THE MARKET

EVERYTHING FOR EVERYONE

DATA PRE-PROCESSING REPORT

DECEMBER 2021

Teacher: Joana Neves

Students: Ana Agostinho, r20201591 | Joana Estrompa, r20201592 | Luis Silvano, r20201479 | Pedro Sousa, r20201611

Index

Abstract

As a team of Data Pre-Processing scientists, we were challenged to both monitor *The Market*'s business and segment the customers, working through processes that transform data into valuable information. We are going to prepare the data for the next analysis team to work on, while giving some first thoughts about the business, dealing with the lack of information about their customer's shopping behaviour *The Market* is facing.

Our main goal is to perform an exploratory analysis – to answer some simple business questions – and to build an analytic-based table (ABT) – to allow the descriptive analysis (segmentation of customers). All our work will prepare the data for the next team to work on.

Introduction

A retail company as *The Market* knows that it operates in a competitive sector. It has some competitive advantage if it has the ability to know the past and to draw customer patterns and profiles and if it works that data wisely. To draw these profiles is our main focus in this project.

To make the data interpretable, we may work with it – transforming it and drawing some customers' profiles. In order to achieve this, our steps will consist on applying the methods we studied on class. We will start by running some statistical analysis on our dataset and then we will treat outliers, missing values and any incoherence we might find. We will draw and analyse the original ABT and we'll take some conclusions about the relevant variables we should keep and also include new variables that may help us understand and provide new insights on our customers.

The final goal is to end up with a table that will contain and reflect all the relevant information about each customer as well as some insights about possible profiles that we might come up after performing some data aggrupation.

Initial data treatment

To be able to perform a correct analysis and to improve our ABT, it is necessary to make sure there are no outliers or missing values - and if there are, work with them and remove them, fixing the problems they would arise if not spotted. It is also needed to have the data giving us a coherent information about the customers and to form groups of categories to provide a better interpretation of our table.

Outliers

Definition: An outlier is a value from a dataset that is considerably far from the other values.

After the initial statistics analysis, we jumped into the outliers treatment, where we removed from our dataset 22 outliers, which represents about 0,44% of our dataset. They were the following: we had some transactional dates from the 80's or from years close to 3000 (these last could have been registering errors); we also found some ratings of 10000 (they should be between 0 and 10).

Missing Values

Definition: When one or more values should be available for analysis but for some reason they aren't.

To treat our missing values we used the *Tree* imputation method - method that uses a trained decision tree with all the other values to determine our missing value; we also deleted the Nationality variable because about 96,5% of its data was missing.

Data Coherence

Definition: data coherence counts with uniformity across data from various sources, as well as logical connections within a single dataset or across different datasets.

Then we worked on data coherence and we concluded that some things didn't make sense:

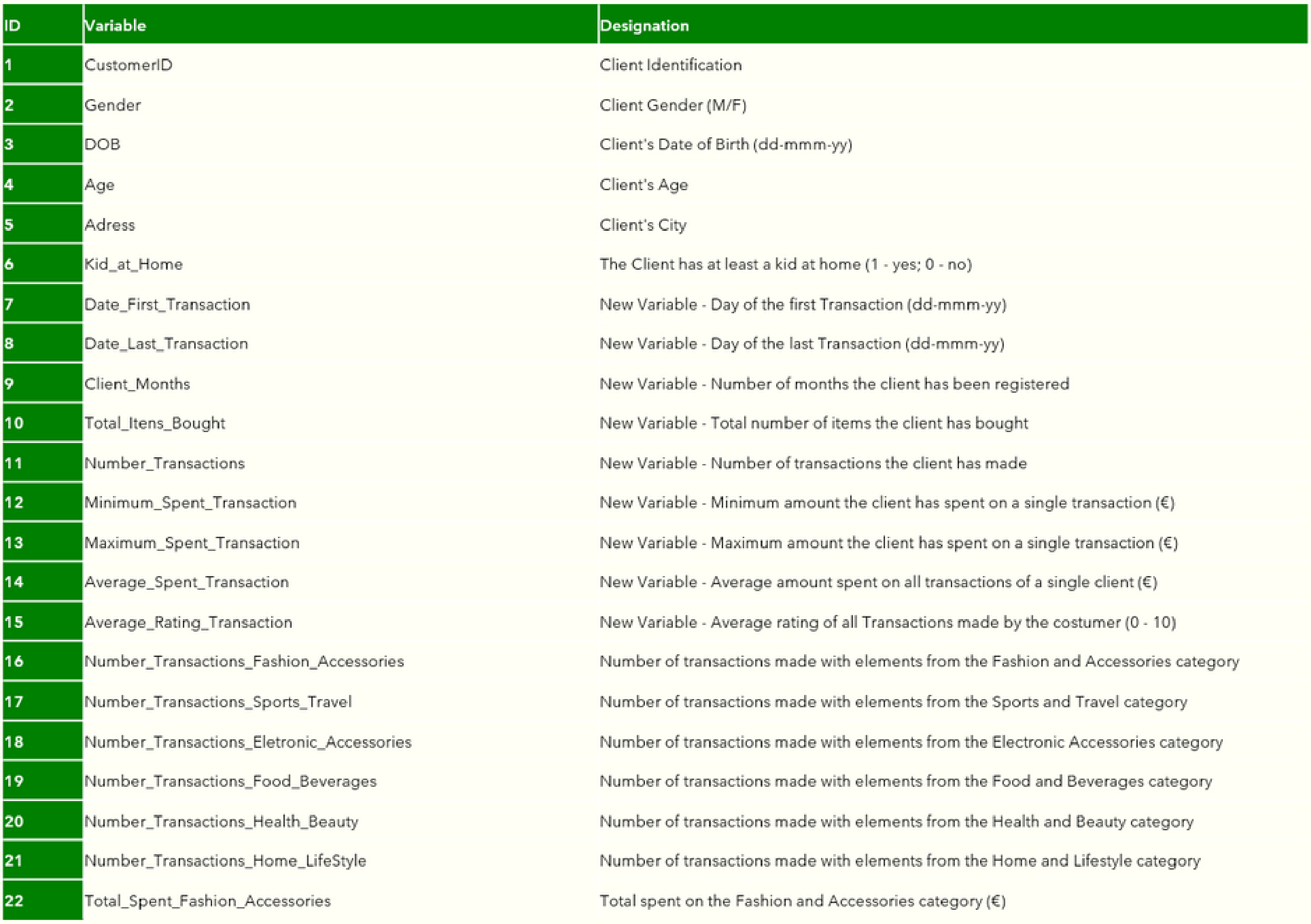
- Gender being different from F or M;

