships will be investigated thoroughly. But first let as clarify, in a cursory manner, the importance of demographics. As noted by Kelly, a demographic study enables us to describe who we are, ethnicity, age/generation, gender, income (disposable or net income), marital status, education, geography (influence needs, wants and access to goods and services), and homeownership allowing to classify us as a part of a group that resembles, of course that our personality is not possible to describe for that others research must be done. But some demographic characteristics is impossible to change and are at the core of our physical being. The definition can be considered as extremely accurate "The characteristics of human populations and population segments, especially when used to identify consumer markets". Retailers will find that most of the demographic data they need to make business decisions can be supported and found on the national official Census data (2016).

Official statistics are a public good, published by government agencies or other public bodies such as international organizations. They provide quantitative or qualitative information on all major areas of citizens' lives, such as economic and social development, living conditions, health, education, and the environment. Official statistics provide a picture of a country or different phenomena through data, and images such as graph and maps. It provides basic information for decision making, evaluations and

assessments at different levels.[4] National statistical bureau data is therefore extremely important, it creates a statistic base that enables to analyse a country from different perspectives: from population, income evolution, ethnicity changes, urban development just to name a few. The fact that this data has geolocation allows us to relate with many external data sources adding extra value to the model.[5] For the purposes of this thesis, we follow Kelly in considering that "demographics is crucial for retailers, if a retailer does not know what the customer wants, or ignores requests, clientele may lose interest and choose to shop elsewhere. Instead, the ideal course of action would be to learn about their consumers and then to learn about their interests. This provides the basis for developing a product mix that would be attractive to clientele". Indeed, this author further notes that, in marketing is very common and find it extremely useful to group consumers into manageable groups, or market segments, but each consumer is specifically triggered to buy, and has a unique set of criteria they use to make judgments about products we want to acquire find the product that serves all is an impossible mission. So, retailers should target and focus on those consumers that more likely will find their product appealing and will want to purchase it. Retailers must target client segments or group of clients that are looking for what they sell and serve, having a clear understanding of the segment desire and what products are suitable is the main goal. A good approach to segment the clients is according with what they spend, the more they spent the more relevant and core for the business activity they are, knowing the customer. New retailers with have a difficult initial period to understand their core or target clients, so understand the trend regarding the products and learning about consumers is a core business activity, so that the retailer can focus on the details that enable the client need and that way the appropriate product mix and promotions can be defined.

A retailer must have a clear understanding about the surrounding socio demographic where the point of sale is located, it enables to optimize stock according to the socio demographic characterization and demand of the population of the surrounding area. Of course, retailer dimension must be considered as well, a big shopping mall will attract population from different locations compared to a small retailer or local shop. We can also consider that psychographics factors, such as an individual's life experiences, personal preferences, and opinions, have a huge impact on our interest and demands. So, is crucial to have a clear understanding about interests and lifestyles (what) as well the demographic characteristics of our clients (who) on our locations these are key drivers for implementing the appropriate and a product mix that fits them (Kelly 2016).[6]

2.5 REGRESSION ANALYSIS

Linear regression will have a huge impact in the following sections and the result, it was the most important analytic model to quantify the impact and relation between the independent variables on the dependent variable. Linear Regression is categorized as a supervised learning technique for continuous variables problem, based in the input and output data, or in other words we are describing the relationship between a set of independent variables and a dependent variable, aiming to find the

[4] Note that, following Misra, "statistics is a mathematical science involving the collection, interpretation, measurement, enumerations or estimation analysis, and presentation of natural or social phenomena, through application of various tools and technique the raw data becomes meaningful and generates the information's for decision making purpose. It is the systematic arrangement of data and information which exhibits the inner relation between things. Statistics plays a vital role in every field of human activity and has important role in determining the existing position of per capita income, unemployment, population growth rate, housing, schooling medical facilities etc. in a country, by which the decision making, and development plans of the government becomes concentric. Now statistics holds a central position in almost every field of research like Industry, Commerce, Trade, Physics, Chemistry, Economics, Mathematics, Biology, Botany, Psychology, Astronomy, management of decision making (2012)".

[5] For more comprehensive details refer to website: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_base_dados&xlang=en

[6] **For more comprehensive details refer to INE website**: https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_base_dados&xlang=en

relationship between them, so that we can: 1. explain the relationship and find the extent to which the dependent is influenced by the independent variables and 2. make predictions for the dependent variable. The quality of multiple linear regression is evaluated by a minimization technique called Ordinary least square (OLS); this method estimates the relationship by minimizing the sum of squares of the difference between the observed dependent variable and those predicted by the linear function. The best fitting line is the one that minimizes the value of the OLS estimator or is the better model that fits the data. Measuring the multiple linear regression quality: 1. Goodness of fit describes how much a statistical model fits a set of observations; we aim to maximize R square (maximize) that is equivalent to minimize the sum of squares. To overcome the fact that a model has more independent variables we have the parameter adjusted R square (maximize), it penalizes model with larger number of independent variables; 2. Statistics significance quality parameters: standard error (closer to 0), t-value (greater than 1.96 for p-value less than 0.05) and p-value. In relation to the rate of profit as one of the most important indicators for stakeholders and shareholders in modern companies, Dospinescu considers that, they identified the relationship between several variables, like, net profit margin and other three composite financial indicators for companies from BET Index, that is the most important Romanian stock market index. It possible to clearly identify from the regression model, with a significance level of 95%, the stockholder's equity, long-term liabilities, provisions, deferred revenues over a year, total liabilities, working capital and current assets. The following parameters supported and enabled the validation: Multiple R, R Square, t test, F test and p-values (2019).

The power of simple regression or multiple regression is huge, in terms of understanding how different variables relate, or not, with a target variable and simultaneously we can quantify how much that impact or relation is. Following Skiera, a simple regression analysis is considered an important statistical method. We can check if a linear relationship can be observed between an dependent and a independent variable or, in other words, endogenous, explained, responde or predicted variable and one or more exogenous, explanatory, control or predictor variable. Linear regression analysis support the process of maarketing decisions, e.g., the derivation of an optimal marketing mix. The scope for using a linear regression is huge, for example it also allows to estimate nonlinear functions such as a multiplicative sales response function (2018).[7]

The fact that we can quantify the relation between target variables and independent variables, we can select the ones that are worth to add to the model. We can also understand if the variables are worth to keep in the model (redundant and redundancy). Some products are typically sold in specific markets niches, the fact that we have an understanding regarding which variables relate with the product it`s a huge competitive advantage. We can target the market more precisely, as discussed during our lectures: Professor Roberto Henriques and Professor Fernando Bação.

There are several studies with the goal of at analyzing demand estimation and forecasting using regression analysis. One specific article that I came across was about estimating demand of Medicam Toothpaste by checking the impact of several internal and external variables like: price, price of Shield toothbrush, price of Colgate toothpaste, advertisement, and total revenue on the demand as well. Regression also supports forecast the next quarter. This was achieved supported by a multiple regression analysis, using Ordinary least Square method applied on time series data of 1st quarter of 2007 to 2nd quarter of 2014 (Lecturer 2015). Afterall simple and multiple regression is a extremly powerfull analytic tool, in the sections below of the master thesis, the impact of this tool will be clearer and an understanding regarding how usefull it is will be more clear.

[7] **For more comprehensive details refer to website**: Data Analysis Using Regression and Multilevel/Hierarchical Models by Andrew Gelman and Jennifer Hill; Data Analysis: A Model Comparison Approach to Regression, ANOVA, and Beyond, Third Edition by Carey Ryan, Charles M. Judd, and Gary H. McClelland

3. METHODOLOGY

3.1 DATA MINING PROCESS

Data Mining focuses on the extraction of knowledge from data. Therefore, data mining is a mix of tools and techniques that supports the process of discovering interesting, patterns and knowledge from large datasets. For that several methodologies are available to support the data mining process, the next paragraphs describe some of the existing ones and compares them.

"CRISP – DM (Cross-Industry Standard Process for Data Mining) is the most popular methodology for analytics, data mining projects, is defined as a comprehensive data mining methodology and process model that provides a complete blueprint for conducting these projects. This methodology was conceived around 1996 and published in 1999, is non-proprietary meaning that is industry and tool neutral, is focus on business issues, has a framework for guidance and is experience base.

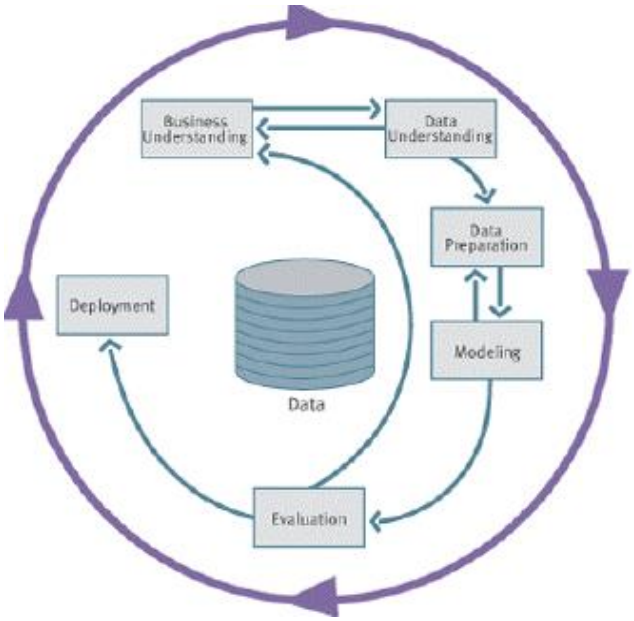

Figure 3-1 - CRISP DM Overview (source: Stellar Consulting)

CRISP – DM break down the life cycle of data mining project into 6 high-level phases: Business understanding of objectives and requirements (what does the business need); data mining problem definition; Data understanding initial data collection and familiarization (data we have), identify data quality issues, Initial and obvious results; Data preparation record and attribute selection (organize data for modeling), data cleansing; Modeling run data mining tools and data mining techniques; Evaluation to determine if results meet business objectives; Identify business issues that should have been addressed on a previously; identify which model best meets the business objectives; Deployment put the resulting models into practice (how stakeholders access the results), set up a model for having a process for continuous data mining."

A major plus of CRISP - DM is that the process is built around a framework that is reliable and repeatable. This research is also developed and supported by the analytic process or methodology from SAS Enterprise Miner (henceforth SEMMA) a tool independently developed by SAS to assist and guide SAS Enterprise miner users for data mining problems. For the purposes of this research, SEMMA is more narrowly focused (is not a comprehensive project management approach) on the technical steps of data mining, as it skips the initial Business Understanding and final Deployment phase from CRISP-DM. The first process step is data sampling processes and focus four stages: Data understanding, Data preparation, Modeling, and evaluation (Data Science Project Management - CRISP DM s.d.)". SEMMA can be considered as a CRISP – DM subset, more details in the following Table 3-1, even so a business understanding from Tasty and sweet was needed, mainly the available dataset so that this master thesis could move forward. Even the SEMMA is not as rigid as CRISP regarding the phases, we can skip some if considered not useful for the problem we have (business questions to be answered).

This software operates using flow diagrams where a specific node represents a specific task that is applied sequentially on the dataset that modified and save. SAS Institute defines data mining as the process of Sampling, Exploring, Modifying, Modeling, and Assessing (SEMMA) large amounts of data to uncover previously unknown patterns which can be utilized as a business advantage. SAS Enterprise Miner software is an integrated product that provides an end-to-end business solution for data mining. A graphical user interface (GUI) provides a user-friendly front end to the SEMMA data mining process: Sample the data by creating one or more data tables. The samples should be large enough to contain the significant information, yet small enough to process; Explore the data by searching for anticipated relationships, unanticipated trends, and anomalies to gain understanding and ideas; Modify the data by creating, selecting, and transforming the variables to focus the model selection process; Model the data by using the analytical tools to search for a combination of the data that reliably predicts a desired outcome; Assess the data by evaluating the usefulness and reliability of the findings from the data mining process.

Nevertheless, applying the SEMMA process that does not mean that we will apply all steps previously mentioned, and we may have to repeat one or more steps several times to achieve the expected and quality expected results. Indeed, the SAS Enterprise Miner interface enables us to use and to create multiple models. Statistical tools include clustering, self-organizing maps (Kohonen maps), variable selection, trees, linear and logistic regression, and neural networking. Data preparation tools include outlier detection, variable transformations, data imputation, random sampling, and the partitioning of data sets (into train, test, and validate data sets). Advanced visualization tools enable you to quickly and easily examine large amounts of data in multidimensional histograms and to graphically compare modeling results (SAS Institute Inc. 2017)".

Table 3-1 - Correspondece between SEMMA and CRISP – DM

SEMMA	**CRISP-DM**
--	Business understanding
Sample	Data understanding
Explore	Data understanding
Modify	Data preparation
Model	Modelling
Assessment	Evaluation
--	Deployment

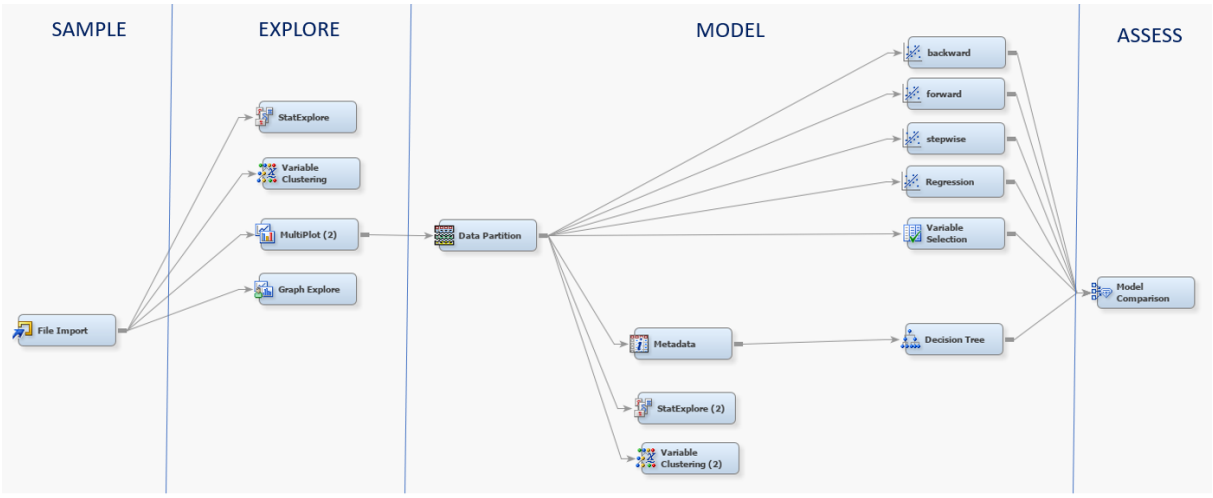

Figure 3-2 - Developed SEMMA model for the master thesis project (source: Enterprise Miner)

As previously stated, in order to meet our analysis purposes, in addition to the original dataset (including all financial and stock transactions done by Tasty and Sweet), more data was added: geographic location of the point of sales, the postal code was transformed into official administrative boundaries and Geographic National Bureau of Statistics data with geographic detailed was added to the model. Coming to this point we are enabled to use data mining analysis and techniques to meet our goals and address our business questions:

1) Are sales and revenue for each point of sales location impacted or related to socio demographic data and how impacting are they?
2) Is the retailer revenue volume impacted by discounts and promotions?

The Tasty and Sweet dataset that will be used to support the analysis, as previously described, is from a retail company. The available data allow to understand the financial meaning of each record, for this model we will use as Annexed Table 7-1 Tasty and Sweet data. In the sections 3.1.1 and 3.1.2, additional details regarding the data on these fields will be given since they are the core of financial transactions regarding revenue: Document type code and document type description are as follows:

Credit (document type 4 - NC): can be from the supplier meaning stock out at cost price, or for the client stock in at price cost, a transaction will be inserted in the database with the value and with VAT value as well.

Shipping guide (document type 3 - GR): can be from the supplier stock in at price cost, or for the client stock out at price cost, for both a transaction will be inserted in the database with the invoice: sell or bought amount, discount amount and VAT amount.

Order (document type 2 - ECL): it`s done for suppliers or the clients, for both a new record will be inserted in the database, when is from the supplier an invoice is received for it and stock in, for the client one or several shipping guide will be created, until all the order is totally complied, stock out and invoice with cost price.

Invoice (document type 4 - FA): financial document is both for the supplier meaning acquisition and stock in; or for the client reflected on one or several shipping guides stock out and price, for those ones several invoices' records will be inserted on the database with sales price and discount value if applied.

