# Data Pre-Processing

## “The Market” Customer Signature Table


**Professor**

Joana Neves

**Elements**

Rafael Sequeira - R20181128

# INDEX

# Introduction

***The Market*** is a retail company based in Portugal that focuses on selling a variety of products that vary from Electronics Accessories to Sports and Travel. The goal of this project is to build a customer signature table from a transactional table in order to help the "Market" to have a deeper knowledge about their customers, therefore, enabling them to gain competitive advantage by increasing customer satisfaction.

The only program used to develop this project was Python using Jupyter Notebook as IDE.

# Data Treatment

Before starting to build the signature table for the company it was important to make sure that the data was in good shape in order to avoid putting wrong or biased information into the signature. To do that some fundamental data preprocessing steps were made, including checking for outliers, filling missing values, and checking for incoherence (inconsistencies).

# Descriptive Statistics

By looking at the descriptive statistics table it's possible to see that in general the data looks goods, however we have some problems in a few variables (cogs and rating) since they present incoherent values. "Cogs" has a minimum value of -99 which is impossible since cogs has to be greater than 0 and "Ratings" have values greater than 10 which is also impossible since the maximum rating is 10.

| | count | mean | std | min | 25% | 50% | 75% | max |
|---|---|---|---|---|---|---|---|---|
| **cust_id** | 5000.0 | 271001.232400 | 2409.033580 | 266783.0000 | 268895.0000 | 271045.0000 | 273094.0000 | 275252.00 |
| **Kidhome** | 4984.0 | 0.721308 | 0.448401 | 0.0000 | 0.0000 | 1.0000 | 1.0000 | 1.00 |
| **Unit price** | 5000.0 | 69.963190 | 30.334670 | 10.1700 | 43.1900 | 83.2950 | 96.1225 | 99.96 |
| **Quantity** | 5000.0 | 5.892600 | 3.004475 | 1.0000 | 3.0000 | 6.0000 | 9.0000 | 10.00 |
| **Tax** | 5000.0 | 20.984337 | 14.880718 | 0.5085 | 7.4065 | 16.0720 | 34.8700 | 49.65 |
| **Total_amt** | 4981.0 | 441.158714 | 312.445226 | 10.6785 | 156.0300 | 338.2155 | 733.6035 | 1042.65 |
| **cogs** | 5000.0 | 418.101940 | 298.380537 | -99.0000 | 145.5000 | 320.2150 | 696.8500 | 993.00 |
| **Rating** | 4968.0 | 15.394082 | 284.165457 | 4.0000 | 5.5000 | 6.8000 | 8.3000 | 10000.00 |

| | count | unique | top | freq |
|---|---|---|---|---|
| **DOB** | 5000 | 758 | 1997-03-31 00:00:00 | 26 |
| **Nationality** | 174 | 4 | PT | 153 |
| **Gender** | 5000 | 2 | M | 2761 |
| **Address** | 5000 | 5 | Lisbon | 3028 |
| **Channel** | 5000 | 3 | Online | 3339 |
| **Type_payment** | 5000 | 4 | MBWay | 3808 |

*Figure 1 - Descriptive Statistics*

# Outliers

The first step of data treatment performed in this project was the outlier's check. This is a crucial step to perform before building a signature because the majority of the transformations that are performed to build a signature are aggregations, meaning that the outliers could be hidden from us after aggregating the data, therefore, creating distortions and bias that would be difficult control later on.

To check for outliers it was used boxplots, which is a good technique to find univariate outliers.