- Kidhome being different from 0 or 1;
- Rating being less than 0 or more than 10;
- Calculations being incorrect:
  - If price per unit times quantity is different than COGS;
  - If COGS plus taxes is different than total;
- If payment methods are different than the established;

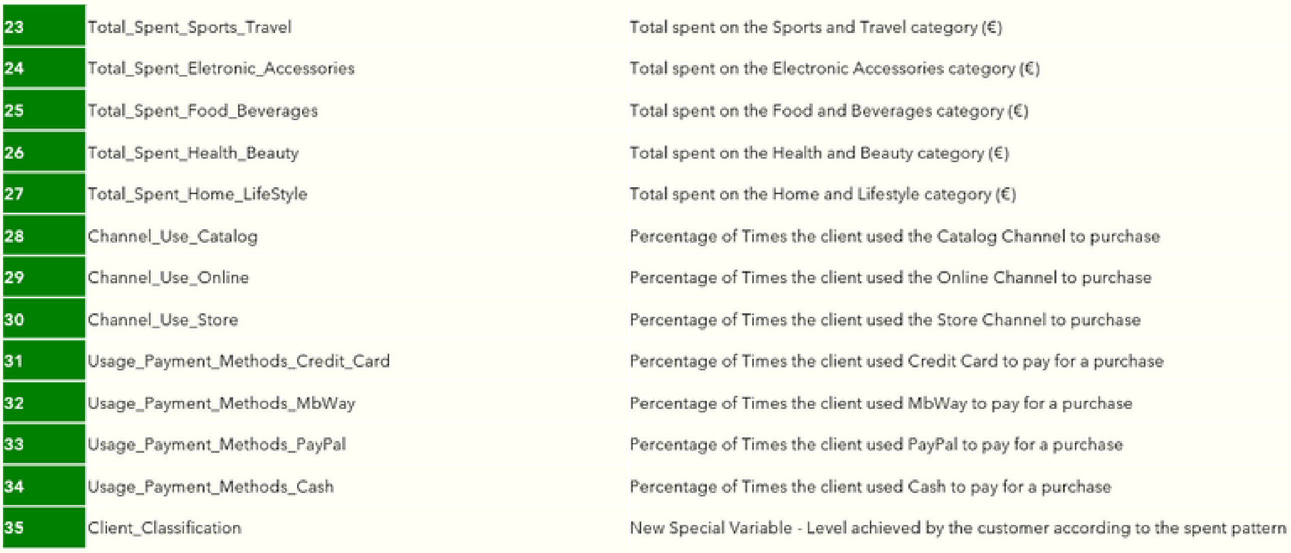
Customer Signature Tabe

After the inicial analysis and data correction, we then proceeded to the building of our Customer Signature Table. We felt the need to create some new variables to help us draw some conclusions.

Follows a table with all the variables in our Customer Signature Table and a brief description:



ID	Variable	Designation
1	CustomerID	Client Identification
2	Gender	Client Gender (M/F)
3	DOB	Client's Date of Birth (dd-mmm-yy)
4	Age	Client's Age
5	Adress	Client's City
6	Kid_at_Home	The Client has at least a kid at home (1 - yes; 0 - no)
7	Date_First_Transaction	New Variable - Day of the first Transaction (dd-mmm-yy)
8	Date_Last_Transaction	New Variable - Day of the last Transaction (dd-mmm-yy)
9	Client_Months	New Variable - Number of months the client has been registered
10	Total_Itens_Bought	New Variable - Total number of items the client has bought
11	Number_Transactions	New Variable - Number of transactions the client has made
12	Minimum_Spent_Transaction	New Variable - Minimum amount the client has spent on a single transaction (€)
13	Maximum_Spent_Transaction	New Variable - Maximum amount the client has spent on a single transaction (€)
14	Average_Spent_Transaction	New Variable - Average amount spent on all transactions of a single client (€)
15	Average_Rating_Transaction	New Variable - Average rating of all Transactions made by the costumer (0 - 10)
16	Number_Transactions_Fashion_Accessories	Number of transactions made with elements from the Fashion and Accessories category
17	Number_Transactions_Sports_Travel	Number of transactions made with elements from the Sports and Travel category
18	Number_Transactions_Eletronic_Accessories	Number of transactions made with elements from the Electronic Accessories category
19	Number_Transactions_Food_Beverages	Number of transactions made with elements from the Food and Beverages category
20	Number_Transactions_Health_Beauty	Number of transactions made with elements from the Health and Beauty category
21	Number_Transactions_Home_LifeStyle	Number of transactions made with elements from the Home and Lifestyle category
22	Total_Spent_Fashion_Accessories	Total spent on the Fashion and Accessories category (€)



23	Total_Spent_Sports_Travel	Total spent on the Sports and Travel category (€)
24	Total_Spent_Eletronic_Accessories	Total spent on the Electronic Accessories category (€)
25	Total_Spent_Food_Beverages	Total spent on the Food and Beverages category (€)
26	Total_Spent_Health_Beauty	Total spent on the Health and Beauty category (€)
27	Total_Spent_Home_LifeStyle	Total spent on the Home and Lifestyle category (€)
28	Channel_Use_Catalog	Percentage of Times the client used the Catalog Channel to purchase
29	Channel_Use_Online	Percentage of Times the client used the Online Channel to purchase
30	Channel_Use_Store	Percentage of Times the client used the Store Channel to purchase
31	Usage_Payment_Methods_Credit_Card	Percentage of Times the client used Credit Card to pay for a purchase
32	Usage_Payment_Methods_MbWay	Percentage of Times the client used MbWay to pay for a purchase
33	Usage_Payment_Methods_PayPal	Percentage of Times the client used PayPal to pay for a purchase
34	Usage_Payment_Methods_Cash	Percentage of Times the client used Cash to pay for a purchase
35	Client_Classification	New Special Variable - Level achieved by the customer according to the spent pattern

We were provided with the information about *The Market*'s clients and there was work done in order to build a Customer Signature Table. This was the table we used to perform our analysis ever since it was built. There are some insights that might need to be explored for a better understanding of this CST.

When building variables 8, 9 and 10, we worked with the given variable *trans_date* from the provided information. Then, we selected (from each customer) the date of the first and of the last transaction registered - variable 8 and 9, respectively - and, as a last step, we performed some calculations to get the number of months the customer has been registered - variable 10.

To build variable 11 we simply summed all the items in every transaction that each customer has ever made and in order to build variable 12, we computed the number of transactions registered by each customer.

Variables 13 and 14 are self explanative.

Variable 15 computes the average amount of money spent on every transactions of each customer (in €) and variable 16 computes the average classification each customer gave to the items he/she bought.

Variables from 17 to 22 gather together and give us an insight about the number of transactions of each one of the categories of products that each customer has made. Variables from 23 to 28 compute the total spent on each category of products by each of the customers.

Variable 35 is a Special New Category that we created in order to group clients according to their spending patterns:

- Silver - clients in one of these situations:
    - with one single transaction;
    - with an average of spending's times number of transactions smaller than 1500€;
    - with two transactions and an average of spending times number of transactions smaller than 2500€;

- Gold - clients in one of these situations:
    - with two transactions and an average of spending times number of transactions equal or bigger than 2500€;
    - with three transactions and an average of spending times number of transactions smaller than 4000€;
    - with four transactions and an average of spending times number of transactions smaller than 3000€;
    - any other situation not specified;

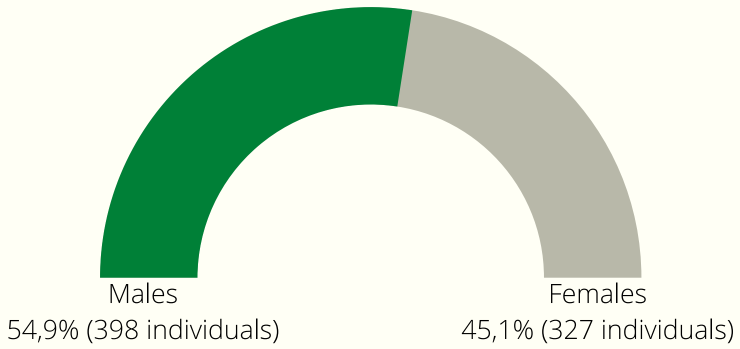
- Diamond - clients is one of these situations:
    - with three or more transactions and an average of spending times number of transactions equal or bigger than 4000€;
    - with five or more transactions and an average of spending times number of transactions equal or bigger than 3000€.

Explained all this, we found need in the building of some visual tools in order to help us understand each main variable and niche of The Market's situation, which would give us sue guidance throughout the rest of our project.

All the graphs showed from now on were built using information of our Customer Signature Table and of our new ABT - with the new variables included.