Quotation (document type 1 – FP and ORC): these records are specific client asks to the supplier if a specific product or set of products are available on a specific date and the price each one of them, no stock in or out and no shipping guide or invoices are created.

Returned items (document type 3 – DEV): these records are item's that clients give back to us.

The following set of document type and document (Table 3-2) were withdrawn from the model because they do not represent sales:

Table 3-2 Excluded document and document type.

Document and document Type	
Document	**Document type**
ORC	1
FP	1
FI	4
NCC	4
NDC	4
NDCP	4
NDD	4
NDOLD	4

3.2 OPERATION TYPE

This feature from the Tasty and Sweet dataset enable to understand the operation, regarding to have only revenue records, the feature was selected as Financial. Other's fulfillment of this variable is described below:

Financial: these are invoices, credits or debt notes from both clients and suppliers.

Stock/Transport reflects stock in or stock out from the point of sales.

Order: meaning request order management if complied shipping guide or invoices will be created reflected stock in and stock out respectively.

Client quotation: these records mean that a specific client asks the supplier if a specific product or set of products are available on a specific date and the price of each one of them.

3.3 LITERATURE REVIEW AND SELECTED ANALYSIS' TOOLS

For this set of data, the objectives are to obtain revenue for each point of sale and determine the main products sold by point of sales. As we discussed in the pages above, one of the main problems for a retailer is to manage the stock properly, due to the financial impact on the company results. Indeed, we followed Vu in considering that the main problem of inventory control is to ensure that the business runs profitably and without any disruption (2018). Geographic location data was added to this model using zip codes from the point of sales of the database, enabling to add to this dataset with an external data source, from the Portuguese Mail (CTT – Locart database), These set of fields are provided by CTT (Portuguese Mail), that supplies a complete list from Portuguese zip codes (total 325k record as is November 2019) and related them with Portuguese administrative boundaries (Distritos or Municipality and Localidade). This way we are enabled to determine revenue and products sold from a geographic location perspective. The next stage will be to proceed with a classification of point of sales regarding revenue and products sold (Marques 2013), so that revenue and products sold patterns are eventually identified.

3.4 INFORMATION AND DATA (NATIONAL BUREAU OF STATISTICS - INE) BY GEOGRAPHY

The Portuguese National Bureau of Statistics (INE) periodically releases Regional Statistical Yearbooks which include a wide range of statistical information at regional and municipal levels (Municipality administrative boundaries). These include population, enterprises and establishments, GDP, just to name a few of them. Regarding geographic boundaries they have their territorial units for Statistics: NUT1, NUT2 and NUT3 developed by Eurostat (European Union) in 1998. Eurostat boundaries are related with the administrative Portuguese boundaries, because Portuguese administrative boundaries (municipality or county) were merged into NUT`s boundaries. The NUTS system subdivides the nation into three levels: NUTS I, NUTS II and NUTS III. In some European partners, as is the case with Portugal, a complementary hierarchy, respectively LAU I and LAU II (posteriorly referred to as NUTS IV and NUTS V) is employed. The LAU, or Local Administrative Units, in the Portuguese context pertains to the 308 municipalities (LAU I) and 3092 civil parishes (LAU II) respectively. In the broadest sense, the NUTS hierarchy, while they may follow some of the borders (municipal or parish) diverge in their delineation.

Table 3-3 Relational table admisnitrative boundaries and nuts

Administrative boundaries and Nut`s		
Field Name	**Field Description**	**Field Type**
COD_DISTRITO	Administrative boundary Distrito id code number	Numeric
DSC_DISTRITO	Administrative boundary Distrito name	Character
COD_CONCELHO	Administrative boundary Municipality id code number	Numeric
DSC_CONCELHO	Administrative boundary Municipality name	Character
DTCC	Administrative composed boundary Distrito and Municipality	Character
NUT1	Administrative boundary for statistical purpose level 1	Character
NUT2	Administrative boundary for statistical purpose level 2	Character
NUT3	Administrative boundary for statistical purpose level 3	Character

The use of this data is crucial, as it allows us to join data from the point of sales associated with the zip code from the distributor, and consequently, it is possible to translate to administrative official boundaries (District and Municipality). Since INE produces information geographically, we can relate the point of sales with social demographic data using the field municipality (Municipality) as Annexed Table 7-2 National Statistical Bureau Information by Municipality.

3.5 EXPLORATION OF THE DATASET

3.5.1 Data Access

The Tasty and Sweet Dataset has 56.002 observations, with a total of 53 variables, these variables and their description are covered in Table 9-1. The INE dataset is described in Tables 9-2; 9-3 and 9-4 in the Annex section below. Firstly, with the aid of Input Data node from Sas Miner, enabled to read the Distribuidor_02dat file available on the created library, for establishing the role and level for the dataset variables as Annexed Table 7-3 Tasty and Sweet variables description

3.5.2 Exploration of the dataset

Aiming at the exploration of primary data, the following nodes from SAS Miner were added to the diagram:

StatExplore node, for dataset statistics display
Multiplot node, for generate several bar and scatter plot chart for target and input variables.
Variable Clustering node, to uncover and find patters and the structure of input variables.
Graph Explore node, enables to visualize univariate and multivariate distribution.

The StatExplore node output is displayed in the following tables for class and interval variables, analyzing it:

93% of sales are done in “Portugal Continental.”
68.92% of the sale are for the family product 00001.
48,5% of the sales are done at “Área Metropolitana de Lisboa.”

Table 3-4 Class Variable Summary Statistics T&S

Variable Name	Missing Values	Mode	Mode Percentage	Mode2	Mode2 Percentage
Family product code	781	00001	68.92	00002	15.71
Zone code	152	06	32.15	01	18.57
Brand	782		14.98	021	7.26
NUT1	16	Portugal Continental	93.27	Região Autónoma dos Açores	3.85
NUT2	16	Região de Lisboa	48.55	Região do Centro	15.94
NUT3	16	Área Metropolitana de Lisboa	48.55	Algarve	13.62
Zone Code	152	06	32.15	01	18.57
Subfamily product 2	783	DESAYUNO / PANADERIA	22.09		15.00
Municipality name	0	Lisboa	33.42	Faro	9.18
Distrito name	0	Lisboa	48	Faro	13.62
District code & Municipality code	0	1106	33.42	0805	9.18
Family product	781	00001	68.92	00002	15.71
Subfamily product	783	00001	22.09		15.00

Unit out	547	C12	21.71	C6	19.60

We should note that of the issues to be addressed from the analysis of table 3-5, the missing values do not present a problem on the interval variables; yet the issue relating to class variables need to be addressed. Furthermore, considering the mean value from the variables, we have a great distance between the values, IVA total vs IEC value (the means are 2,041 and 342,9 respectively) or the discount values vs Total Revenue. We address the different unit measures in the dataset (monetary values measure in euros vs years) to procced with data normalization. Finally, looking at the mean value of interval variables vs the maximum value of them, is not clear that we in the presence of outliers. Even comparing the median with the maximum value of each interval variable.

Table 3-5 Interval Variable Summary Statistics

Variable Name	Mean	Standard Deviation	Non-Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Year	2018	0.5	5,219	0	2018	2019	2019	-0.29329	-10.3791
Month	7	3.17	5,219	0	1	7	12	-0.3452	-0.87817
Iva total	2,041	7,171.9	5,219	0	-41,505.9	349.75	154,590.4	10.16405	157.0827
Merchandise total	17,081	67,727.2	5,219	0	-236,972	2,805.9	1.915,095	14.63282	328.8853
Discount 1	2.062	9.39	5,219	0	0	0.235	180.3546	11.77257	167.3
Discount 2	1.2537	5.93	5,219	0	0	0.0208	112.5	9.734795	116.8718
Discount 3	0.19	3.1	5,219	0	0	0	176.4	41.94596	2137.388
Total Revenue	13,468	4,1319.8	5,219	0	-221,966	2,571.9	826,719	8.559851	116.9871
IEC value	342.9	2.240.3	5,219	0	-7.470,75	0	60,973.5	14.65187	302.3542

An extra The StatExplore node was added to explore the cross-tabulation report, the variables NUT3 vs zone and NUT3 vs Subfamily product 2. It was possible to extract the following information:

Sales at NUT3 “Área Metropolitana de Lisboa”, accounts for 48% of total sales

Sales at NUT3 “Algarve”, accounts for 14% of total sales

Sales at NUT3 “Área Metropolitana de Lisboa”, about 12% are for “DESAYUNO / PANADERIA”

Sales at NUT3 “Área Metropolitana de Lisboa”, about 7% are for “GALLETAS /CHOCOLATES / DULCES”

Sales at NUT3 “Algarve”, about 3.5% are for “DESAYUNO / PANADERIA”

Sales at NUT3 “Área Metropolitana do Porto”, about 2.6% are for “DESAYUNO / PANADERIA”

The Multiplot node and Graph Explore node output enables to visualize variable distribution, that way outliers, “strange values” and missing values can be identified to assist us with that several graphs were created and made available on the annex section, Fig. 9.1 to Fig 9.11, from these graphs: Revenue, Merchandise and IEC eventually we are in the presence of outliers (highlighted in red), as we are dealing with sales and they have different volumes/quantities and can be product devolutions as

well, in that way the outliers detection task is significantly harder. Variable Clustering node points out how the data is structured and produces hierarchical clustering with numeric variables. The dendrogram allows to identify that three clusters' accounts for 80% of the variance in the dataset Figure 3.1. Examining the relationship between pairs of internal variables Figure 3.2, we can identify coefficient with absolute value above 0.8, in that there are several variables correlated allowing to discard one in favor of other.

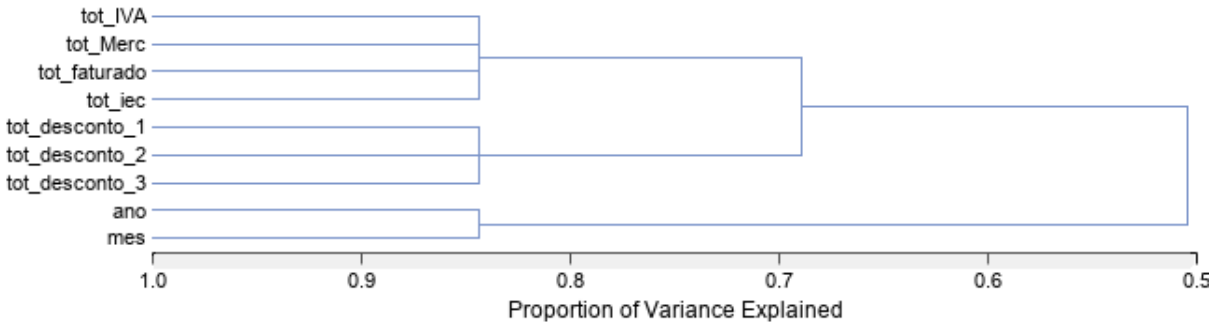

Figure 3-3 Dendrogram tasty and sweet (source: Enterprise Miner)

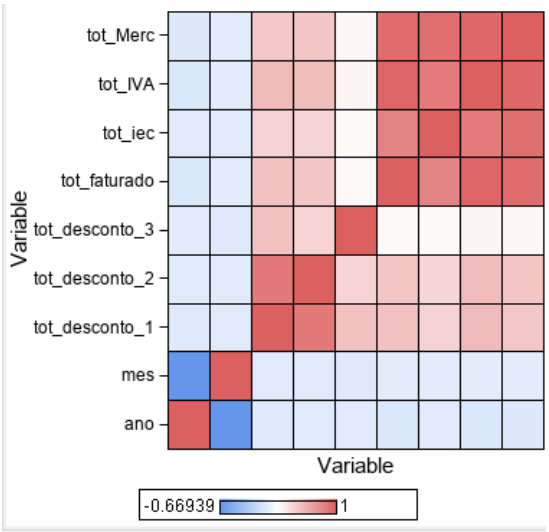

Figure 3-4 Variable Correlation Tasty and Sweet (source: Enterprise Miner)

Table 3-6 Variable Correlation Tasty and Sweet Detail (source: Enterprise Miner)

Variable 1	Variable 2	Correlation Value
Discount 1	Discount 2	0.86
Discount 2	Discount 1	0.86
IEC total	IVA total	0.86
IEC total	Merchandise total	0.92
IEC total	Total Revenue	0.82
IVA total	Merchandise total	0.96
IVA total	Total Revenue	0.97
IVA total	IEC total	0.86
Merchandise total	IVA total	0.96
Merchandise total	Total Revenue	0.93
Merchandise total	IEC total	0.92
Total Revenue	IVA total	0.97
Total Revenue	Merchandise total	0.93
Total Revenue	IEC total	0.82

Furthermore, the Pearson correlation Table 3-7 Pearson Correlation Coefficients Tasty and Sweet, was produced to reinforce variable correlation analysis.

Table 3-7 Pearson Correlation Coefficients Tasty and Sweet (source: Enterprise SAS Guide)

Pearson Correlation Coefficients							
	Merchandise Total	**IVA total**	**IEC total**	**Discount 1**	**Discount 2**	**Discount 3**	**Total Revenue**
Merchandise Total	**1**	**0.96**	**0.92**	0.46	0.47	0.21	**0.93**
IVA total	**0.96**	**1**	**0.86**	0.53	0.52	0.22	**0.97**
IEC total	**0.92**	**0.86**	**1**	0.40	0.39	0.20	**0.82**
Discount 1	0.46	0.53	0.40	**1**	**0.86**	0.49	0.48
Discount 2	0.47	0.52	0.39	**0.86**	**1**	0.39	0.47
Discount 3	0.21	0.22	0.20	0.49	0.39	**1**	0.20
Total Revenue	**0.97**	**0.97**	**0.82**	0.48	0.47	0.2	**1**

At this stage and analysing the output from the variable correlation and Pearson correlation, some decisions were taken regarding the variables: merchandise total, IVA total and IEC total were dropped for the future model due to having correlation above 0.8 but Total Revenue is kept, regarding the variables discount 1 and discount 2 only discount 1 was kept, overall, there was a shrink in the variables from this dataset.

3.5.3 Exploration of INE

The StatExplore node output is displayed in the following tables for class and interval variables the variable with more missing values is Rend_09 (Mean monthly earning of employees in establishments by municipality and according to level of education PhD, 2016). The variables Rend_02 and Rend_06 looks to have the same data, from the analysis of the following data Table 3-8. If confirmed one of them will be dropped from the model.

Table 3-8 Class Variable Summary Statistics INE (source: Enterprise Miner)

Variable Name	Mean	Median	Non-Missing	Missing	Minimum	Maximum
Ocup_01	49,111	8,965	347	0	182	3,596,827
Ocup_02	79,126	10,628	347	0	191	5,942,131
Pop_01	81,652	9,278	347	0	305	6,152,409
Pop_02	88,275	9,971	347	0	252	6,655,715
Pop_19	9,889	1,235	347	0	25	744,923
Pop_20	20,591	2,565	347	0	61	1,536,298
Pop_21	9,491	1,088	347	0	36	713,129
Pop_22	9,145	996	347	0	26	687,843
Pop_23	10,413	968	347	0	43	792,369
Pop_24	7,725	527	347	0	20	596,589
Rend_01	69,170	7,621	338	9	207	5,108,104
Rend_02	1,210,558	111,187	338	9	3,827	90,360,666
Rend_03	1,061,842	100,917	338	9	3,470	79,094,031
Rend_04	15,134	14,672	328	19	10,700	27,263
Rend_05	98,070	11,056	338	9	289	7,220,626
Rend_06	1,210,558	111,187	338	9	3,827	90,360,666
Rend_07	1,061,842	100,917	338	9	3,470	79,094,031
Rend_08	10,522	10,242	328	19	7,176	20,069
Rend_09	1,951	1,810	118	229	544	9,671
Rend_10	1,391	1,340	267	80	869	2,740
Rend_11	1,399	1,318	309	38	946	5,271
Rend_12	1,477	1,389	280	67	741	6,403
Rend_13	918	890	309	38	699	3,273

The Multiplot node and Graph Explore node were not explored deeply for this dataset, only a few graphs were generated, they are available in the annex section in Figure 7-12 Estimates of housing stock by municipality - buildings, 2017 to Figure 7-17 Mean declared gross income by fiscal household € 2017, the reason for this decision is because this data source is produced and managed by National Statistics Central Bureau and due to that only extreme cases will be handled properly.

Variable Clustering node: the dendrogram allows to identify that two clusters' accounts for 80% of the variance in this dataset (Figure 3-5 Dendrogram INE). Examining the relationship between pairs of internal variables (Figure 3-6 Variable Correlation INE), we can identify coefficient with absolute value above 0.8, there are several variables correlated allowing to discard one in favor of other.