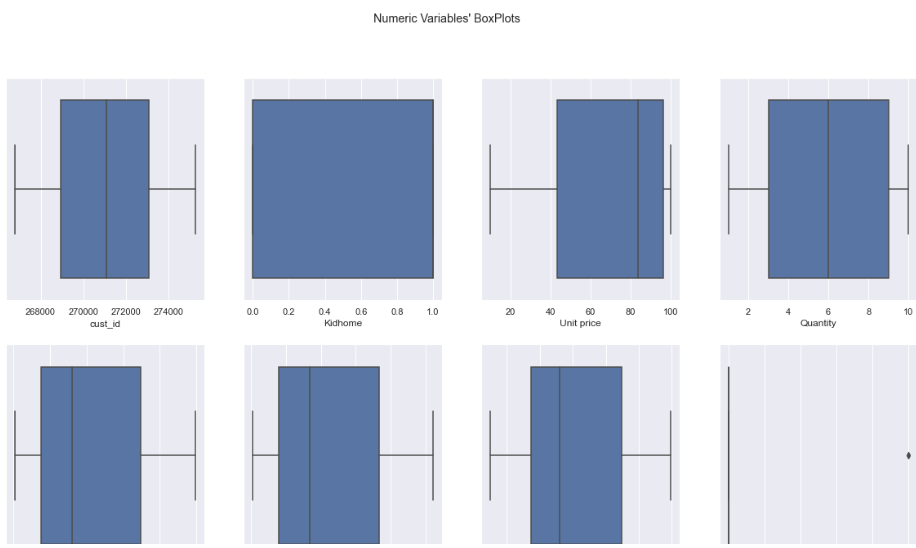



*Figure 2 - Boxplots before outliers removal*

By looking at the boxplots above it possible to see that the only problem that we have regarding outliers is in Ratings, so every observation with a rating greater than 10 was removed (on total **7 observations** were removed during this step).

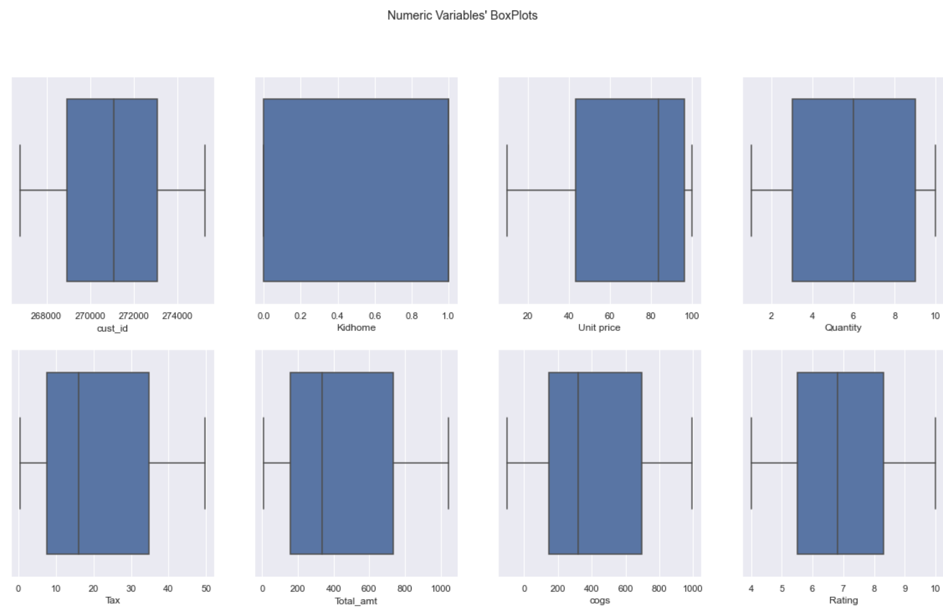



*Figure 3 - Boxplots after outlier removal*

# Missing Values

After removing the outliers from the transaction tables, it was important to check for missing values since most algorithms don't work when there is something missing in the dataset. By looking at the missing values report we can see that there are both numerical and categorical variables with missing values. In terms of numerical variables, we have Total_amt and Ratings with missing values, and regarding categorical and dummy variables we have Nationality (categorical) with a major amount of missing data and Kidhome (dummy/ binary)

```
cust_id           0
tran_date         0
DOB               0
Kidhome          16
Nationality    4819
Gender            0
Address           0
Channel           0
Type_payment      0
Product line      0
Unit price        0
Quantity          0
Tax               0
Total_amt        19
cogs              0
Rating           32
dtype: int64
```

```
cust_id         0
Kidhome         0
Unit price      0
Quantity        0
Tax             0
Total_amt       0
cogs            0
Rating          0
tran_date       0
DOB             0
Nationality     0
Gender          0
Address         0
Channel         0
Type_payment    0
Product line    0
```