Analysis of the variables through visual elements

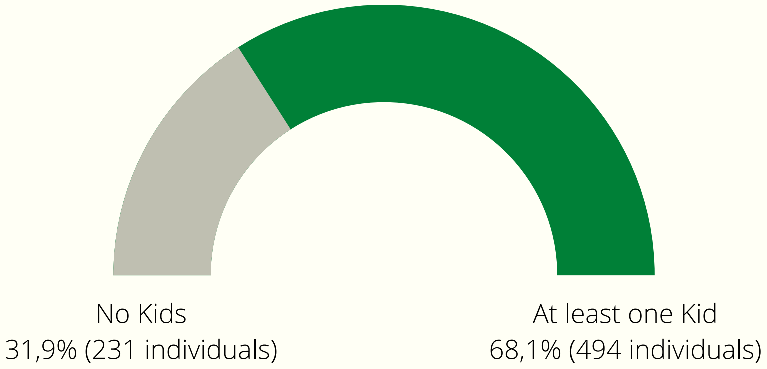
graph 1 - **Customers by Gender**



We started by working with the understanding of the division of The Market's clients by gender.

There are not many important conclusions to draw with this graph, but it is interesting to have an insight of our customer's distribution.

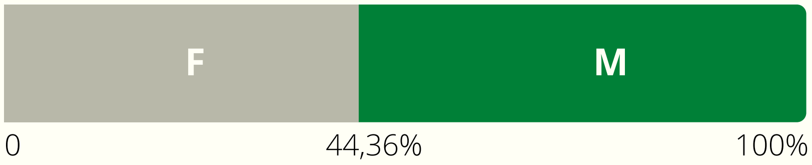
graph 2 - **Kids at Home**



Then we found important to understand the household composition of our customers and, for doing that, we used the *Kid_at_Home* variable (variable 6). We can, then, find clients in one of two situations: with *No Kids* or with *At least one Kid*.

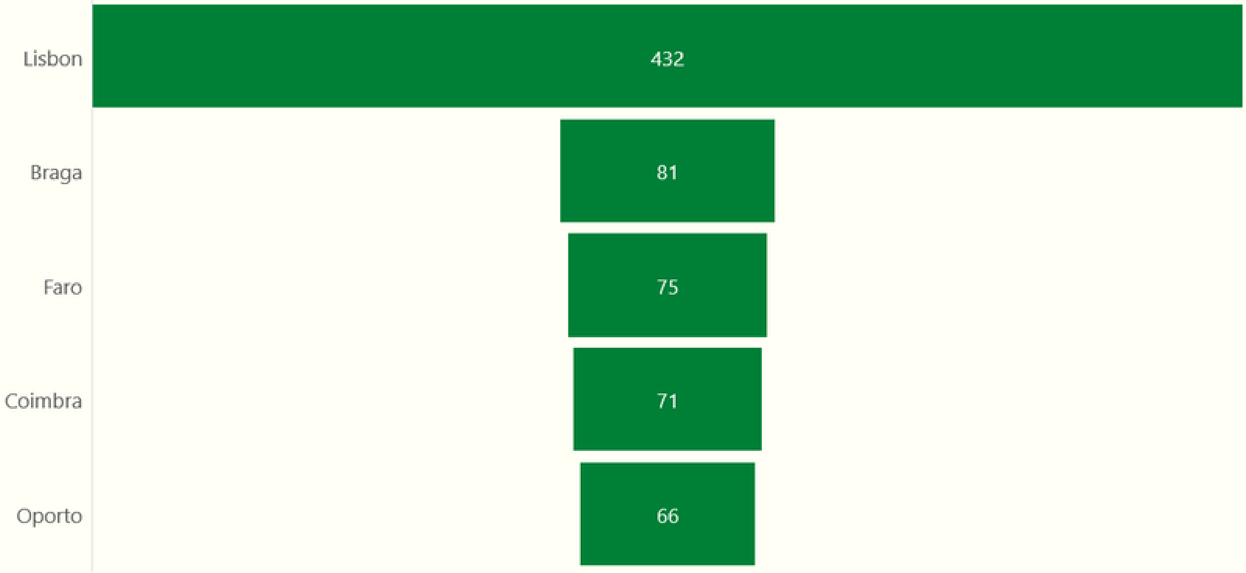
This variable might be important to draw some consume patterns, once it is known that families with kids tend to spend more in some specific categories.

graph 3 - **Spending per Gender on % of Total Spent**



The spending amount per gender can indicate us which gender represents the most valuable target market. However, the difference in *The Market* is not so relevant, so we proceeded with the analysis only keeping this graph in mind if needed in the future.

graph 4 - **Distribution of Clients per City**

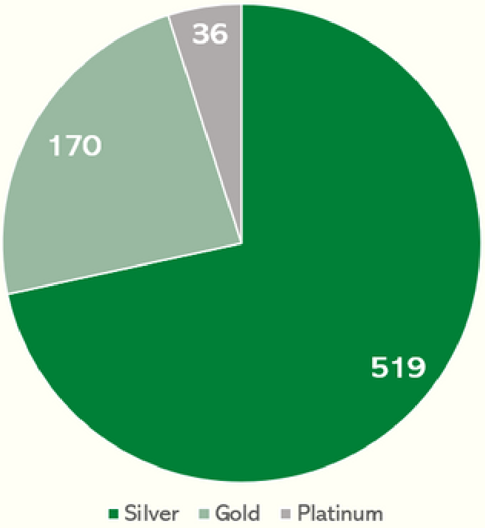


We then built a funnel graph to help us understand the origin of our clients. With this variable, there was not the need to create any agroupations, since we were working with only 5 cities.