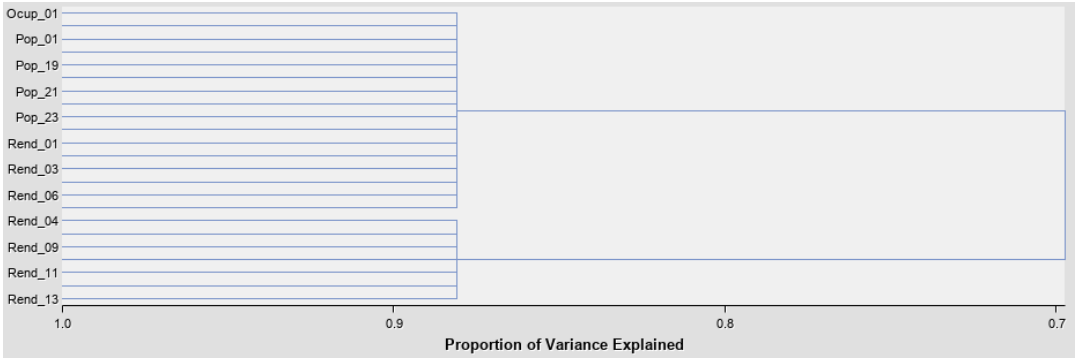

Figure 3-5 Dendrogram INE (source: Enterprise Miner)

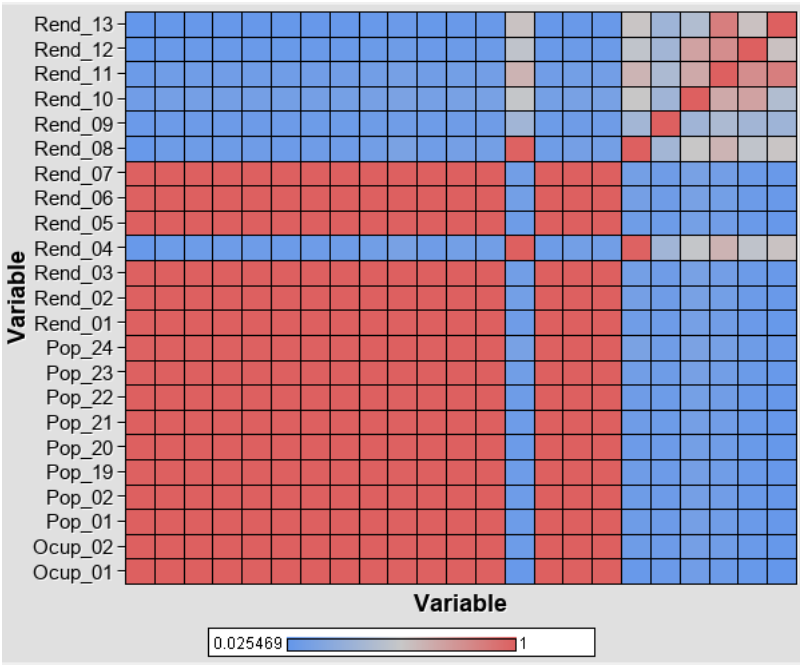

Figure 3-6 Variable Correlation INE (source: Enterprise Miner)

Furthermore, the Pearson correlation table, was produced to reinforce variable correlation analysis, firstly by type of data: "Ocup", "Pop" and "Rend" aiming to select the appropriate variables for the future model.

Table 3-9 Pearson Correlation Coefficients INE – “Ocup” (source: Enterprise SAS Guide)

Pearson Correlation Coefficients Ocup		
	Ocup_01	**Ocup_02**
Ocup_01	**1**	0.99
Ocup_02	0.99	**1**

The high correlation between the variables “Ocup” was confirmed from the Pearson table.

Table 3-10 Pearson Correlation Coefficients INE – “Pop” (source: Enterprise SAS Guide)

Pearson Correlation Coefficients Pop								
	Pop_01	**Pop_02**	**Pop_19**	**Pop_20**	**Pop_21**	**Pop_22**	**Pop_23**	**Pop_24**
Pop_01	**1**	0.99999	0.99928	0.99926	0.99983	0.99985	0.9984	0.99215
Pop_02	0.99999	**1**	0.99928	0.99917	0.99983	0.99987	0.99855	0.99254
Pop_19	0.99928	0.99928	**1**	0.99933	0.99921	0.99922	0.99708	0.99104
Pop_20	0.99926	0.99917	0.99933	**1**	0.99936	0.99911	0.99596	0.98773
Pop_21	0.99983	0.99983	0.99921	0.99936	**1**	0.99992	0.99844	0.9921
Pop_22	0.99985	0.99987	0.99922	0.99911	0.99992	**1**	0.99879	0.99309
Pop_23	0.9984	0.99855	0.99708	0.99596	0.99844	0.99879	**1**	0.99725
Pop_24	0.99215	0.99254	0.99104	0.98773	0.9921	0.99309	0.99725	**1**

All the "Pop” variables are highly correlated, like the variables “Ocup” (0.99).

Table 3-11 Pearson Correlation Coefficients INE – “Rend” (source: Enterprise SAS Guide)

Pearson Correlation Coefficients Rend													
	Rend_01	**Rend_02**	**Rend_03**	**Rend_04**	**Rend_05**	**Rend_06**	**Rend_07**	**Rend_08**	**Rend_09**	**Rend_10**	**Rend_11**	**Rend_12**	**Rend_13**
Rend_01	**1**	0.99937	0.99966	0.0793	0.99997	0.99937	0.99966	0.08475	0.07236	0.11256	0.08983	0.05744	0.04108
Rend_02	0.99937	**1**	0.99995	0.10082	0.99911	1	0.99995	0.10772	0.07916	0.12383	0.1005	0.0658	0.04934
Rend_03	0.99966	0.99995	**1**	0.09597	0.99947	0.99995	1	0.10242	0.07761	0.12125	0.0981	0.06399	0.04749
Rend_04	0.0793	0.10082	0.09597	**1**	0.07611	0.10082	0.09597	0.99116	0.32222	0.49362	0.61108	0.46753	0.54261
Rend_05	0.99997	0.99911	0.99947	0.07611	**1**	0.99911	0.99947	0.08104	0.0711	0.11034	0.08793	0.05592	0.03972
Rend_06	0.99937	1	0.99995	0.10082	0.99911	**1**	0.99995	0.10772	0.07916	0.12383	0.1005	0.0658	0.04934
Rend_07	0.99966	0.99995	1	0.09597	0.99947	0.99995	**1**	0.10242	0.07761	0.12125	0.0981	0.06399	0.04749
Rend_08	0.08475	0.10772	0.10242	0.99116	0.08104	0.10772	0.10242	**1**	0.33291	0.51519	0.60614	0.47243	0.52694
Rend_09	0.07236	0.07916	0.07761	0.32222	0.0711	0.07916	0.07761	0.33291	**1**	0.31763	0.37669	0.32629	0.30987
Rend_10	0.11256	0.12383	0.12125	0.49362	0.11034	0.12383	0.12125	0.51519	0.31763	**1**	0.66013	0.69102	0.39978
Rend_11	0.08983	0.1005	0.0981	0.61108	0.08793	0.1005	0.0981	0.60614	0.37669	0.66013	**1**	0.78859	0.86061
Rend_12	0.05744	0.0658	0.06399	0.46753	0.05592	0.0658	0.06399	0.47243	0.32629	0.69102	0.78859	**1**	0.54767
Rend_13	0.04108	0.04934	0.04749	0.54261	0.03972	0.04934	0.04749	0.52694	0.30987	0.39978	0.86061	0.54767	**1**

For the “Rend” variables the correlations have different values (from 0.01 to 0.99) opposite to “Ocup” and “Pop”. We should note that the correlation analysis was extremely time consuming, and yet critical because decisions at this phase will have a huge impact in the future model. Analytics models must be fed with variables that are not redundant (feed with the important and relevant ones) and relevant. Also, we must consider the size of the input space, sparse, known as the curse of dimensionality, as this grows it becomes harder to find groups and patterns. Some techniques for Normalization or standardization aiming to have common scale for the variables, were applied to the variables population and “Rend”.

4. RESULTS AND DISCUSSION

4.1 ARE SALES AND REVENUE IMPACTED OR RELATED TO GEOGRAPHIC SOCIO DEMOGRAPHIC DATA?

Coming to this point, several decisions regarding the variables to use and the transformations to be done on them have to been made, so that we can move forward with a dataset that enables to make valid and insightful conclusions, only ten variables were selected for the model and seven of them transformed. The following Table 4-1 and Figure 4-1, resumes the selection, descriptive statistics and actions made and identifies the worth of the chosen set on the target variable:

Table 4-1 Variables and transformations (source: Enterprise Miner)

Variable Name	Mean	Median	Missing	Non-Missing	Minimum	Maximum	Variable Transformation
Municipality name							No
Territorial structure by municipality - weight of resident population, 2011	0.023378	0.0169	0	56002	0.001323	0.053159	Yes
Total Revenue	242.7148	123.3684	0	56002	0	17723.69	No
Indicators of enterprises by municipality N. º/km2 - enterprise density, 2016	449.8616	315.1	0	56002	11.8	1046.4	No
Indicators of enterprises by municipality - total turnover, 2016	520.7759	530.2	0	56002	93.8	1161.5	No
Distribution of declared gross income less individual tax income paid of tax households by municipality (less 5k), 2016	4.692368	5	0	56002	2	6	Yes
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 5k and 10k), 2016	1.033677	1	0	56002	1	2	Yes
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016	4.710242	4	0	56002	2	6	Yes

Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016	4.17228	5	0	56002	3	6	Yes
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 19k and 32.5k), 2016	2.494625	2	0	56002	2	4	Yes
Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016	3.896807	3	0	56002	1	6	Yes

The Territorial structure by municipality - resident population, 2011 was transformed by weighted by the total population, and all variables named Distribution of declared gross income less individual tax income paid of tax households by municipality were ranked between 1 and 6 according to their value by municipality.

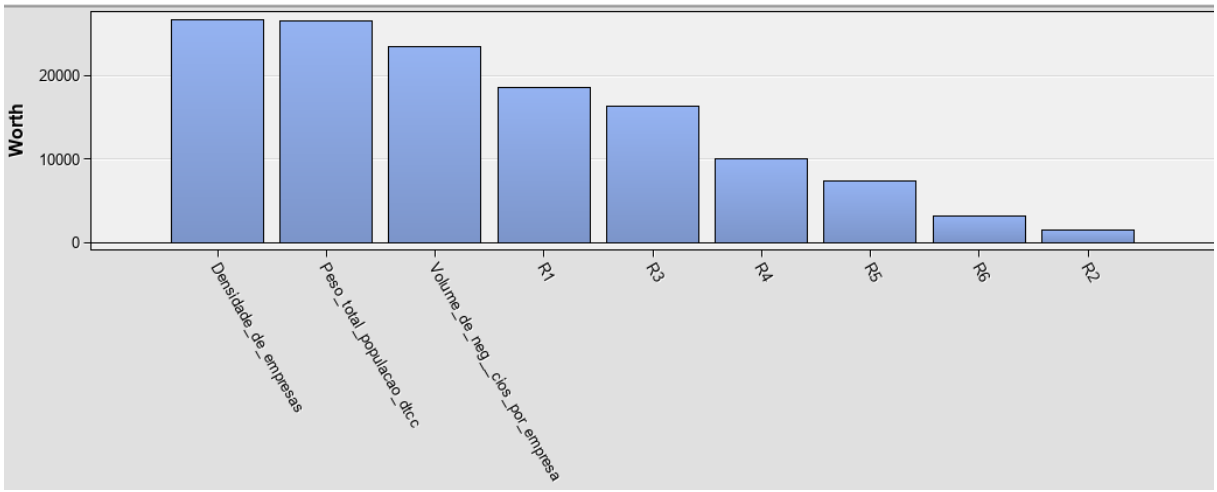

Figure 4-1 SAS Variable worth on target total revenue variable (source: Enterprise Miner)

From the SAS Variable Clustering node, variable correlation we can check the correlation between the model variables.

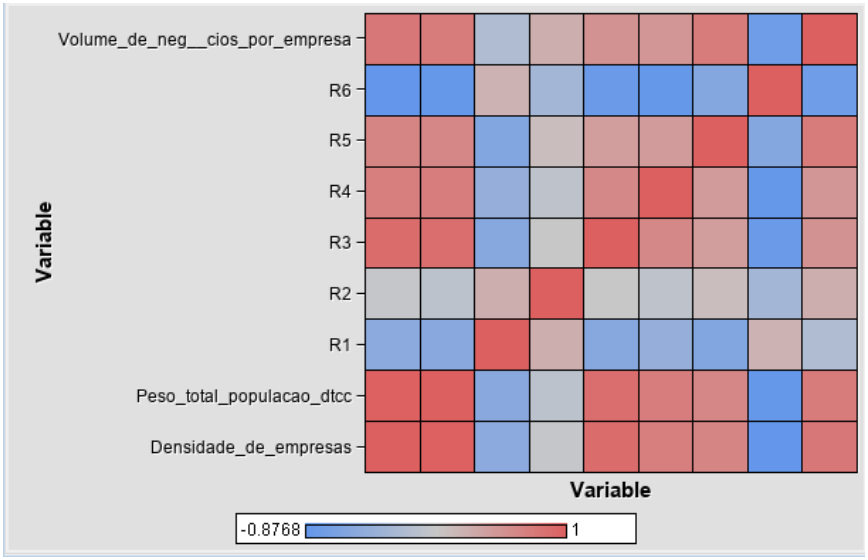

Figure 4-2 Variable correlation before partition (source: Enterprise Miner)

From the analysis of Variable correlation (Figure 4-2) we can point out that "Territorial structure by municipality - weight of resident population, 2011" and "Indicators of enterprises by municipality N.º/km2 - enterprise density, 2016" have a high value of correlation 0.99; Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016 (R3) and Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016 (R4) have 0.63 correlation value; Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016 (R4) and Distribution of declared gross income less individual tax income paid of tax households by municipality (between 19k and 32.5k), 2016 (R5) have a 0.47 correlation value; a negative correlation (-0.77) was identified between the variables Indicators of enterprises by municipality - total turnover, 2016 and Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016 (R6).

To further support the variable selection process, and to determine which variables impact the target variable, several models were applied on the dataset: Regression (regression, regression stepwise, regression forward and regression backward) and a Decision tree. Before applying the models, the dataset was submitted to data partition in three exclusive datasets, with the following distribution: 70% training, 20% validation and 10% test. Partition method selected is simple random, so that every observation in the data set has the same probability of being written to one of the partitioned data sets. Looking at Fig. 4-3, some variables worth changed: Territorial structure by municipality - weight of resident population, 2011 became the top one, and Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016 (R4) became less relevant than Distribution of declared gross income less individual tax income paid of tax households by municipality (between 19k and 32.5k), 2016 (R5).

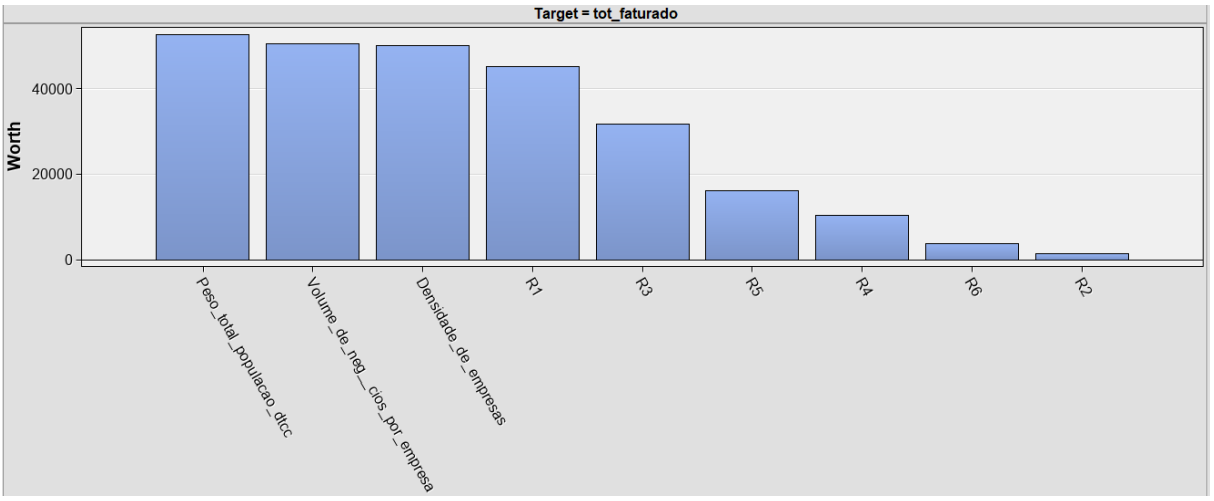

Figure 4-3 Variable worth after data partition (source: Enterprise Miner)

Looking at Figure 4-3 some variables worth changed after the previous operation: Territorial structure by municipality - weight of resident population, 2011 became the top one, and Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016 (R4) became less relevant then Distribution of declared gross income less individual tax income paid of tax households by municipality (between 19k and 32.5k), 2016 (R5), the overall the impacts are limited.