*Figure 4 - Missing values before and after imputation*

For the categorical variables (Nationality), since there are a lot of missing values (4819), the missing values were filled by using the mode of the category, in this case, ‘PT’. A similar approach was performed for the Kidhome variable, however, in this, there were a lot fewer missing values which makes the imputation more robust.
Regarding the numerical variables with missing data (*Total_amt and Ratings*) in the case of Total_amt the imputation of the missing values was performed using the formula Tot_amt = cogs + tax, so in this case, we don’t need to worry about the robustness of the imputation since Tot_amt is a linear combination of 2 other variables. However regarding Ratings, since there is no information to help perform the imputation it was used KNN impute that performs the imputation of the missing values based on the nearest neighbors of the observation in question (in this case K=5 meaning that the algorithm will take into consideration the 5 nearest neighbors to perform the imputation, the algorithm will also give more weights to the nearest neighbors when compared to more distant ones).

# Incoherence Check

After imputing the missing values, it was performed an incoherence check in order to detect and remove possible observations that didn’t make sense in terms of consistency.

The following inconsistencies were considered:

**1st –** Observations with Date of Birth greater than the date of the transaction

**2nd –** Observation where Cogs is smaller than 0 (it’s impossible because cogs reflect the total price paid by the customer excluding tax)

On total, 28 observations were removed by the coherence checking.

# Building the Signature

Before starting to build the signature it’s important to have an overview of what might be the company needs in terms of its customers since every company is different and might need different types of variables in a signature.

```
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   cust_id       4965 non-null   float64
 1   Kidhome       4965 non-null   float64
 2   Unit price    4965 non-null   float64
 3   Quantity      4965 non-null   float64
 4   Tax           4965 non-null   float64
 5   Total_amt     4965 non-null   float64
 6   cogs          4965 non-null   float64
 7   Rating        4965 non-null   float64
 8   tran_date     4965 non-null   datetime64[ns]
 9   DOB           4965 non-null   object
 10  Nationality   4965 non-null   object
 11  Gender        4965 non-null   object
 12  Address       4965 non-null   object
 13  Channel       4965 non-null   object
 14  Type_payment  4965 non-null   object
 15  Product line  4965 non-null   object
dtypes: datetime64[ns](1), float64(8), object(7)
```

*Figure 5 - Original features of the transactional table and respective data types*

By looking at the variables present at the transactional table (see metadata) it's possible to see that the company has some interesting features that might be useful to create more knowledge about the situation of the company. In a very succinct way, it's reasonable to say that the variable DOB can be used to calculate the customer age, the variables Channel, Type_payment, and Product_Line can be used to create new variables that assess how the customers interact with the company (eg: which customers buy more using MBWay, etc), which can be done by aggregating (groupby) each category of each variable with the customer ID and then summing the Total_amt (this type of information can be very useful for customer segmentation). Besides that, we can also calculate how frequently a customer visits the store and how much they spend on average on each transaction.

After aggregating the transactional table, the customer signature ended up with 800 rows, meaning that there are 800 customers in the signature table (in a signature, there is one row per element in study, in this case, customers). The initial transactional table had 801 customers indicating that 1 customer was eliminated during the Pre-Processing stage. In terms of features, the signature table ended up with 27 variables. The features that start with Fav_ indicate the categories where the client spends more money (eg: if Fav_Payment_Method = MBway, then it means that the client mainly pays using MBway).