We were able to realize that Lisbon is the town of more than half of all The Market's clients. It is also possible to understand that it holds more than 5 times more customers than any other city.

This is an important information when deciding local campaigns, for example.

graph 5 - **Classification of The Market Clients**



After performing the initial analysis of The Market clients according to their personal information, we then started the same process of analysing and grouping clients according to and considering their shopping patterns.

Therefore, we felt the need to understand if the profits of The Market are from a low volume of customers who spend large amounts of money or from a large volume of clients who spend little in The Market - or a middle term of both extremes.

So, we proceeded with (as explained on previous steps of our report) the Classification of the clients according to their spending patterns (repeating):

- Silver - clients in one of these situations:
  - with one single transaction;
  - with an average of spending's times number of transactions smaller than 1500€;
  - with two transactions and an average of spending times number of transactions smaller than 2500€;

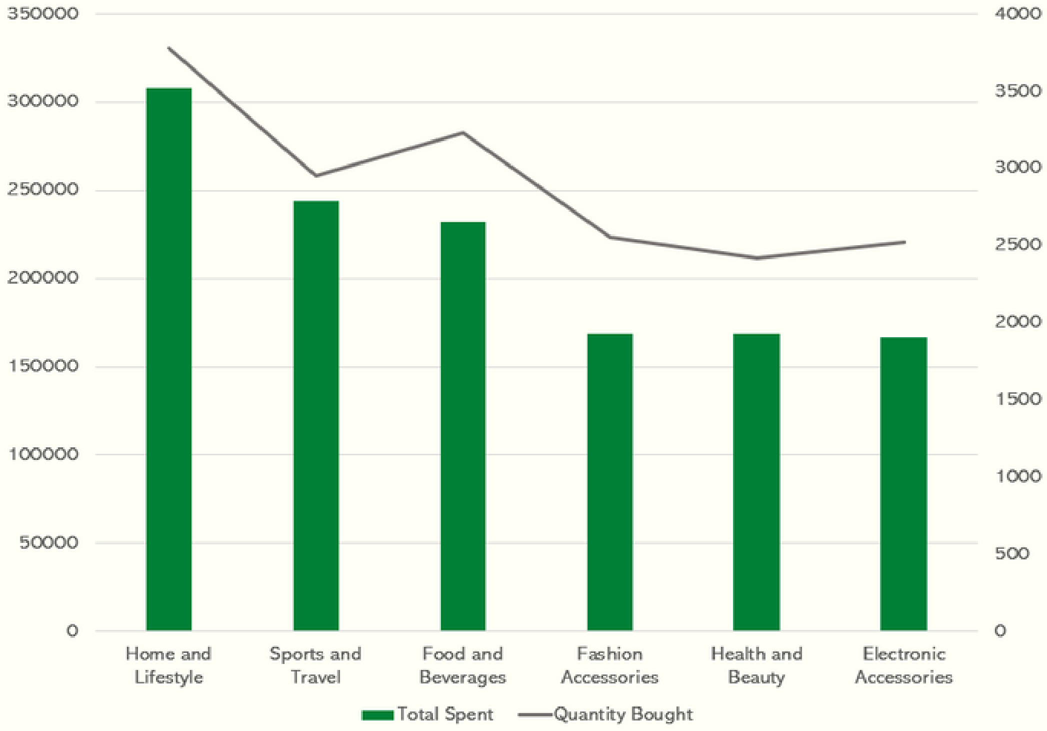
- Gold - clients in one of these situations:
  - with two transactions and an average of spending times number of transactions equal or bigger than 2500€;
  - with three transactions and an average of spending times number of transactions smaller than 4000€;
  - with four transactions and an average of spending times number of transactions smaller than 3000€;
  - any other situation not specified;

- Diamond - clients is one of these situations:
  - with three or more transactions and an average of spending times number of transactions equal or bigger than 4000€;
  - with five or more transactions and an average of spending times number of transactions equal or bigger than 3000€.

After creating these groups, we got the following results: 519 of the 707 registered clients are Silver Clients, representing 71,59% of *The Market* overall customer portfolio; 170 registered clients are Gold Customers, representing 23,45% of the total customers of *The Market*; finally, 36 clients are Diamond Customers, which represents only 4,97% of *The Market* clients.

So, answering the question we made earlier, we can conclude that the majority of *The Market* sales are made to a large amount of clients who show low and small spending and consuming patterns. This information is very relevant when deciding which type of clients to target in future campaigns.