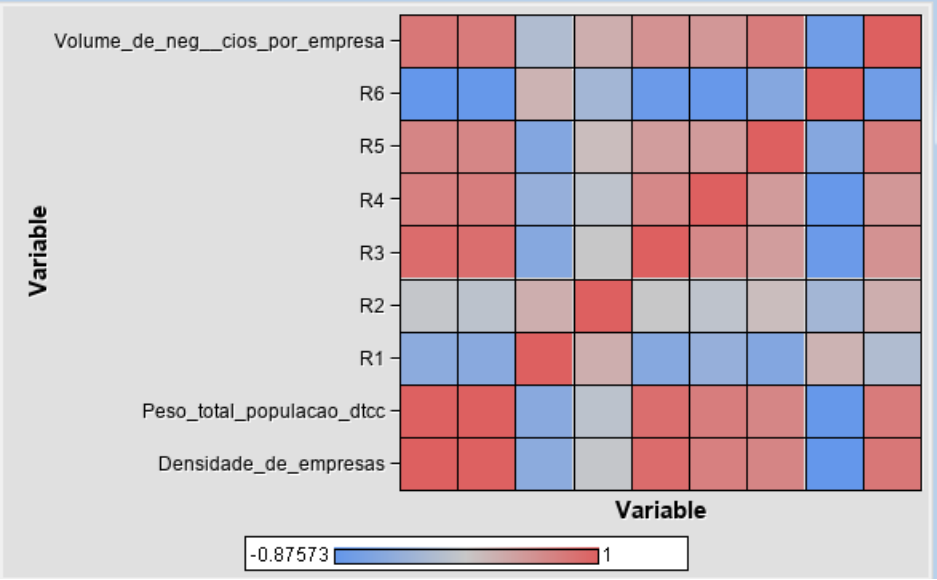

Figure 4-4 Variable correlation after partition (source: Enterprise Miner)

From the analysis of Figure 4-4 Variable correlation after partition, no relevant changes were detected, “Territorial structure by municipality - weight of resident population, 2011” and “Indicators of enterprises by municipality N. º/km2 - enterprise density, 2016” have a high value of correlation 0.99; Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016 (R3) and Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016 (R4) kept 0.63 as correlation value; a negative correlation (-0.77) was identified between the variables Indicators of enterprises by municipality - total turnover, 2016 and Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016 (R6).

To further assist in the variable selection process, and to determine how much the independents variables impact the target variable, several models were applied on the dataset: Regression (regression, regression stepwise, regression forward and regression backward) and a Decision tree.

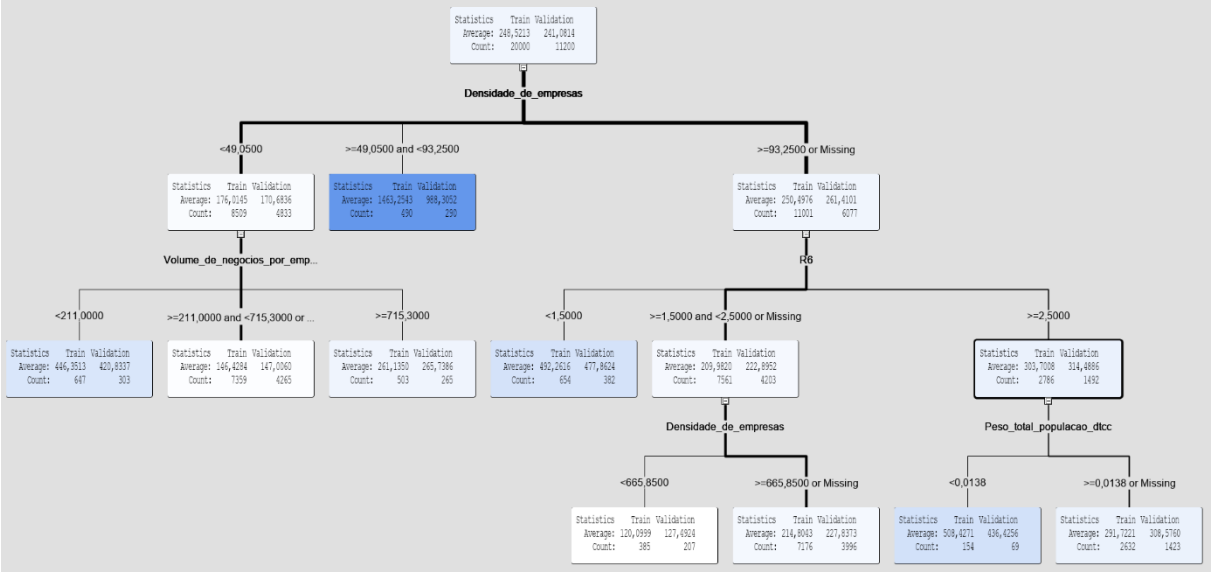

Figure 4-5 - Generated Decison tree (source: Enterprise Miner)

From the Decision tree output, we have a clear understanding of variable importance or variable ranking. Decision trees have a lot of advantages in terms of interpretability, but the results are not as good as other more robust models, decision tress allow us to have a feeling from the data and how the data is related intrinsically, which variables are more related and more important. From the generated output we can consider the most relevant are: enterprises by municipality N. º/km2 - enterprise density, 2016; enterprises by municipality - total turnover, 2016; Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016; Territorial structure by municipality - weight of resident population, 2011.

For further support more models were produced, allowing us to get a clear understanding of the relevance and importance of which socio-demographic variables impact on the dependent variable total revenue. To quantify this impact the regressions models were deeper analyzed, and the decision tree model was considered only helpful because for most problems it will not produce good models as regressions. We should note that **linear regressions** are evaluated by:

1) Goodness of fit, that describe how the model fit a set of observations or how good the estimated model is. For that we will be assisted by the following parameters:

   R-squared (goal to maximize)
   Adjusted R-squared (goal to maximize)
   F-Statistic (goal to maximize)
   Sum of square (goal to minimize)

In the followings tables 4-2 and 4-3, we have the previous parameter values for each regression model used.

Table 4-2 R-square, Adjusted R-square and F value by regression

Model	R-Square	Adjusted R-Square	F Value
Regression Backward	0.1433	0.1431	819.60
Regression Forward	0.1433	0.1431	819.60
Regression Stepwise	0.1433	0.1431	**936.44**
Regression	0.1433	0.1431	819.60

All the linear regressions: forward, backward, stepwise and regression returned the same values for the parameters R-squared and Adjusted R-square, but regarding the F parameter the best model is regression stepwise.

Table 4-3 Regression Sum of squares

Model	Sum of square
Regression Backward	1428240480
Regression Forward	1428240480
Regression Stepwise	**1427884776**
Regression	1428240480

Regarding the parameter Sum of square, the regression stepwise has the lowest value, pointing out that for the goodness of fit this is the best model.

2) Model statistical significance, for regression stepwise will be assisted by the parameters: standard error, t value and p value. The following Table 4-4 support the analysis to determine the most impacting variables, pointed out by the most relevant statistic value or p value.

Table 4-4 Regression Stepwise p value statistics (source: Enterprise Miner)

Variable Name	**Pr>|t| Regression Stepwise**
Indicators of enterprises by municipality N. º/km2 - enterprise density, 2016	<.0001
Territorial structure by municipality - weight of resident population, 2011.	<.0001
Distribution of declared gross income less individual tax income paid of tax households by municipality (less 5k), 2016	<.0001
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 5k and 10k), 2016	<.0001
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016	<.0001
Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016	<.0001
Indicators of enterprises by municipality - total turnover, 2016	<.0001

From the above table 4-4 and if P the value for above variables is below previously selected threshold (0.05), then all these variables are considered statically significant.

To have a solid selection for the best regression model, the node model comparison was added to the project. The only updated parameter was the model selection statistic to Gini Coefficient. This node enables us to compare the performance of competing models, for the selection criteria the average squared error on the validation dataset was elected, and the one with the lowest value is the regression stepwise.

Table 4-5 Average squared error value for validation dataset (source: Enterprise Miner)

Model	Averaged Squared Error for Validation Dataset
Regression stepwise	**176794,20**
Regression forward	176851,88
Regression backward	176851,88
Regression	176851,88

4.2 IS THE RETAILER REVENUE VOLUME IMPACTED BY DISCOUNTS AND PROMOTIONS?

One of the business questions of this thesis, is to understand the relation between retailer revenue and promotions and discounts done. Supported by a simple linear regression model, that relates the continuous target revenue, and to measure how much is influenced by the independent variables discount. Several graphs were done to highlight the “relation” between these two variables:

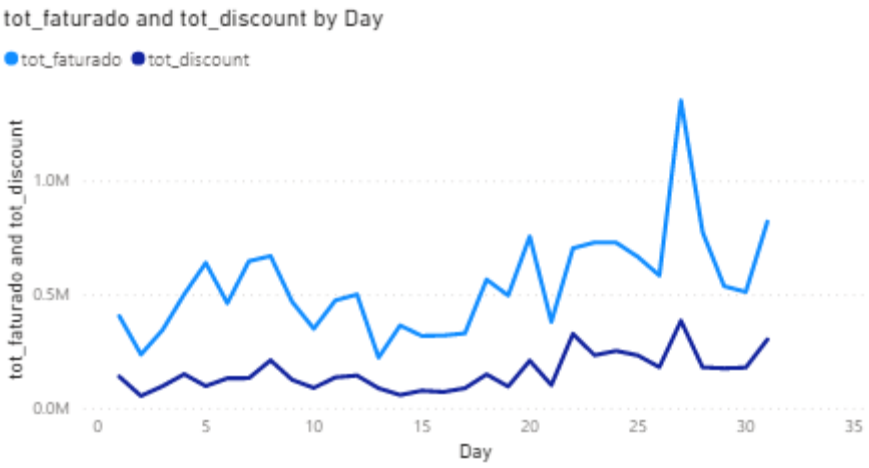

Figure 4-6 Revenue vs Discount (source: Microsoft Power BI)

Firstly, a graph chart was produced to plot the variation between these two variables, throughout a specific time frame, it`s clear that the increases in revenue (tot_facturado) is related and impacted by the increase in discount and promotions (tot_discount).

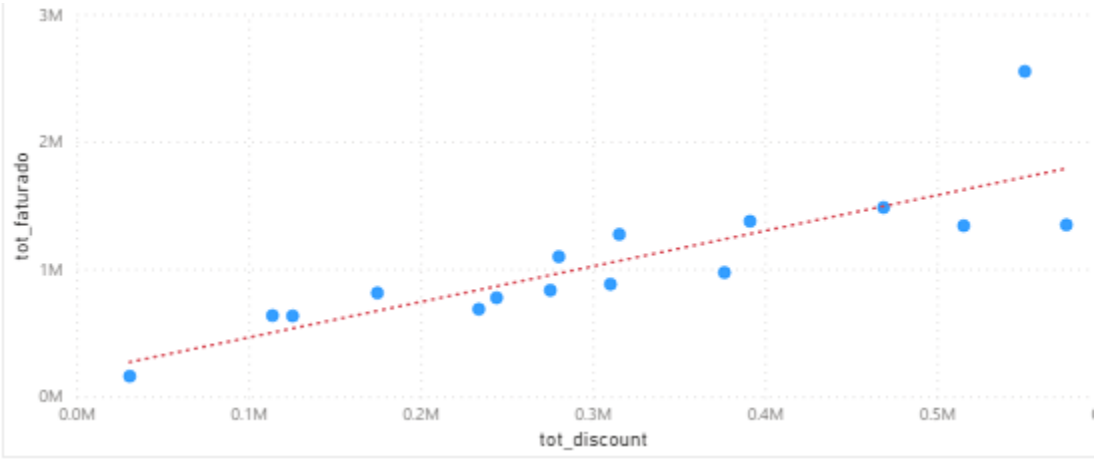

Figure 4-7 – Scatter plot Revenue and discounts (source: Microsoft Power BI)

Secondly, a scatter plot with the continuous variable's total revenue and total discount, this graph allows to understand the linear relation between these two variables, from the trendline is clear we have a positive relation between them (both increase simultaneously). This graph is useful to interpreter the type of correlation between these variables.

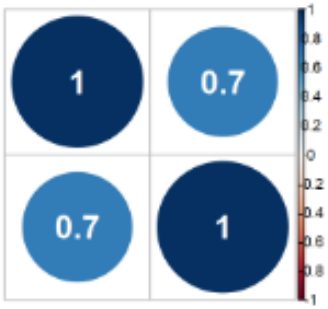

Figure 4-8 Correlation plot Revenue and discounts (source: Microsoft Power BI)

Thirdly, a correlation plot was produced to determine the dependence between these variables revenue and discounts, that is supported by the correlation coefficient value between them, the obtained value is 0.7, pointing out a moderate relationship (moderate correlation is usually a value between 0.5 and 0.7), represented by the blue scale color that translates in to a moderate correlation value.

5. CONCLUSIONS AND RECOMMENDATIONS FOR FURTHER RESEARCH

This concluding section will provide, in a summarized fashion, the main findings of the research conducted, some of the obstacles encountered as well as some of its limitations. In the concluding paragraphs, several suggestions for further research are put forward in the hope that this important area of study and practice be taken forward. This thesis focused on, aimed, and supported by a data science project, to analyze and relate different and various data sources: sales, and socio-demographic data. To that end, the following methods were used: data mining "CRISP – DM (Cross-Industry Standard Process for Data Mining) and methodology from SAS Enterprise Miner (henceforth SEMMA) a tool independently developed by SAS to assist and guide SAS Enterprise miner users for data mining problems.

Our main findings, as discussed in the chapters above were firstly, determine how the data is related intrinsically and which variables are more related: Relative to the business questions answers, supported by this master`s thesis: Firstly, identify the most impacting variables on the target revenue variable: Indicators of enterprises by municipality N. º/km2 - enterprise density, 2016; Territorial structure by municipality - weight of resident population, 2011; Distribution of declared gross income less individual tax income paid of tax households by municipality (less 5k), 2016. Secondly quantify if the retailer revenue volume impacted by discounts and promotions supported by graphs analysis.

The applied research conducted for this master's thesis faced several roadblocks but most of them were overcome. The first one was the interpretation of the retailer dataset T&S, mainly due to the lack of experience with this data and to get a deeper understanding of it and how it related took many hours and raised many questions. After reaching this understanding, the way forward was clearer. Geotagging the retailer point of sales based on their zip codes, for that a geographic master table was developed, supported by the CTT dataset, that relates the zip codes and geographic boundaries. This phase was less difficult to overcome, but all was done from scratch, and due to my professional experience and know-how this phase ran smoothly. The collection of a great deal of socio-demographic data, some of it irrelevant for the purposes of this research, had a considerable impact in the model development phase. Many variables have clear input errors while others do not, so that bringing any value for the future model, required that a huge cleaning process be done before the modelling stage. Many tests and interpretation were needed to produce the final output. Another huge challenge was the fact that the variables had different scales, so to have a model that was not polluted by the different variable scale many transformations, standardization, were done in the dataset so that the future models were not affected and impacted.

All the previous steps led to a dataset that, in my understanding, makes sense and produced an insightful and reliable model. From my understanding the main goals were, indeed, achieved, supported by relating different datasets and produce actionable and insightful information that answered the questions in the beginning of these master's thesis: firstly, which are the socio demographic with more impact variables on revenue; secondly is revenue impacted by promotions and discounts. For the first question it was possible to identify the following variables: Distribution of declared gross income less individual tax income paid of tax households by municipality (between 5k and 10k), 2016; Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016; Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016; Indicators of enterprises by municipality - total turnover, 2016 supported by several models (decision trees and regression), the outcome from the regression and the parameters pointed out what to consider; The second question was also possible to answer, supported by graphic analysis from the available data, the retailer revenue is in fact impacted by discounts and promotions throughout time.