| Column Names | Description |
|---|---|
| Cust_id | Customer ID |
| Nationality | Nationality |
| YOB | Year of Birth |
| DOB | Date of Birth |
| Gender | Gender |
| Address | City of residence of the customer |
| Kidhome | Dummy takes value 1 if customer has children, 0 otherwise |
| Age | Age of the customer |
| frequency | Number of times the customer made a transaction in the store |
| Rating | Average Rating given by the customer |
| Average_expense | Average expense of the customer |
| Catalog | Amount spend via Catalog |
| Online | Amount spend via Online |
| Store | Amount spend via Store |
| Fav_channel | Channel where customer spend more money, takes value 'Mixed' if there is a draw |
| Electronic accessories | Amount spend in Electronic Accessories |
| Fashion accessories | Amount spend in Fashion accessories |
| Food and beverages | Amount spend in Food and beverages |
| Health and beauty | Amount spend in Health and beauty |
| Home and lifestyle | Amount spend in Home and lifestyle |
| Sports and travel | Amount spend in Sports and travel |
| Fav_prod_line | Product Line where customer spend more money, takes value 'Mixed' if there is a draw |
| Cash | Amount of money paid by the customer in Cash |
| Credit Card | Amount of money paid by the customer using a Credit Card |
| MBWay | Amount of value paid by the customer using MBWay |
| Paypal | Amount of value paid by the customer using PayPal |
| Fav_Payment_Method | Payment method that the customer used more (in monetary terms) |
| Tot_Amount | Total Amount of money spend by the customer in the store (including tax) |

*Figure 6 - MetaData*

# Validation check

Before performing data visualization to get to know better the data it can be important to perform a validation test to check if the results of the signature match the ones from the transaction table. (The sum of the total amount should be the same in both the signature and the transactional table).

```
Testing Channel section:
Transaction Table:
Channel
Catalog     2.552009e+05
Online      1.441392e+06
Store       4.758447e+05
Name: Total_amt, dtype: float64

Customer signature Table:
Catalog     2.552009e+05
Online      1.441392e+06
Store       4.758447e+05
```

```
Testing Type_payment section:
Transaction Table:
Type_payment
Cash           2.471899e+03
Credit Card    4.813863e+05
MBWay          1.665517e+06
Paypal         2.306278e+04
Name: Total_amt, dtype: float64

Customer signature Table:
Cash           2.471899e+03
Credit Card    4.813863e+05
MBWay          1.665517e+06
Paypal         2.306278e+04
```

```
Transaction Table:
Product line
Electronic accessories    383052.8415
Fashion accessories       363476.4000
Food and beverages        350969.4405
Health and beauty         241458.1260
Home and lifestyle        417069.3030
Sports and travel         416411.7195
Name: Total_amt, dtype: float64

Customer signature Table:
Electronic accessories    383052.8415
Fashion accessories       363476.4000
Food and beverages        350969.4405
Health and beauty         241458.1260
Home and lifestyle        417069.3030
Sports and travel         416411.7195
```

*Figure 7 - Validation check between transactional table and signature*

**Note:** the values on the report aren't exactly the same as in the notebook because there was a change on the outlier treatment (the previous outlier removal also removed the observations where Rating was null). The updated values are the ones on the notebook

# Data Visualization and discussion

## Visualizations from the transaction table:

By looking at the graphic below, it's possible to observe that by far December is the most lucrative month for the company, this can be considered normal since it's the month where people tend to buy more things and spend more money.

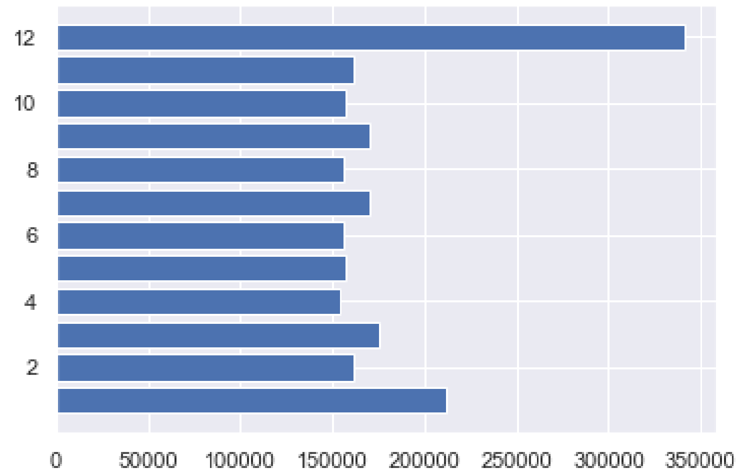



*Figure 8 - Total Amount Grouped By month (not taking into consideration the Year)*