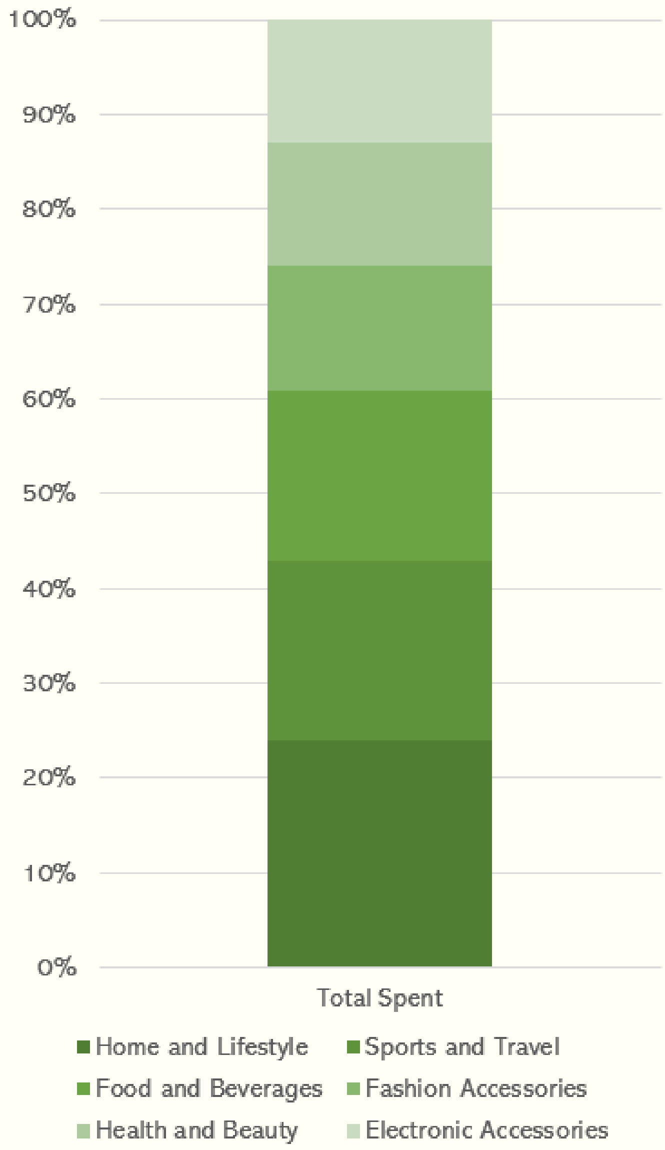
graph 6 - **Total Spent and Quantity Bought per Category**



Analysing the sales of *The Market*, we started with a columns and lines graph, who measured the Total Amount Spent and Quantity Bought per category.

This helped us understand, for example, that the category with highest sales volume is also the category where customers spent more money (Home and Lifestyle).

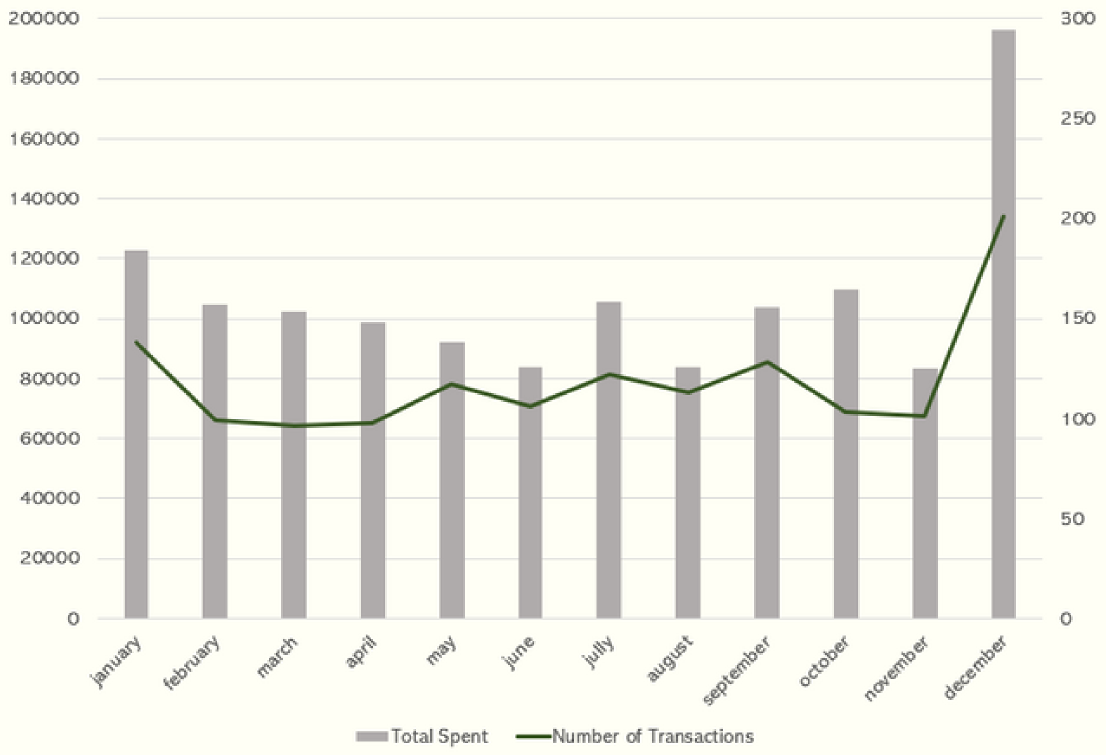
graph 7 - **Total Spent per Category**



Following the same thinking line, we then proceeded to analyse the contribution of each category to the total of sales of *The Market* with the construction of a pilled column graph.