Throughout the development of this dissertation, and regarding the technical component, it was possible to relate different datasets. The main dataset has daily sold item's operations, for a specific data frame, by geography from the retailer; the second dataset had socio demographic variables by geography. Geography was the connection point for both datasets, allowing to understand the business by geography, relating with other information sources so that business decisions could be taken.

My personal understanding is that this dissertation will be a plus for academic institutions as well for enterprises. The fact that it was possible to relate different dataset is a plus, because new and unknown data relations (correlation) were identified with the available variables. As this specific subject is not in a mature stage, regarding previous research, this one is a guideline and can assist other retailers to use, as support, to identify the best potential new locations and deploy new point of sales supported by a specific method to identify them, that way revenue streams and high value areas are found supported by an analytic method as previously discussed in this dissertation.

For future works it will be key to find the model that explains, with best accuracy, the relationship between the dependent and independent variables, this can be achieved with software available on the market for this goal. Optimizing the warehouse stock based on real demand is another study that should be addressed, that will impact all the supply chain, optimizing it, and will contribute, eventually, towards increasing revenue streams from avoiding disruption of products on any point of sale. Also, optimization of the retail footprint will be a plus, and for that purpose it would be crucial to have methodology that enables to capture this, so that is possible to rethink the design of the physical store network, following a four-stage methodology: Customer and Competitor Profiling to have a clear understanding of who shops in the stores - demographic and socioeconomic characteristics of the consumers; Store Performance Modelling understand which factors impact store performance; Headroom Estimation how much headroom exists in the market and how much growth can occur before cannibalization becomes a significant problem; Store Network Optimization determine the right combination of stores to close, expand, relocate, build, or improve.

6. BIBLIOGRAPHY

Bateson, Chris. 2019. The Impact Of Geolocation Data On Retail Marketing Strategy. 21 de May. Acedido em 11 de June de 2020. https://medium.com/@chrisbateson80/the-impact-of-geolocation-data-on-retail-marketing-strategy-871642d48a8e.

Boston Consulting Group. 2020. *Retail Footprint Optimization.* Acedido em 26 de January de 2020. https://www.bcg.com/en-pt/capabilities/big-data-advanced-analytics/retail-footprint-optimization.aspx.

Campo, Katia et al. 2000. "The impact of location factors on the attractiveness and optimal space share of product categories." *The impact of location factors on the attractiveness and optimal space share of product categories*, December: 255-279.

Dias, Mauro André. 2014. *Gestão e planeamento de stocks numa empresa de distribuição de suplementos alimentares.* Master thesis, Lisbon: Lisboa School of Economics & Management.

Dokić, Aleksa. 2018. "Regional retail sector development in Serbia - key performance and development indicators' GAP." *Regional retail sector development in Serbia - key performance and development indicators' GAP*, November: 335-356.

Dospinescu, Nicoleta et all. 2019. "A PROFITABILITY REGRESSION MODEL IN FINANCIAL COMMUNICATION OF ROMANIAN STOCK EXCHANGE'S COMPANIES." *Ecoforum.*

Dusanka, Dakic et al. 2017. "A Comparison of Contemporary Data Mining Tools." XVII International Scientific Conference on Industrial Systems. Novi Sad, Serbia: Department for Industrial Engineering and Management. 1.

Erol, Rizvan. 2017. "A new hybrid model for dynamic pricing strategies of perishable products." *A new hybrid model for dynamic pricing strategies of perishable products.* Adana: Cukurova University. 85-88.

Fox, Edward, Steven Postrel, Amanda McLaughlin. 2007. "The Impact of Retail Location on Retailer Revenues: An Empirical Investigation." *The Impact of Retail Location on Retailer Revenues: An Empirical Investigation*, February: 4 - 7.

Integrative Business & Economics. 2014. "A Topsis Model for Chain Store Location Selection." 410 - 416.

Kelly, Kathy. 2016. *Understanding Your Customers: How Demographics and Psychographics Can Help.* 16 de June. Acedido em 11 de 06 de 2020. https://extension.psu.edu/understanding-your-customers-how-demographics-and-psychographics-can-help.

Laoh, Lidya. 2019. "Dividend Payout Forecast : Multiple Linear Regression vs Genetic Algorithm-Neural Network." *CogITo Smart Journal.*

Lecturer, Furrukh et al. 2015. "Regression Analysis as a Managerial Research Tool: A Practical Illustration of Demand Analysis in Pakistan." *European Journal of Business and Management.*

Makridakis, Robin M. Hogarth and Spyros. 1981. "Forecasting and Planning: An Evaluation." *Management Science*, 2 de February: 115- 131.

Marques, Viriato. 2013. "Comparison of data mining techniques and tools for data classification." *Comparison of data mining techniques and tools for data classification.* Coimbra: Polytechnic Institute of Coimbra. 1.

Mc Kinsey & Company. 2013. "The power of advanced analytics." The power of advanced analytics. Acedido em 11 de 06 de 2020. https://www.mckinsey.com/~/media/mckinsey/dotcom/client_service/retail/articles/perspectives%20-%20winter%202013/power_of_advanced_analytics_in_revenue_management_periscope%20vf.ashx.

Mc Kinsey & Company. 2016. *Big data: The next frontier for innovation.* London: Mckinsey Global Institute.

McKone, Robert and Dan Haslehurst. 2015. *The Potential of Geolocation for Revolutionizing Retail.* 13 de November. Acedido em 11 de 06 de 2020. https://hbr.org/2015/11/the-potential-of-geolocation-for-revolutionizing-retail.

Mild, Andreas. 2006. "Retail Revenue Management." *Retail Revenue Management*, January: 90-103.

Misra, Srikant. 2012. "Importance of Statistics in Social Sciences & Research." *National Seminar on Recent trends in official statistics, At DEOARTMENT OF STATISTICS LUCKNOW UNIVERSITY LUCKNOW.* LUCKNOW: UNIVERSITY LUCKNOW.

O`Kelly, Morton E. 2009. *Applied retail location models using spatial interaction tools.* Ohio: Ohio State University.

Rigby, Darrell K and Vijay Vishwanath. 2006. "Localization: The revolution in consumer markets." *Harvard business review* 82-92; 148.

Rosu, Lucian. 2013. "GEOMARKETING -A NEW APPROACH IN DECISION MARKETING." *CASE STUDY – SHOPPING CENTRES IN IASI*, October: 123 - 132.

SAS Institute Inc. 2017. *SAS Enterprise Miner 14.3: Reference Herlp.* Cary,NC: SAS Institute Inc. 2017.

Skiera, Bernd et al. 2018. *Regression Analysis.* Frankfurt: Springer International Publishing AG.

Stubbs, Agnes Teh. 2018. *How to Use Retail Data Analysis to Boost Sales.* 24 de April. Acedido em 11 de June de 2020. https://www.softwareadvice.com/resources/retail-data-analysis-to-boost-sales/.

Turóczy, Zsuzsanna et al. 2012. "Multiple Regression Analysis of Performance Indicators in the Ceramic Industry." *Procedia Economics and Finance* 509–514.

Turóczy, Zsuzsanna, et al. 2012. "Multiple Regression Analysis of Performance Indicators in the Ceramic Industry." *Procedia Economics and Finance* 509–514.

Vu, Hien. 2018. *Inventory Management in Retail Industry - Application of Big Data Analytics.* Research, Auckland: University of Auckland.

Vyt, Dany. 2008. "Retail network performance evaluation: A DEA approach considering retailers' geomarketing." *The International Review of Retail Distribution and Consumer Research*, May: 235-253.

Wang, Jiangbo, Junhui Liu, Tiantian Li, Shuo Yin and Xinhui He. 2018. "The Research of Regression Method for Forecasting Monthly Electricity Sales Considering Coupled Multi-factor." *IOP Conference Series: Earth and Environmental Science.* Zhengzhou.

Ziliani, Cristina. 2000. "Retail micro-marketing: Strategic advance or gimmick?" *The International Review of Retail Distribution and Consumer Research* , October: 355-368.

7. ANNEXES

Table 7-1 Tasty and Sweet data

Tasty and sweet data		
Field Name	**Field Description**	**Type**
Tipo Documento	Document type code	Numeric
Documento	Document type description	Character
NumDoc	Document id	Numeric
Tipo de Operação	Operation type	Character
Vendas	Sale value	Numeric
CodVendedor	Seller id	Character
CodCliente	Client id	Character
CodArtigo	Article code	Character
CodFamília	Family product code	Character
CodZona	Zone Code	Character
Encomendas	Orders	Numeric
Margem	Margin	Numeric
Marca	Brand	Character
Unidade Saida	Unit out	Character
Categoria IEC	IEC category	Character
Taxa Álcool	Alcohol volume	Numeric
Família	Family product	Character
Subfamília Tipo Artigo	Subfamily product	Character
Subfamília2	Article type	Character
Descrição	Subfamily product 2	Character
Data	Operation date	Numeric
TotalMerc	Merchandise total	Numeric
TotalIVA	Iva total	Numeric

TotalDesc	Discount total	Numeric
TotalOutros	Others total	Numeric
Cód. Postal	Sale zip code	Character
Localidade Cód. Postal	Locality zip code	Character
Quantidade	Quantity	Numeric
Total Ilíquido	Invoice (pre-tax)	Numeric
Valor IEC	IEC value	Numeric
Total IEC (MBase)	IEC total (Mbase)	Numeric
TotalMerc (MBase)	Merchandise total (Mbase)	Numeric
TotalDesc (MBase)'	Discount total (Mbase)	Numeric
Total IEC	IEC total	Numeric
Desc. Financeiro (MBase)	Financial discount (Mbase)	Numeric
Desc. Linha (MBase)	Line discount (Mbase)	Numeric
Desconto Comercial (MBase)	Comercial discount (Mbase)	Numeric
Desconto 1 (%)	Discount 1 Percentage	Numeric
Desconto 2 (%)	Discount 2 Percentage	Numeric
Desconto 3 (%)	Discount 3 Percentage	Numeric
CP	Zip code	Character
cod_distrito	Distrito code (from zip code)	Character
dsc_distrito	Distrito description (from zip code)	Character
cod_concelho	Municipality code (from zip code)	Character
dsc_concelho	Municipality name (from zip code)	Character
NUT1	NUT1 description	Character
NUT2	NUT2 description	Character
NUT3	NUT3 description	Character

Dtcc	District code & Municipality code	Character

Table 7-2 National Statistical Bureau Information by Municipality

National Statistical Bureau Information by Municipality		
Field Name	**Field Description**	**Field Type**
NUTS_DTMN	Administrative composed boundary Nut and Municipality id	Character
DTMN	Administrative boundary Municipality id	Character
Edificios 2017	Estimates of housing stock by municipality - buildings, 2017	Numeric
Alojamentos 2017	Estimates of housing stock by municipality - accommodation, 2017	Numeric
População residente Homens 2017	Resident population by municipality and according to age groups and sex on 31/12/2017 - total male	Numeric
População residente Mulheres 2017	Resident population by municipality and according to age groups and sex on 31/12/2017 - total female	Numeric
Menos de 5 000 €	Distribution of declared gross income less individual tax income paid of tax households by municipality (less 5k), 2016	Numeric
De 5 000 a menos de 10 000 €	Distribution of declared gross income less individual tax income paid of tax households by municipality (between 5k and 10k), 2016	Numeric
De 10 000 a menos de 13 500 €	Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016	Numeric
De 13 500 a menos de 19 000 €	Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016	Numeric
De 19 000 a menos de 32 500 €	Distribution of declared gross income less individual tax income paid of tax households by municipality (between 19k and 32.5k), 2016	Numeric

32 500 € ou mais	Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016	Numeric
Agregados fiscais 2017	Declared Gross income thousands € 2017	Numeric
Agregados Rendimento bruto declarado Milhares € 2017	Declared Gross Income deducted from IRS thousands € 2017	Numeric
Agregados Rendimento bruto deduzido do IRS liquidado Milhares € 2017	Mean declared gross income by fiscal household € 2017	Numeric
Agregados Rendimento bruto declarado médio por agregado fiscal 2017 €	Taxable Persons	Numeric
Sujeitos Passivos	Taxable persons declared gross income thousands € 2017	Numeric
Sujeitos Passivos Rendimento bruto declarado Milhares € 2017	Taxable persons declared gross income deducted from IRS thousands € 2017	Numeric
Sujeitos Passivos Rendimento bruto deduzido do IRS liquidado Milhares € 2017	Taxable persons mean declared gross income by taxable person thousands € 2017	Numeric
Sujeitos Passivos Rendimento bruto declarado médio por sujeito passivo e 2017	Mean monthly earning of employees in establishments by municipality and according to level of education phd, 2016	Numeric
Ganho médio mensal dos/das trabalhadores/as por conta de outrem nos estabelecimentos por município, segundo o nível de habilitações, 2016 Doutoramento	Mean monthly earning of employees in establishments by municipality and according to level of education Master, 2016	Numeric
Ganho médio mensal dos/das trabalhadores/as por conta de outrem nos estabelecimentos por município, segundo o nível de habilitações, 2016 Mestrado	Mean monthly earning of employees in establishments by municipality and according to level of education Graduate, 2016	Numeric
Ganho médio mensal dos/das trabalhadores/as por conta de outrem nos estabelecimentos por município, segundo o nível de habilitações, 2016 Licenciatura	Mean monthly earning of employees in establishments by municipality and according to level of education Bachelor, 2016	Numeric
Ganho médio mensal dos/das trabalhadores/as por conta de outrem nos estabelecimentos por município, segundo o nível de habilitações, 2016 Bacharelato	Mean monthly earning of employees in establishments by municipality and according to level of education Secondary Education, 2016	Numeric

Ganho médio mensal dos/das trabalhadores/as por conta de outrem nos estabelecimentos por município, segundo o nível de habilitações, 2016 Ensino secundário	Declared Gross income thousands € 2017	Numeric

Table 7-3 Tasty and Sweet variables description

Variable	Role	Level
Family product code	Input	Nominal
Zone Code	Input	Nominal
Brand	Input	Nominal
NUT1	Input	Nominal
NUT2	Input	Nominal
NUT3	Input	Nominal
Zone	Input	Nominal
Year	Input	Interval
Subfamily product 2	Input	Nominal
Municipality name	Input	Nominal
Distrito name	Input	Nominal
Dtcc	Input	Nominal
Family product	Input	Nominal
Month	Input	Interval
Subfamily product	Input	Nominal
Iva total	Input	Interval
Merchandise total	Input	Interval
Discount 1	Input	Interval
Discount 2	Input	Interval
Discount 3	Input	Interval
Total Revenue	Input	Interval

Table 7-4 INE variables description

Variable	Role	Level	Code
Municipality administrative identification code	Input	Nominal	
Estimates of housing stock by municipality - buildings, 2017	Input	Interval	Ocup_01
Estimates of housing stock by municipality - accommodation, 2017	Input	Interval	Ocup_02
Resident population by municipality and according to age groups and sex on 31/12/2017 - total male	Input	Interval	Pop_01
Resident population by municipality and according to age groups and sex on 31/12/2017 - total female	Input	Interval	Pop_02

Distribution of declared gross income less individual tax income paid of tax households by municipality (less 5k), 2016	Input	Interval	Pop_19
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 5k and 10k), 2016	Input	Interval	Pop_20
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016	Input	Interval	Pop_21
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 13.5k and 19k), 2016	Input	Interval	Pop_22
Distribution of declared gross income less individual tax income paid of tax households by municipality (between 19k and 32.5k), 2016	Input	Interval	Pop_23
Distribution of declared gross income less individual tax income paid of tax households by municipality (above 32.5k), 2016	Input	Interval	Pop_24
Fiscal Household 2017	Input	Interval	Rend_01
Declared Gross income thousands € 2017	Input	Interval	Rend_02
Declared Gross Income deducted from IRS thousands € 2017	Input	Interval	Rend_03
Mean declared gross income by fiscal household € 2017	Input	Interval	Rend_04
Taxable Persons	Input	Interval	Rend_05
Taxable persons declared gross income thousands € 2017	Input	Interval	Rend_06
Taxable persons declared gross income deducted from IRS thousands € 2017	Input	Interval	Rend_07
Taxable persons mean declared gross income by taxable person thousands € 2017	Input	Interval	Rend_08
Mean monthly earning of employees in establishments by municipality and according to level of education phd, 2016	Input	Interval	Rend_09
Mean monthly earning of employees in establishments by municipality and according to level of education Master, 2016	Input	Interval	Rend_10
Mean monthly earning of employees in establishments by municipality and according to level of education Graduate, 2016	Input	Interval	Rend_11
Mean monthly earning of employees in establishments by municipality and according to level of education Bachelor, 2016	Input	Interval	Rend_12
Mean monthly earning of employees in establishments by municipality and according to level of education Secondary Education, 2016	Input	Interval	Rend_13