Total amount grouped by both months and years. For the years 2018 and 2020 the results aren't that interesting due to lack of data from most months, in 2018 we only have data for November and December, and for 2020 we only have data for January (not shown in the report because is redundant).

The results from 2019 (figure 10), the only year with complete data, are similar to the results shown above with December being the most lucrative month for the company followed by March (3), July (7), and September (9). The differences between December and the other months are much smaller in the 2019 graphic when compared to the graphic shown above because in the

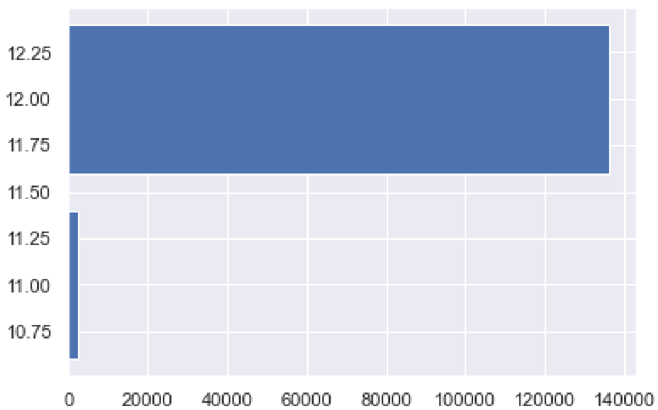



*Figure 9 - Total Amount Grouped by Year (2018)*

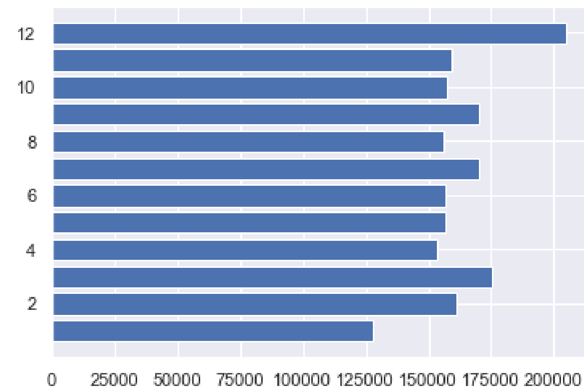



*Figure 10 - Total Amount Grouped by Year (2019)*

graphic above (figure 7), we are also taking into consideration the sales of 2018 that have high value of sales in December.

The graphics below (figures 11-13) calculate the total amount of money received by the company taking into consideration the different sectors that the company presents (Product Line, Type of payment, and Channel).

Regarding the Product Line, the most popular categories among customers (categories where customers spend the most money) are the 'Sports and Travel' and 'Home and Lifestyle' and the least popular category, by far, is Health and Beauty. When looking at payment methods, it's easy to see that the most lucrative payment method for the company is by far MBWay, and regarding channels, it's possible to see that most of the customers buy online making it the most lucrative channel that the company has.