Since we had already analysed the *graph 6 - Total Spent and Quantity Bought per Category*, there are no surprises about what category contributes more to the total of Sales, but we managed to understand that the variable Home and Lifestyle represents a little bit more than 20% of the total of Sales.

graph 8 - **Total Spent and Number of Transactions per Month**



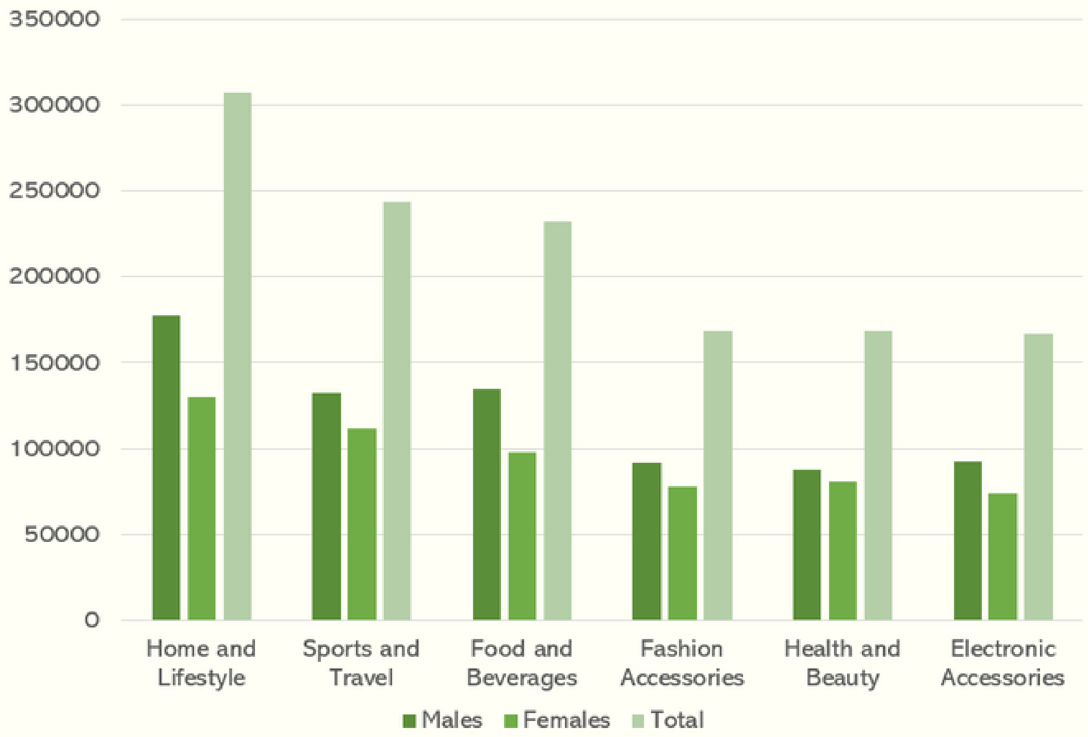
We found interesting to understand the monthly fluctuation of *The Market*'s business, and for doing so, we built a combined graph of columns and a line who indicates us the total Spent and the amount of transactions of each month, respectively.

The weakest months when considering Total Sold were June and November, both selling just a little over 83K monetary units. However, the months that registered the owest volume of Transactions were February, March and April, with 99, 96 and 98 registered transactions, respectively.

We didn't find any explanation for this variables, since we would need some more details about the company's campaigns over the year.

There is also a consideration to do on the best month of activity: December. This month registered 201 transactions and sold over 196K monetary units, standing out from the others. This may be explained by the Christmas Time, which is a good sales moment (for most of the businesses like *The Market*).

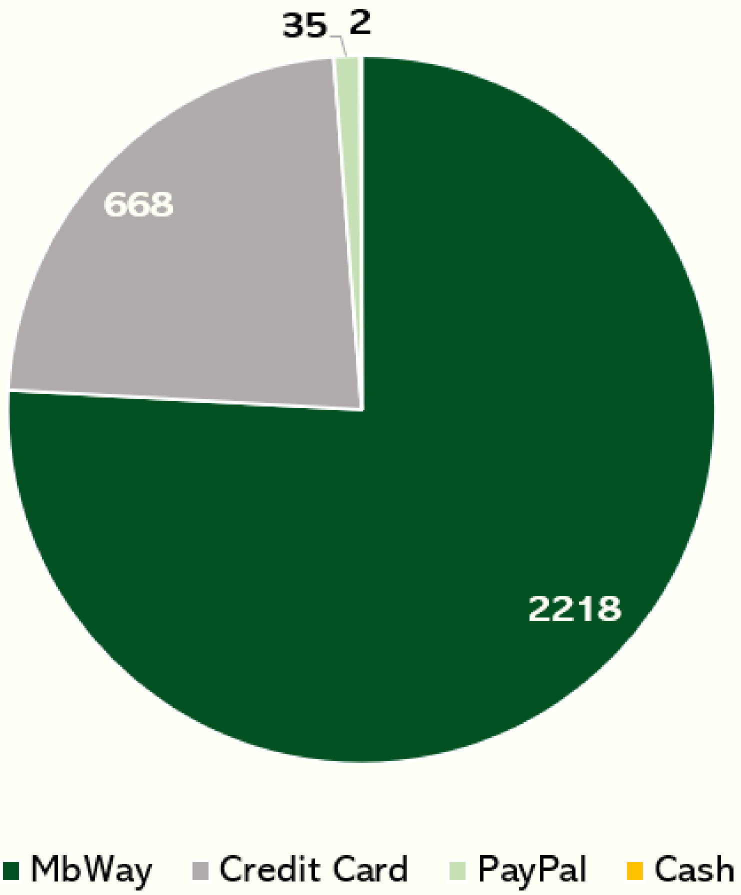
graph 9 - **Total Spent per Category by Gender**