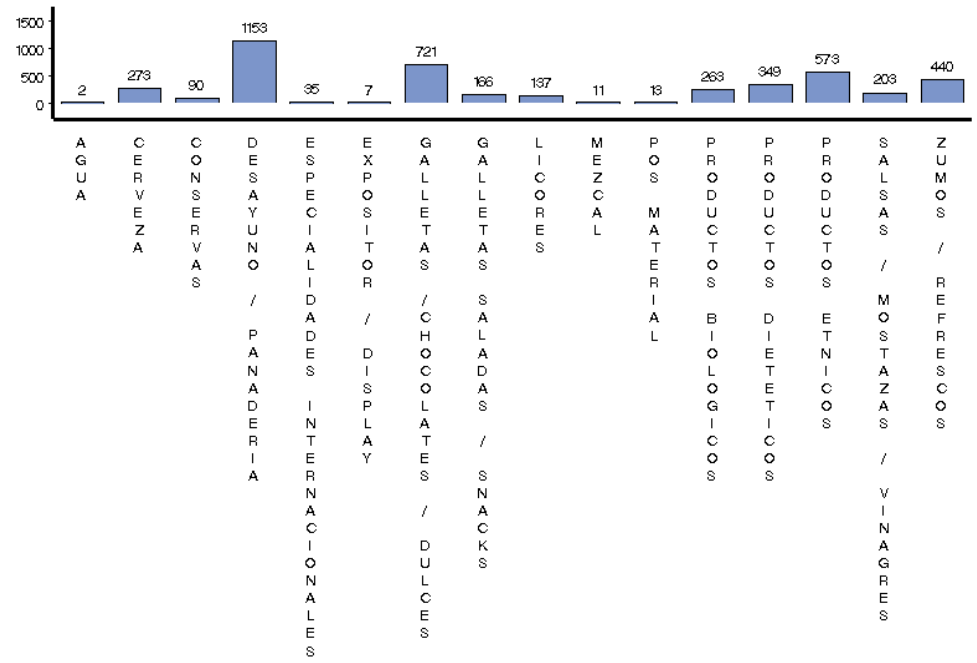

Figure 7-1 Subfamily product 2 Distribution

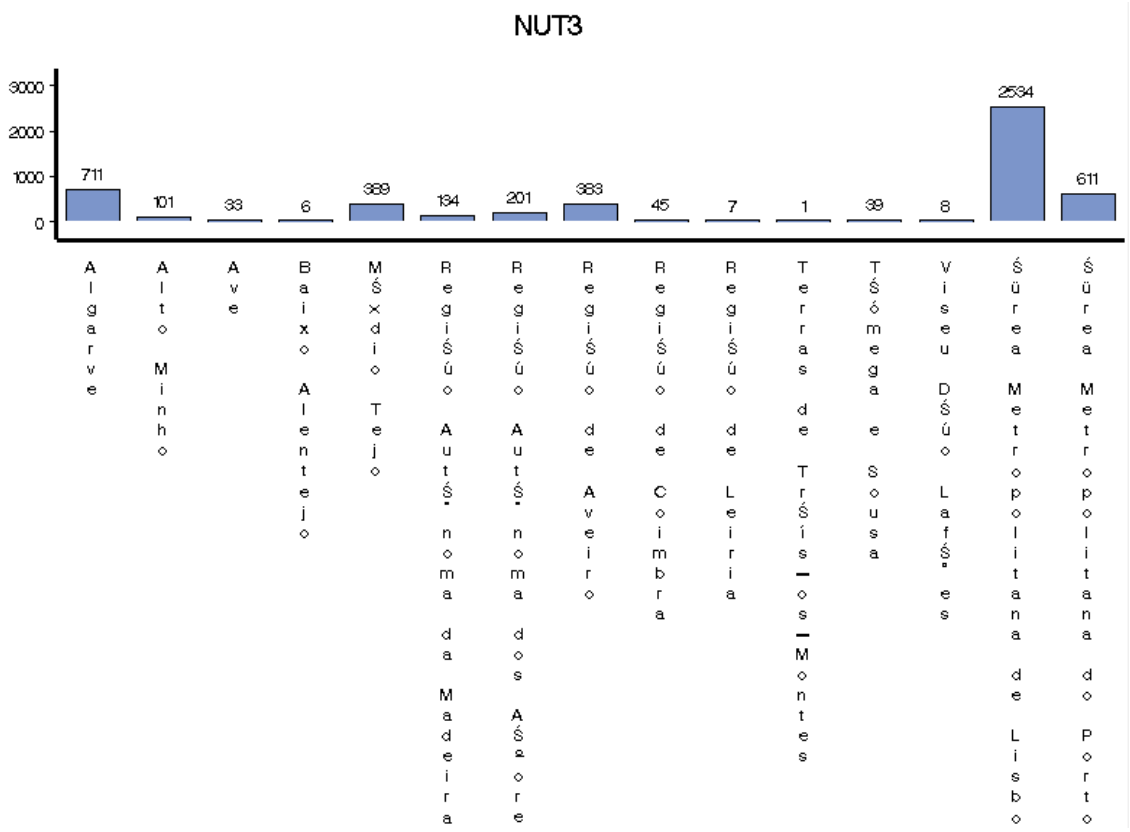

Figure 7-2 NUT3 description Distribution

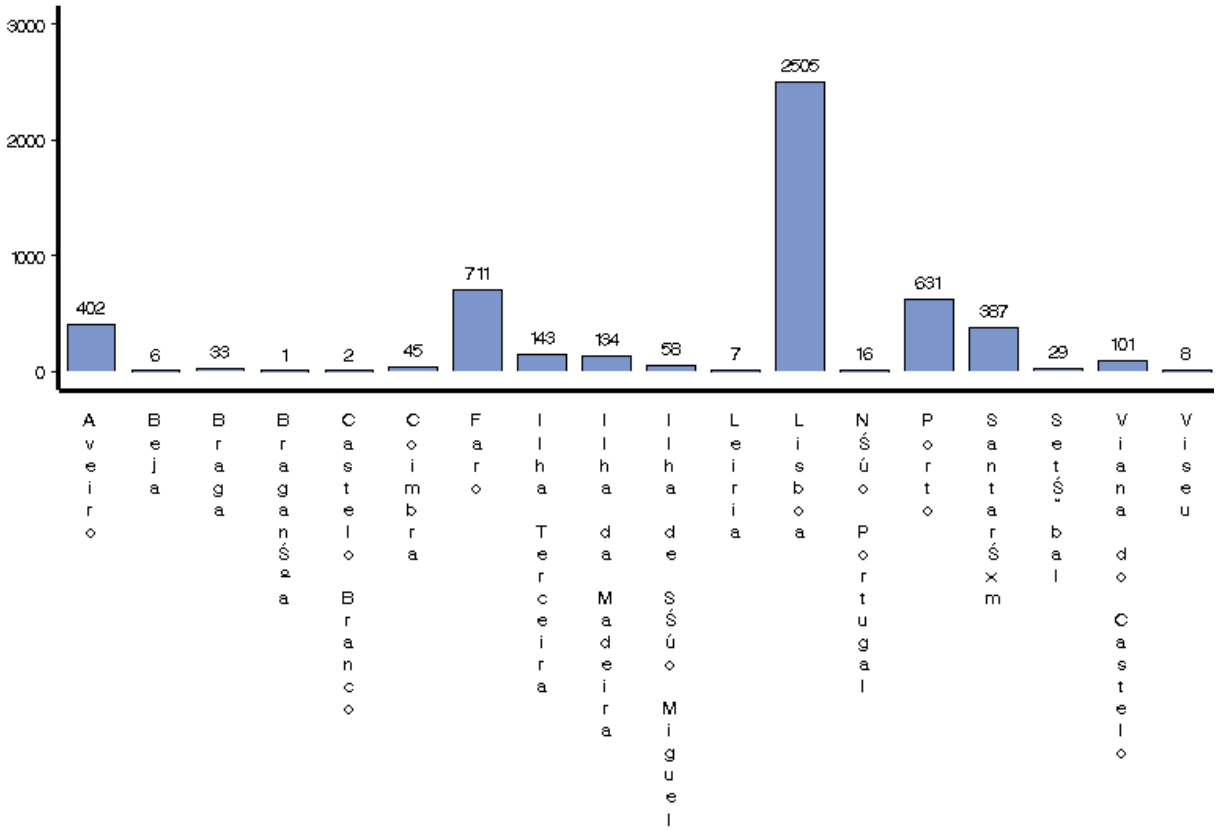

Figure 7-3 Distrito Distribution

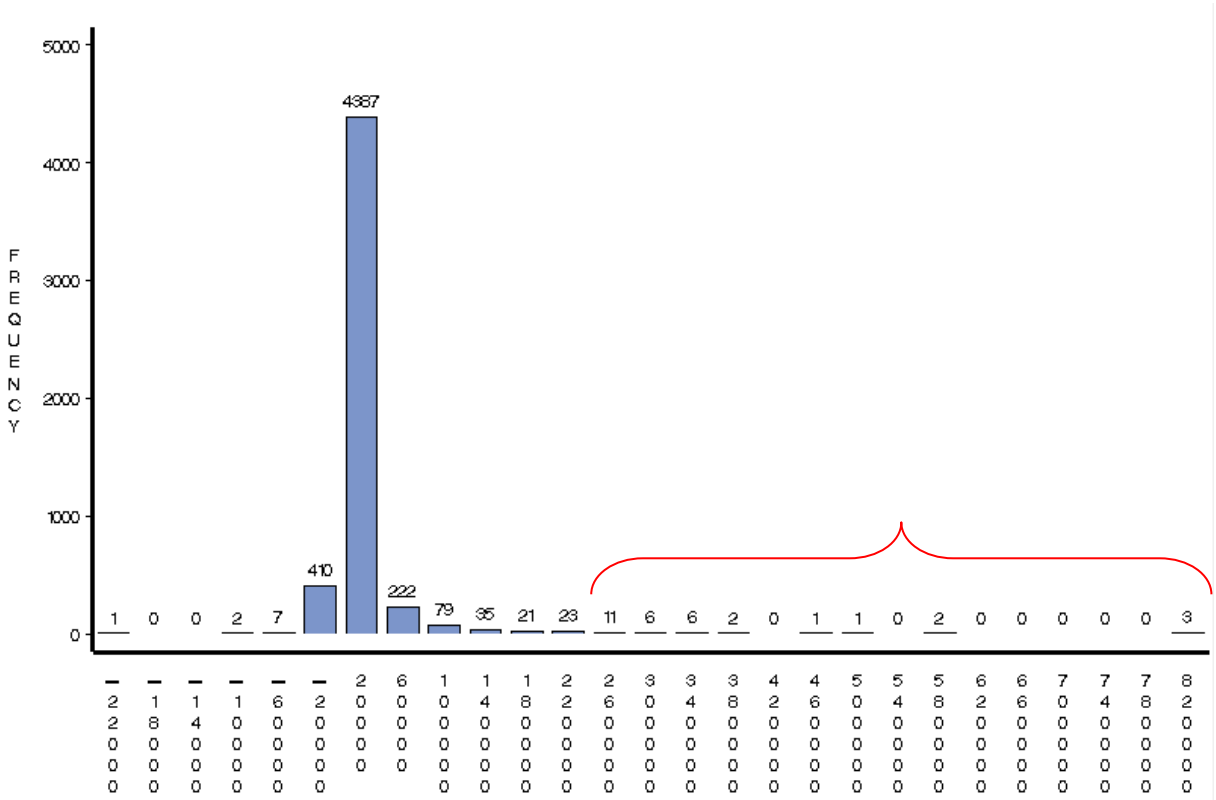

Figure 7-4 Total Revenue Distribution

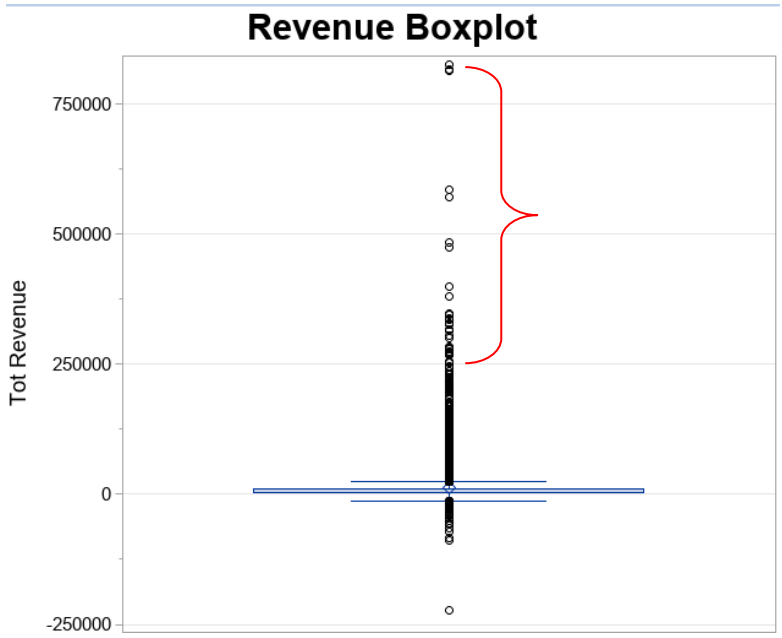

Figure 7-5 Revenue Boxplot

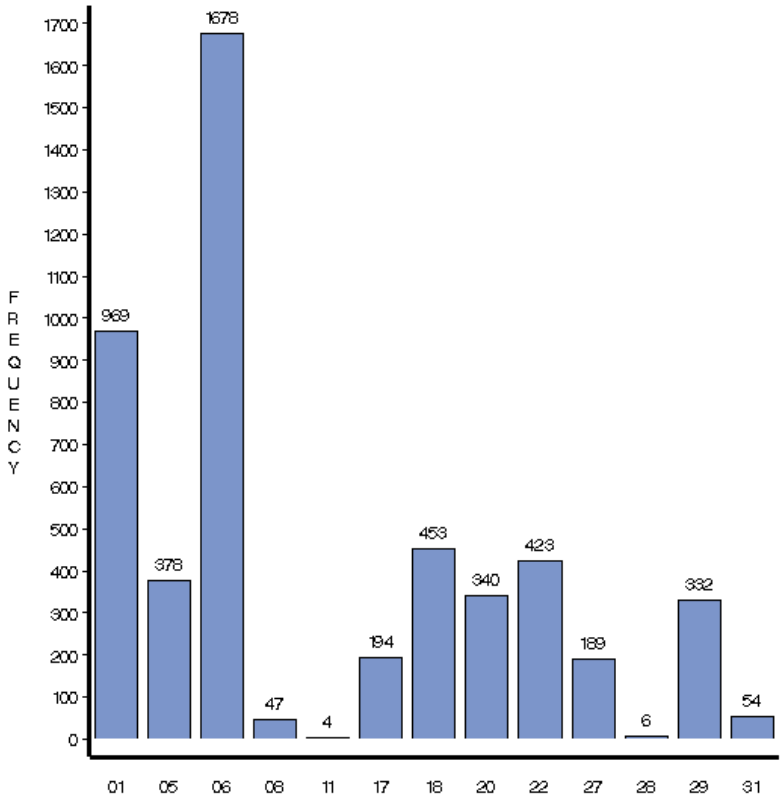

Figure 7-6 Zone Distribution

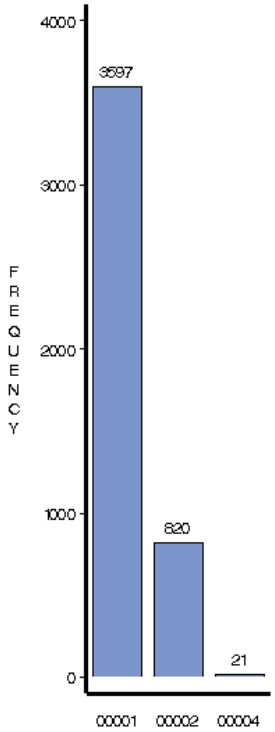

Figure 7-7 Family Product Distribution

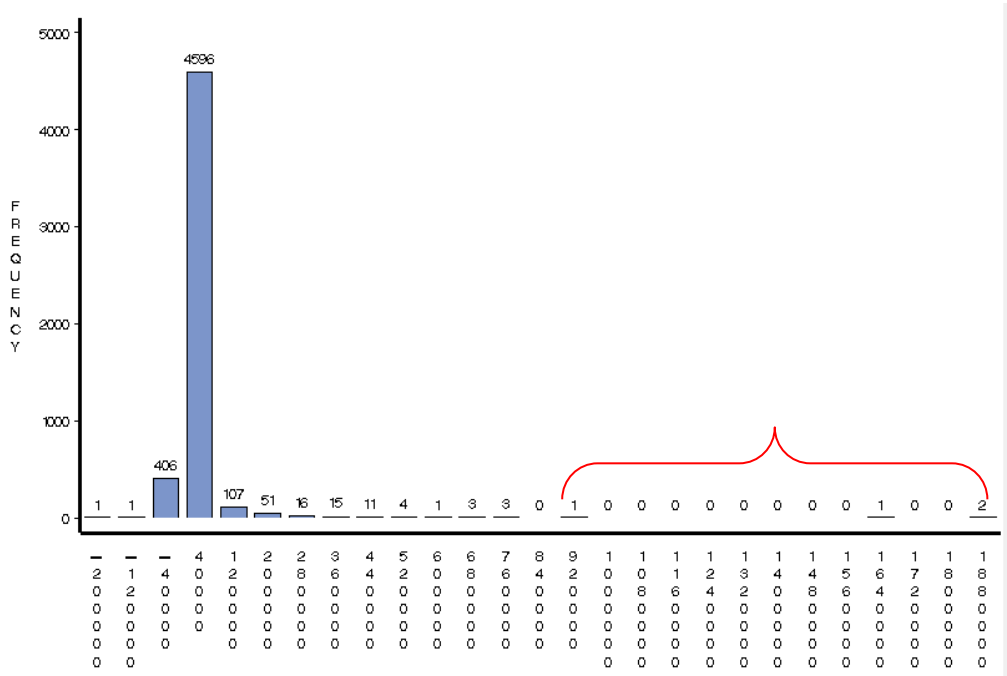

Figure 7-8 Merchandise Distribution

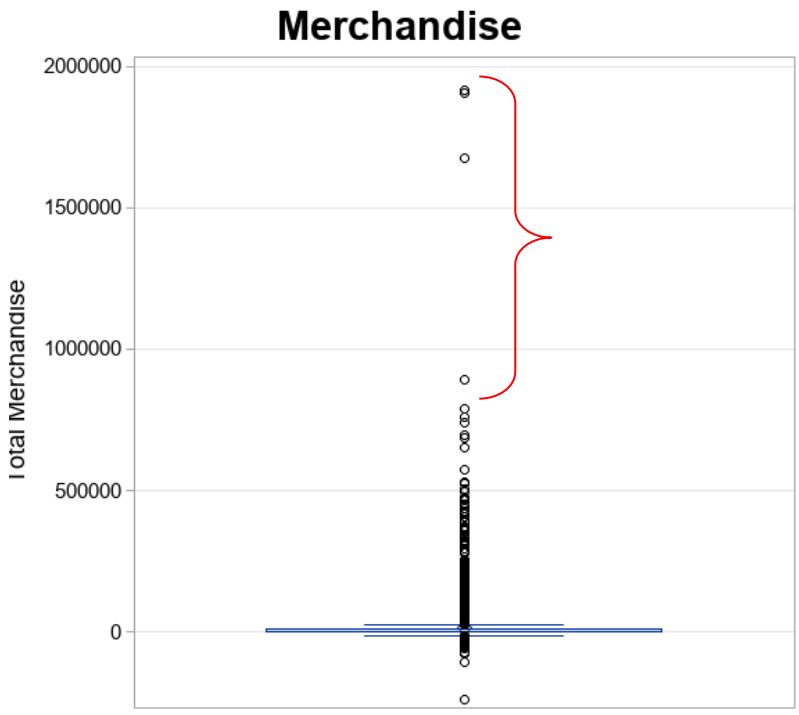

Figure 7-9 Merchandise Boxplot