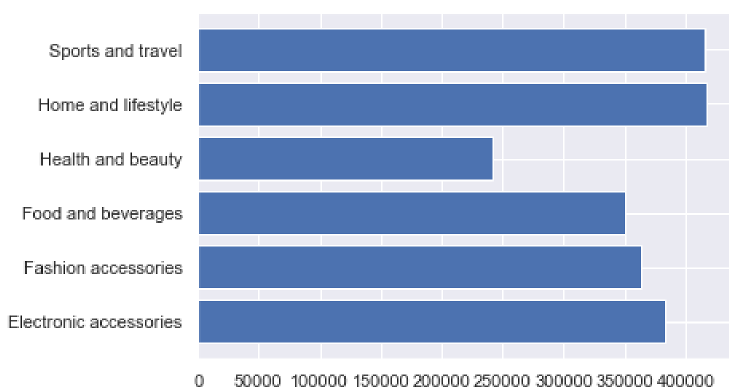



*Figure 11- Total Amount Group By Product Line*

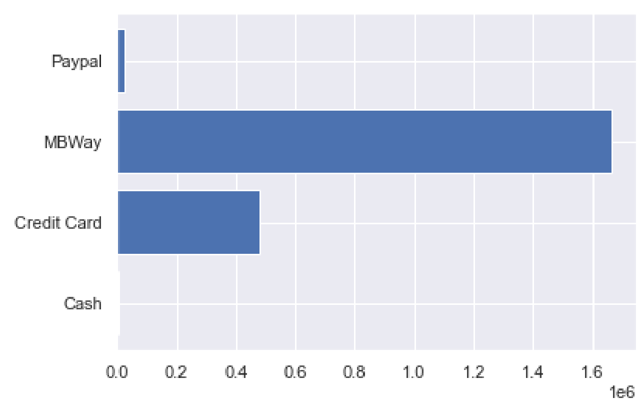



*Figure 12 - Total Amount Group by Payment method*

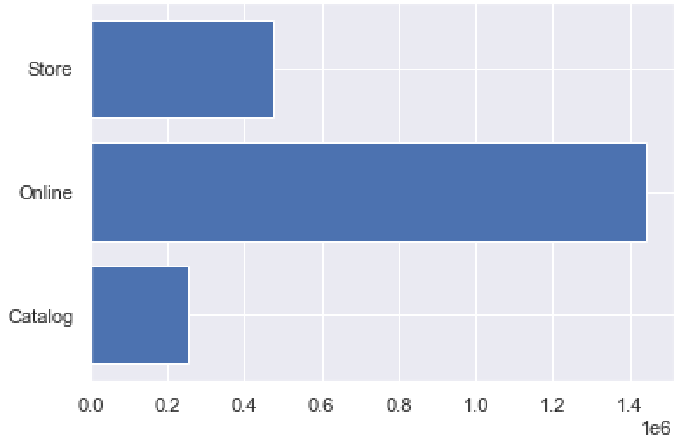



*Figure 13 - Total Amount Grouped by channel*

<u>Visualization from the signature table:</u>

The graphics below count the number of customers that fall into different categories with the goal of getting to know better the customers and how they interact with the store.

The 1st graphic represents the distribution of the clients regarding their gender, and it's possible to see that the store has more male customers than female ones, although this difference isn't very substantial. (1)

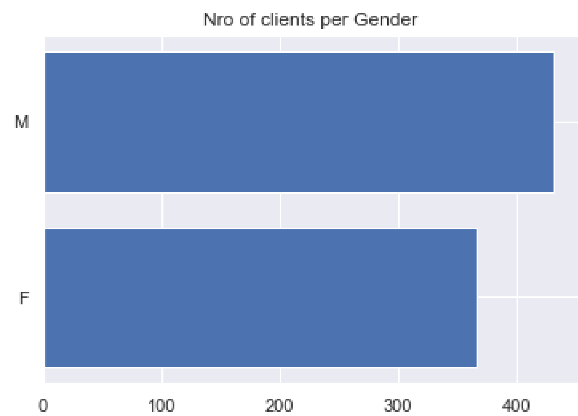
The 2nd graphic represents the distribution of the company regarding Nationality, by looking at the graphic it's possible to see that the majority of the customers are from Portugal. (note: it's important to take into consideration that the variable Nationality had a lot of missing values in the Pre-Processing stage which might bias the results towards having more Portuguese people when compared to other countries). (2)

The 3rd graphic represents how the companies' customers are distributed around the country, being possible to observe that the majority pf the clients lives in Lisbon and all the other cities have similar values