When performing an analysis a little bit more detailed, we started with the analysis of the Total Spent per Category by Gender. We were able to confirm that in all Categories, Males spend more than Females. The results were not surprising, since we already had seen that the Male gender spends more than the Female on *The Market*.

We were able to reinforce the note that was made when analysing *graph 3 - Percentage of Total Spent per Gender*, when deciding future campaigns' targets.

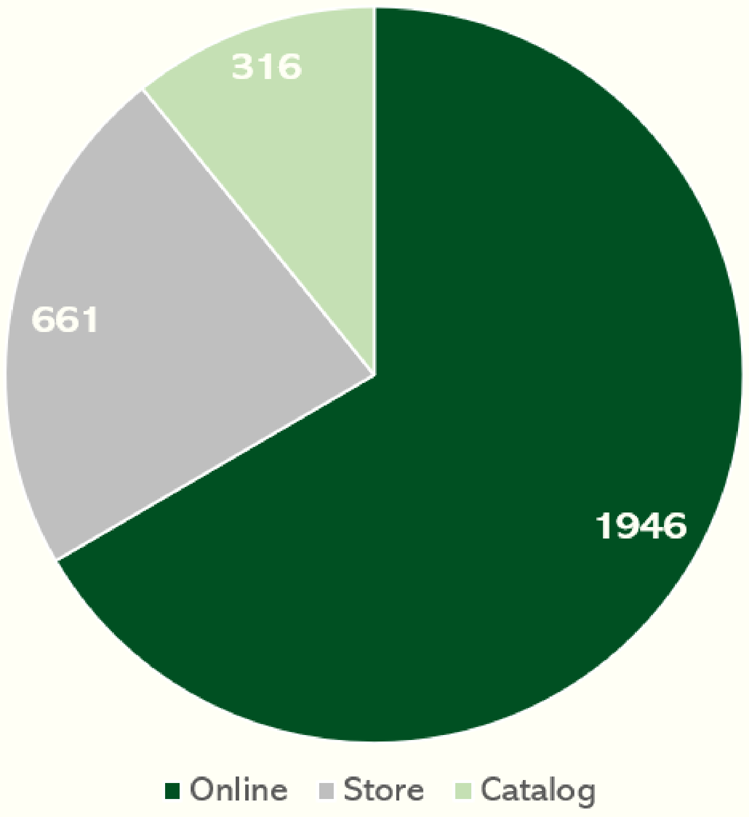
graph 10 - **Number of Transactions per Payment Method**



When analysing the *graph 10 - Number of Transactions per Payment Method,* we concluded that the most used payment method is MbWay - with 2218 transactions - and the payment method with less usage is Cash - with only 2 transactions.

So, we can conclude the clients prefer use MbWay, than the other types of payment (Cash, Credit Card our Paypal).

graph 11 - **Number of Transactions per Channel**



Noting the *graph 11 - Number of Transactions per Channel,* the clients tend to use more the Online Channel for do their transactions, with 1946 transactions and the Channel that they used less, was the Store, with only 316 transactions.

Concluding: the clients prefer do their transactions Online, because the majority wants to avoid confusion from physical shops.

Final notes and conclusions

Once we finished the data treatment through various methods and its brief analysis through data visualization methods, we were able to draw some conclusions, even aknowleging that that was not the purpose of our team.

The first conclusion we were able to make was that the part of The Market's clients that has at least one kid (representing 68,14% of the total of clients) is the group that represents almost 80% of the Total Sales of the company.

We would suggest the creation of directed marketing campaigns to atract new customers without Kids, who are the minority of The Market.

We would also suggest the creation of incentives to Diamond clients, once they only represent less tha 5% of the total client list but they hold more than 17% of Total Sales. This means that these group of individuals is a valuable group for the business to focus on, with (perhaps) premium client support or iniciatives that would make these customers to feel special.

It might be interesting to keep in mind that clients from Lisbon are more than half of *The Market*'s customers, so this is a very valuable location for the business.

Considering that tha category Home and Lifestyle represents more than 20% of *The Market*'s sales, it would be a good idea to focus on this category to improve its value and sales volume.

To finish, the rating average of satisfaction given by the customers (6,87) is also something that would be interesting focusing on getting better results. The Market should evaluate possible details that would improve the quality of the clients perception of the business.