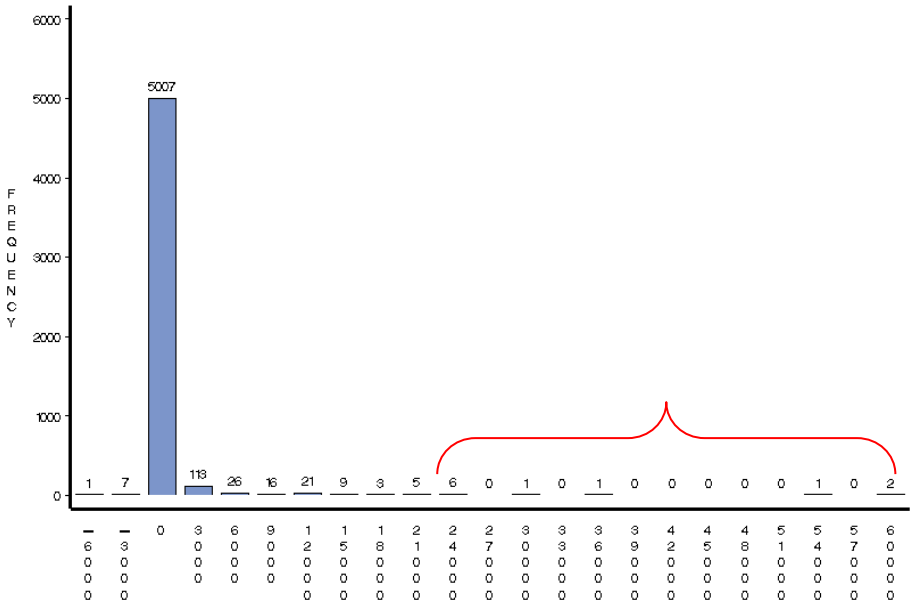

Figure 7-10 IEC Value

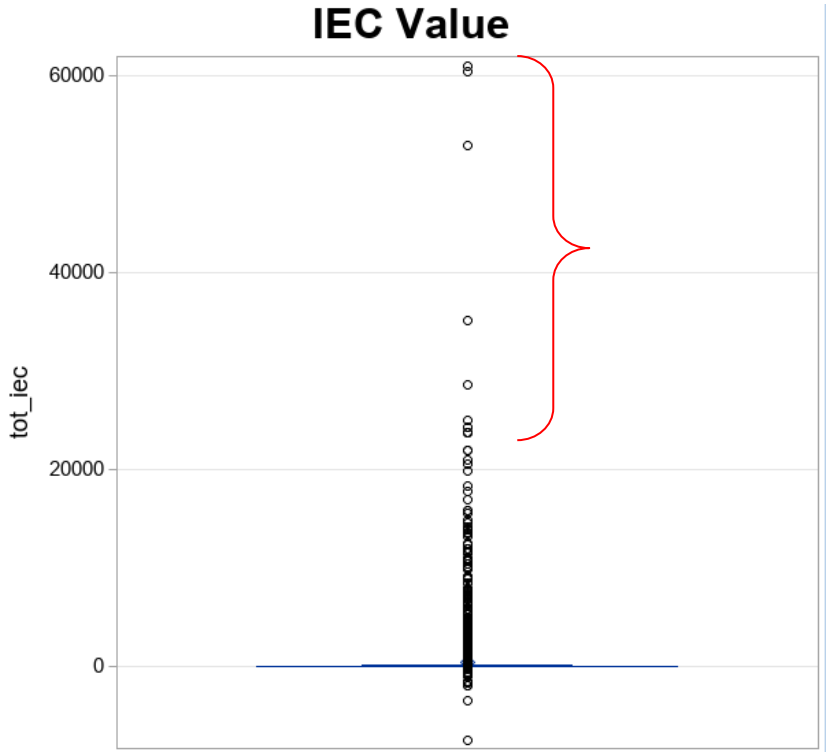

Figure 7-11 IEC Boxplot

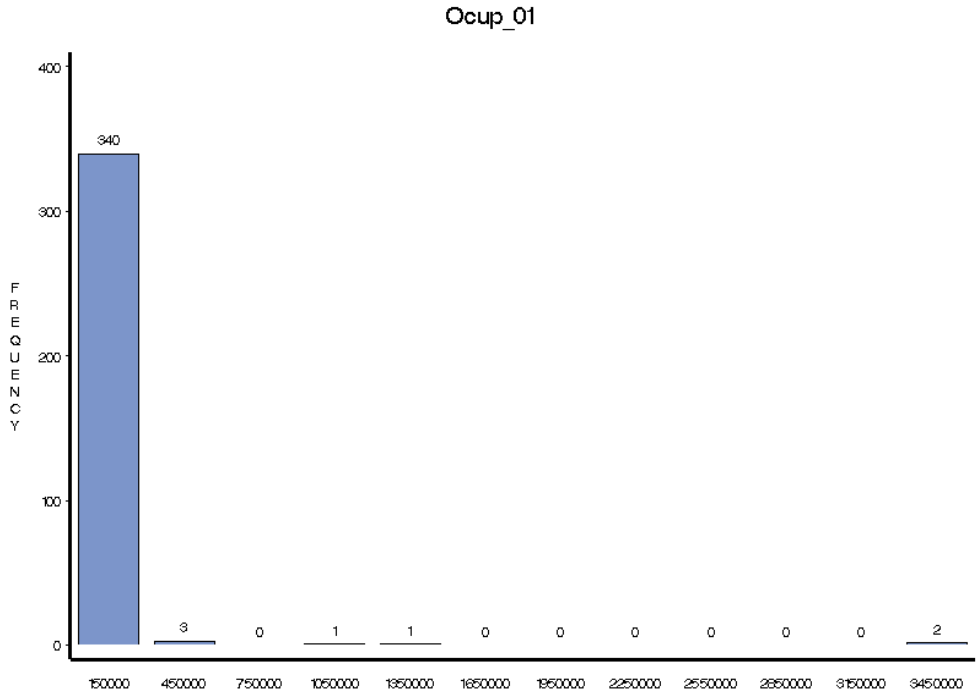

Figure 7-12 Estimates of housing stock by municipality - buildings, 2017

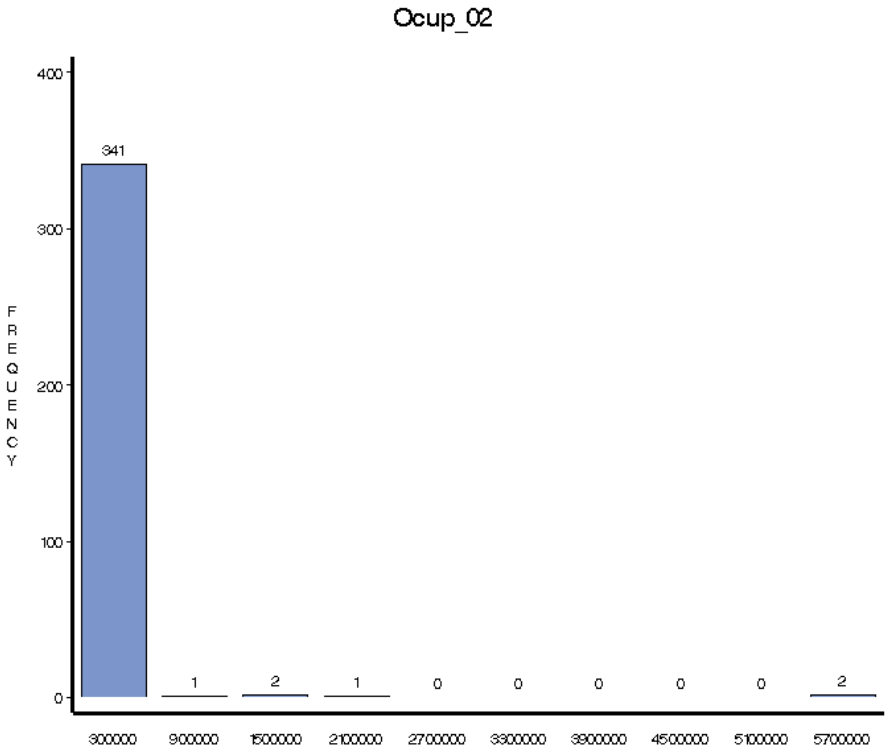

Figure 7-13 Estimates of housing stock by municipality - accommodation, 2017

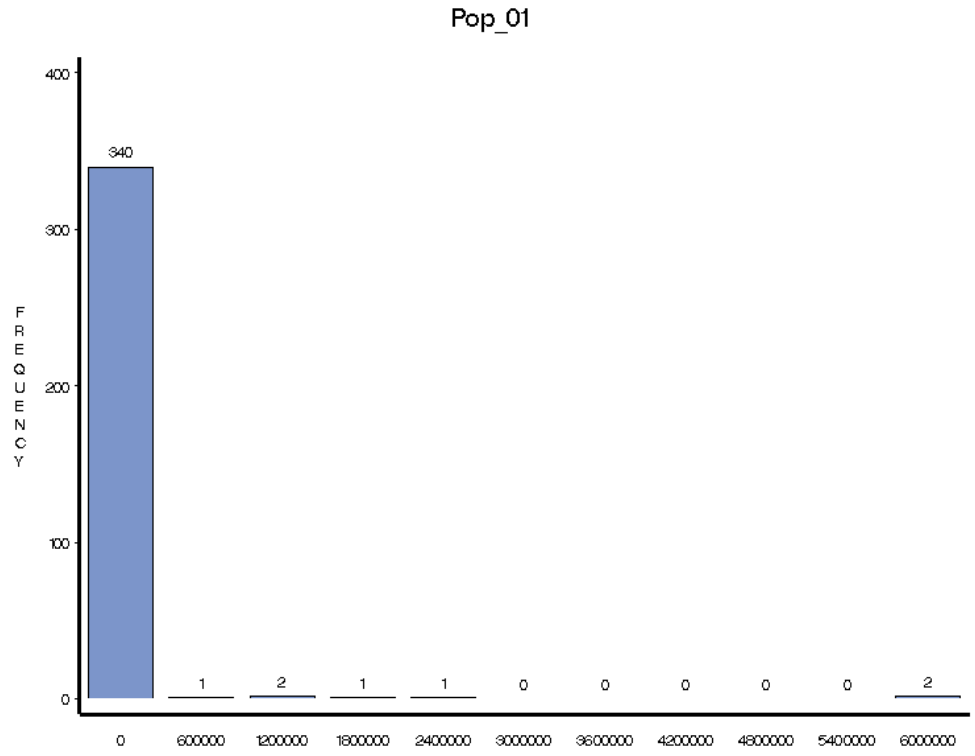

Figure 7-14 Resident population by municipality and according to age groups and sex on 31/12/2017 - total male

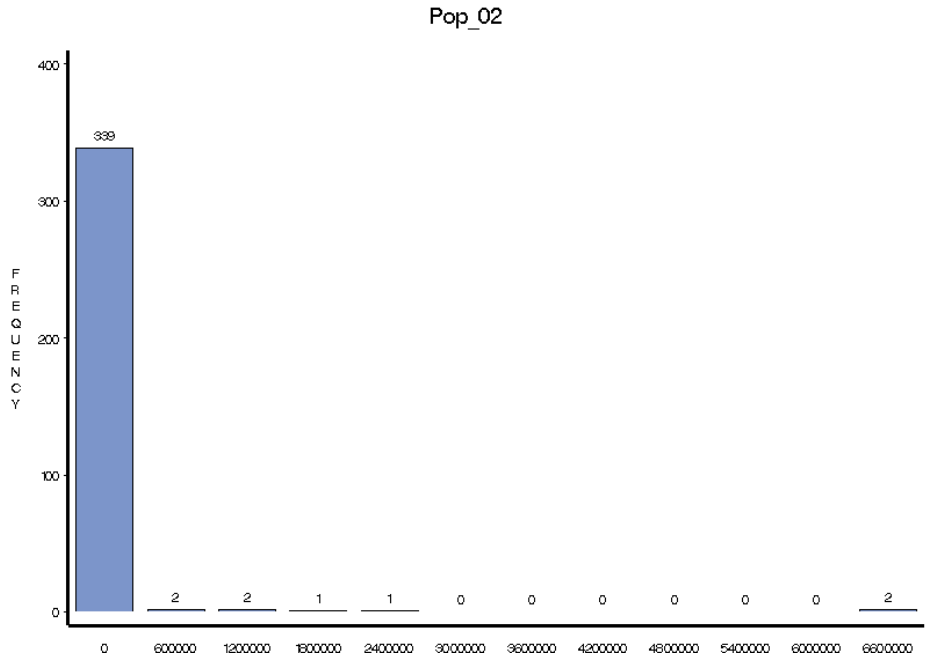

Figure 7-15 Resident population by municipality and according to age groups and sex on 31/12/2017 - total female