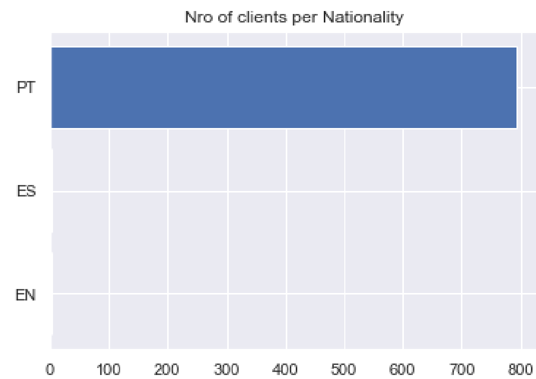
The last three graphics represent the distribution of the customers regarding their favorite sector of the company (eg: The 1st graphic counts the number of clients by their favorite Product Line).

By looking at graphic 4.1, it's possible to see that the Home and Lifestyle category is the most famous among customers, however, this difference isn't very huge to the other categories. By looking at the following two graphics (graphic 4.2 and graphic 4.3) it's possible to detect huge differences in preferences since customers tend to choose MBWay to pay to conclude their transactions and most of the customers prefer to use the online store to buy products from the company.

Graphic nr 1




Graphic nr 2

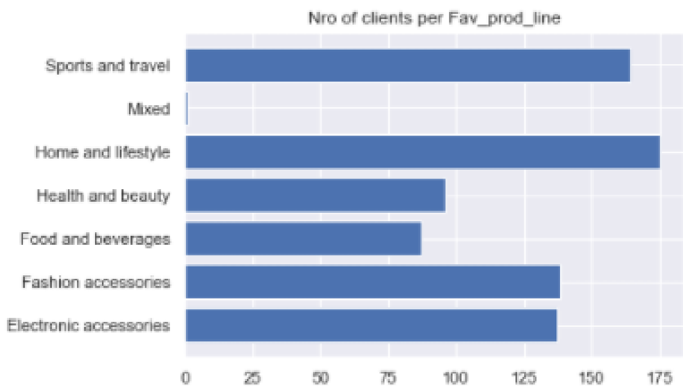



Graphic nr 3

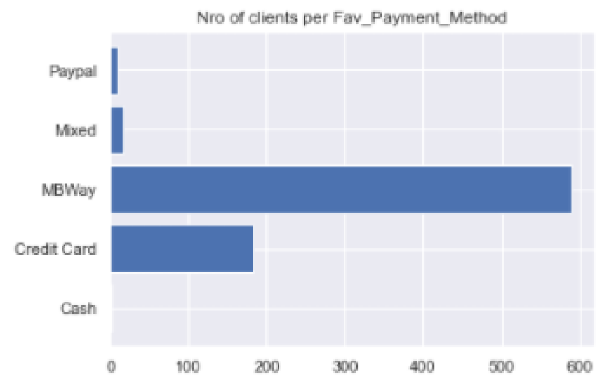

*Figure 14 - Graphics regarding the signature table (pt1)*

**Note:** in the graphics below, where is 3.1, 3.2, 3.3 it should be 4.1, 4.2, 4.3

Graphic nr 3.1




Graphic nr 3.2




Graphic nr 3.3

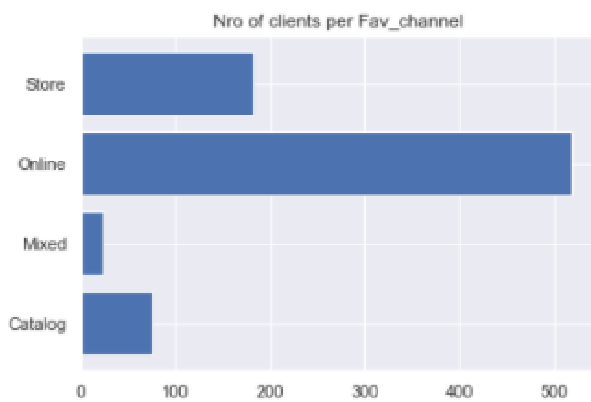



*Figure 15 - Graphics regarding the signature table (pt2)*