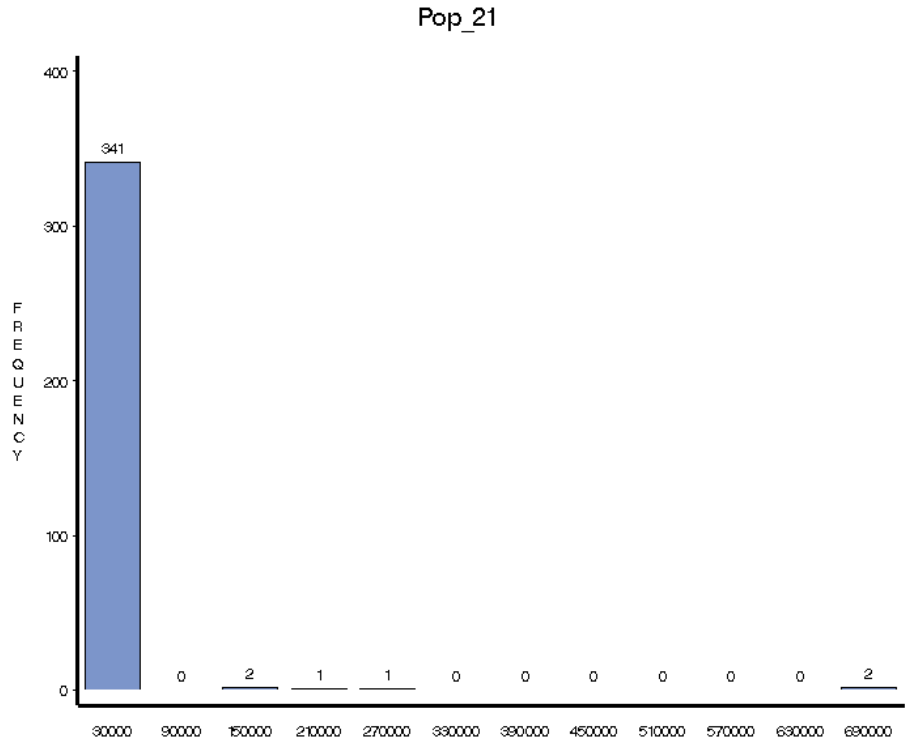

Figure 7-16 Distribution of declared gross income less individual tax income paid of tax households by municipality (between 10k and 13.5k), 2016

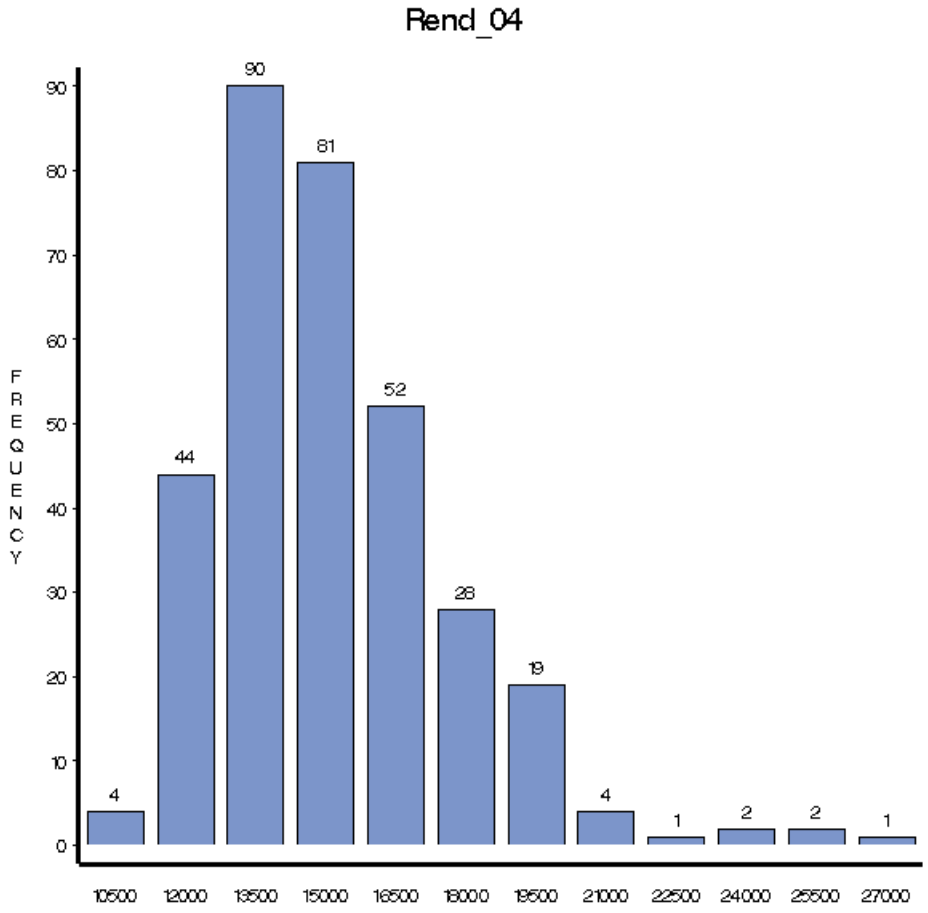

Figure 7-17 Mean declared gross income by fiscal household € 2017

Table 7-5 Variable Correlation INE Detail

Variable 1	Variable 2	Correlation Value
Ocup_01	Ocup_02	0.98
Ocup_03	Ocup_02	0.91
Ocup_03	Ocup_01	0.82
Ocup_07	Ocup_01	0.79
Ocup_07	Ocup_06	-0.87
Ocup_08	Ocup_07	0.93
Ocup_08	Ocup_01	0.79
Ocup_09	Ocup_02	0.99
Ocup_09	Ocup_01	0.97
Ocup_09	Ocup_03	0.89
Ocup_10	Ocup_09	0.96
Ocup_10	Ocup_02	0.96
Ocup_10	Ocup_03	0.95
Ocup_10	Ocup_01	0.89
Ocup_11	Ocup_02	1.00
Ocup_11	Ocup_09	0.99
Ocup_11	Ocup_01	0.98
Ocup_11	Ocup_10	0.96
Ocup_11	Ocup_03	0.89
Ocup_11	Ocup_07	0.76
Ocup_12	Ocup_02	1.00
Ocup_12	Ocup_09	0.99
Ocup_12	Ocup_01	0.98
Ocup_12	Ocup_10	0.95
Ocup_12	Ocup_03	0.89
Ocup_12	Ocup_07	0.76
Ocup_13	Ocup_03	0.98
Ocup_13	Ocup_10	0.97
Ocup_13	Ocup_02	0.91
Ocup_13	Ocup_11	0.89
Ocup_13	Ocup_09	0.89
Ocup_13	Ocup_12	0.89
Ocup_13	Ocup_01	0.81
Pop_01	Ocup_12	0.99
Pop_01	Ocup_11	0.98
Pop_01	Ocup_01	0.98
Pop_01	Ocup_02	0.98
Pop_01	Ocup_09	0.97
Pop_01	Ocup_10	0.90
Pop_01	Ocup_03	0.85

Pop_01	Ocup_13	0.83
Pop_01	Ocup_07	0.80
Pop_01	Ocup_08	0.78
Pop_02	Pop_01	1.00
Pop_02	Ocup_12	0.99
Pop_02	Ocup_11	0.99
Pop_02	Ocup_02	0.99
Pop_02	Ocup_01	0.98
Pop_02	Ocup_09	0.98
Pop_02	Ocup_10	0.92
Pop_02	Ocup_03	0.88
Pop_02	Ocup_13	0.86
Pop_02	Ocup_07	0.78
Pop_02	Ocup_08	0.76
Pop_03	Pop_01	1.00
Pop_03	Pop_02	0.99
Pop_03	Ocup_01	0.97
Pop_03	Ocup_12	0.97
Pop_03	Ocup_11	0.96
Pop_03	Ocup_02	0.96
Pop_03	Ocup_09	0.94
Pop_03	Ocup_10	0.87
Pop_03	Ocup_03	0.82
Pop_03	Ocup_07	0.82
Pop_03	Ocup_08	0.80
Pop_03	Ocup_13	0.79
Pop_04	Pop_01	1.00
Pop_04	Pop_02	0.99
Pop_04	Ocup_01	0.98
Pop_04	Ocup_12	0.97
Pop_04	Ocup_11	0.97
Pop_04	Ocup_02	0.96
Pop_04	Ocup_09	0.95
Pop_04	Ocup_10	0.87
Pop_04	Ocup_03	0.82
Pop_04	Ocup_07	0.81
Pop_04	Ocup_08	0.80
Pop_04	Ocup_13	0.79
Pop_05	Pop_02	1.00
Pop_05	Ocup_12	1.00
Pop_05	Pop_01	1.00
Pop_05	Ocup_11	1.00
Pop_05	Ocup_02	0.99
Pop_05	Ocup_01	0.99
Pop_05	Ocup_09	0.98

Pop_05	Pop_04	0.98
Pop_05	Pop_03	0.98
Pop_05	Ocup_10	0.93
Pop_05	Ocup_03	0.88
Pop_05	Ocup_13	0.87
Pop_05	Ocup_07	0.77
Pop_06	Ocup_02	1.00
Pop_06	Ocup_12	1.00
Pop_06	Ocup_11	1.00
Pop_06	Pop_05	1.00
Pop_06	Pop_02	0.99
Pop_06	Ocup_09	0.99
Pop_06	Pop_01	0.99
Pop_06	Ocup_01	0.98
Pop_06	Pop_04	0.97
Pop_06	Pop_03	0.97
Pop_06	Ocup_10	0.96
Pop_06	Ocup_03	0.91
Pop_06	Ocup_13	0.90
Pop_07	Ocup_10	0.99
Pop_07	Ocup_13	0.98
Pop_07	Ocup_03	0.95
Pop_07	Ocup_02	0.93
Pop_07	Ocup_09	0.93
Pop_07	Pop_06	0.92
Pop_07	Ocup_11	0.92
Pop_07	Ocup_12	0.91
Pop_07	Pop_05	0.89
Pop_07	Pop_02	0.88
Pop_07	Pop_01	0.85
Pop_07	Ocup_01	0.84
Pop_07	Pop_04	0.81
Pop_07	Pop_03	0.81
Pop_08	Pop_07	1.00
Pop_08	Ocup_10	0.98
Pop_08	Ocup_13	0.98
Pop_08	Ocup_03	0.95
Pop_08	Ocup_02	0.93
Pop_08	Ocup_09	0.92
Pop_08	Pop_06	0.91
Pop_08	Ocup_11	0.91
Pop_08	Ocup_12	0.91
Pop_08	Pop_05	0.88
Pop_08	Pop_02	0.87
Pop_08	Pop_01	0.85

Pop_08	Ocup_01	0.83
Pop_08	Pop_04	0.80
Pop_08	Pop_03	0.80
Pop_19	Pop_02	1.00
Pop_19	Pop_05	0.99
Pop_19	Pop_06	0.99
Pop_19	Ocup_12	0.99
Pop_19	Pop_01	0.99
Pop_19	Ocup_11	0.99
Pop_19	Ocup_02	0.99
Pop_19	Ocup_09	0.98
Pop_19	Pop_04	0.98
Pop_19	Ocup_01	0.98
Pop_19	Pop_03	0.98
Pop_19	Ocup_10	0.94
Pop_19	Ocup_03	0.90
Pop_19	Pop_07	0.90
Pop_19	Pop_08	0.89
Pop_19	Ocup_13	0.88
Pop_19	Ocup_07	0.79
Pop_19	Ocup_08	0.75
Pop_20	Pop_19	1.00
Pop_20	Pop_05	1.00
Pop_20	Pop_02	1.00
Pop_20	Pop_06	0.99
Pop_20	Ocup_12	0.99
Pop_20	Ocup_11	0.99
Pop_20	Pop_01	0.99
Pop_20	Ocup_02	0.99
Pop_20	Ocup_09	0.98
Pop_20	Ocup_01	0.98
Pop_20	Pop_04	0.98
Pop_20	Pop_03	0.98
Pop_20	Ocup_10	0.94
Pop_20	Pop_07	0.89
Pop_20	Ocup_03	0.89
Pop_20	Pop_08	0.88
Pop_20	Ocup_13	0.87
Pop_20	Ocup_07	0.78
Pop_20	Ocup_08	0.75
Pop_21	Pop_06	1.00
Pop_21	Ocup_11	1.00
Pop_21	Pop_05	1.00
Pop_21	Pop_20	1.00
Pop_21	Ocup_12	1.00

Pop_21	Ocup_02	0.99
Pop_21	Pop_19	0.99
Pop_21	Ocup_09	0.99
Pop_21	Pop_02	0.99
Pop_21	Pop_01	0.98
Pop_21	Ocup_01	0.98
Pop_21	Pop_04	0.97
Pop_21	Pop_03	0.96
Pop_21	Ocup_10	0.95
Pop_21	Pop_07	0.91
Pop_21	Pop_08	0.91
Pop_21	Ocup_03	0.90
Pop_21	Ocup_13	0.89
Pop_21	Ocup_07	0.75
Pop_22	Ocup_11	1.00
Pop_22	Pop_21	1.00
Pop_22	Ocup_12	1.00
Pop_22	Pop_05	0.99
Pop_22	Ocup_09	0.99
Pop_22	Pop_06	0.99
Pop_22	Pop_20	0.99
Pop_22	Ocup_02	0.99
Pop_22	Pop_02	0.99
Pop_22	Pop_19	0.98
Pop_22	Pop_01	0.98
Pop_22	Ocup_01	0.98
Pop_22	Pop_04	0.97
Pop_22	Pop_03	0.96
Pop_22	Ocup_10	0.95
Pop_22	Pop_07	0.91
Pop_22	Pop_08	0.90
Pop_22	Ocup_03	0.87
Pop_22	Ocup_13	0.87
Pop_22	Ocup_07	0.76
Pop_23	Pop_22	1.00
Pop_23	Ocup_11	0.99
Pop_23	Pop_21	0.99
Pop_23	Pop_05	0.99
Pop_23	Ocup_09	0.99
Pop_23	Ocup_02	0.99
Pop_23	Pop_06	0.99
Pop_23	Pop_02	0.98
Pop_23	Pop_20	0.98
Pop_23	Pop_01	0.97
Pop_23	Pop_19	0.97

Pop_23	Ocup_01	0.97
Pop_23	Pop_04	0.95
Pop_23	Pop_03	0.95
Pop_23	Ocup_10	0.95
Pop_23	Pop_07	0.91
Pop_23	Pop_08	0.91
Pop_23	Ocup_13	0.87
Pop_23	Ocup_03	0.87
Pop_24	Pop_23	0.99
Pop_24	Pop_22	0.99
Pop_24	Ocup_09	0.99
Pop_24	Ocup_11	0.98
Pop_24	Ocup_12	0.98
Pop_24	Pop_21	0.98
Pop_24	Ocup_02	0.97
Pop_24	Pop_06	0.97
Pop_24	Pop_05	0.97
Pop_24	Pop_20	0.97
Pop_24	Pop_02	0.96
Pop_24	Pop_19	0.96
Pop_24	Pop_01	0.95
Pop_24	Ocup_10	0.95
Pop_24	Ocup_01	0.95
Pop_24	Pop_04	0.93
Pop_24	Pop_03	0.93
Pop_24	Pop_07	0.92
Pop_24	Pop_08	0.91
Pop_24	Ocup_13	0.86
Pop_24	Ocup_03	0.85
Tipo_17	Pop_02	1.00
Tipo_17	Pop_05	1.00
Tipo_17	Pop_20	1.00
Tipo_17	Pop_06	1.00
Tipo_17	Pop_19	1.00
Tipo_17	Ocup_12	1.00
Tipo_17	Pop_21	0.99
Tipo_17	Ocup_11	0.99
Tipo_17	Pop_01	0.99
Tipo_17	Ocup_02	0.99
Tipo_17	Pop_22	0.99
Tipo_17	Pop_04	0.98
Tipo_17	Ocup_01	0.98
Tipo_17	Pop_23	0.98
Tipo_17	Ocup_09	0.98
Tipo_17	Pop_03	0.98

Tipo_17	Pop_24	0.97
Tipo_17	Ocup_10	0.94
Tipo_17	Pop_07	0.90
Tipo_17	Ocup_03	0.89
Tipo_17	Pop_08	0.89
Tipo_17	Ocup_13	0.88
Tipo_17	Ocup_07	0.77