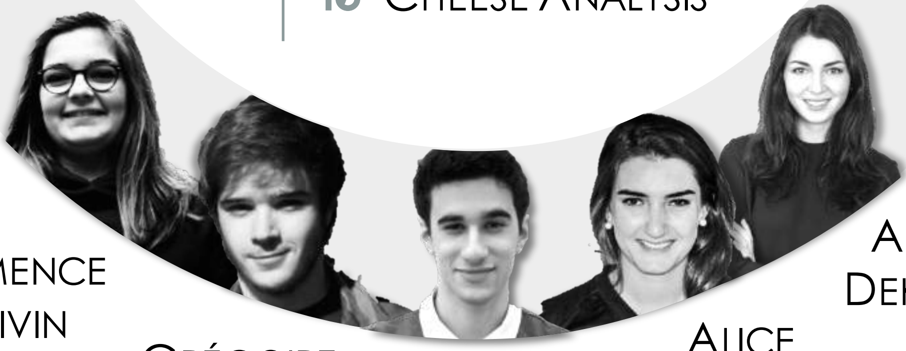
dataiku

CASE STUDY

Group 1

Toulouse Business School
Think & Create

TABLE OF CONTENTS



CLÉMENCE BOIVIN

GRÉGOIRE GALLIER

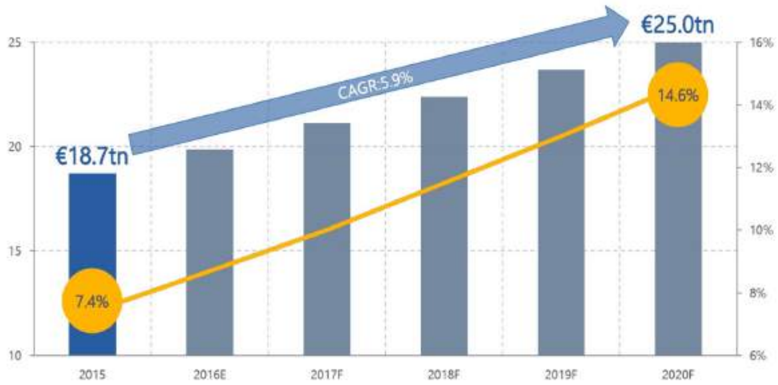
SALIM AMARTI

ALICE DE LATAULADE

ANDRÉA DEKNEUVEL

Global Analysis

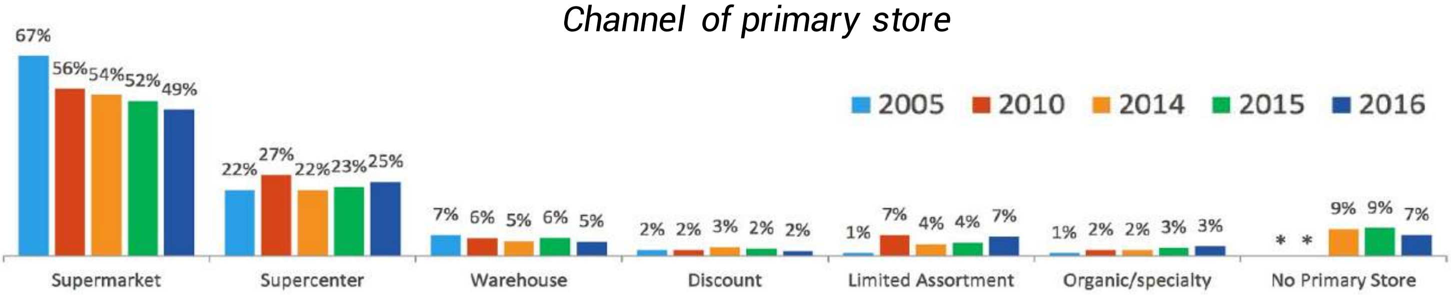
Global retail sales and e-commerce share



The beginning of the 1900's is important in the history of retail since it marks the transition from mom-and-pop stores (atelies familiaux/artisanaux) to department stores. M. Selfridges redefined ***retail: from a necessary errand to a form of entertainment*** (19th century). The retail has known several evolution, especially since the rise of new technologies in 1990s. We have moved **from brick-and-mortar shopping experience to a click-and-mortar approach**: mobility has driven changes in shopping behaviour with major retailers seeking to adapt their business approaches (important investments in e-commerce). The e-commerce represents 18,7 trillion of euros in 2015. The global retail market is set to grow strongly.

Focus on the United States

Channel of primary store



The data we used

•We began with the **transaction data that we joined with the data set of product**

•We have created **additional columns** for the grocery and product variables and columns with the quantity and the value of the product from this variables,

•We have specified also two columns with loyalty and non loyalty card.

•We **excluded from this analysis the basket with a quantity equal to 0**

•With the variable quantity we have observed significant extreme values due to the product gasoline. We decided to create other variables in which the **gas product** would be recoded with a quantity equal to one like one passage to cash.

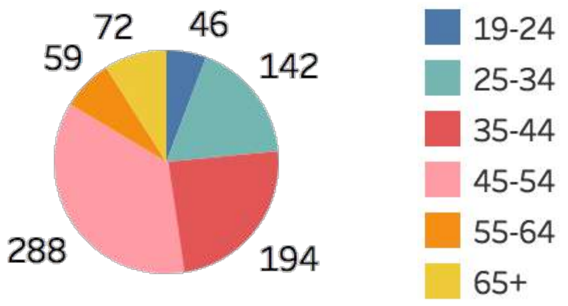
•We created the variables **discounted sales** to have the number of product with a reduction

In general we **haven't exclude many variables** but we realize several precision with the creation of new variables in order to have more information to realize different groups.

In order to identify the level of customer loyalty we use **the approach RFM.** For the recency we use the variables **day min and max** for the frequency **day distinct** and for monetary value we choose the variables to have the **maximum number of product bought** by household, **the money spent during all visits** and the average of money spent during a period.

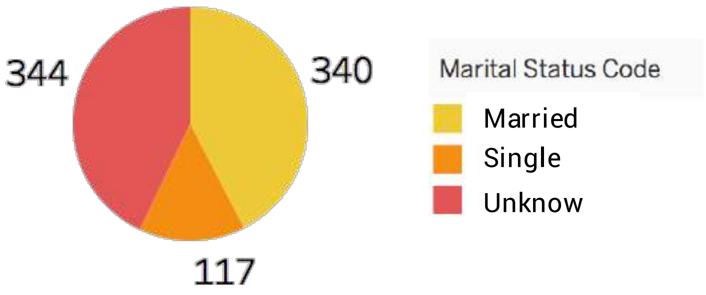
HOUSEHOLD ANALYSIS

AGE DESCRIPTION



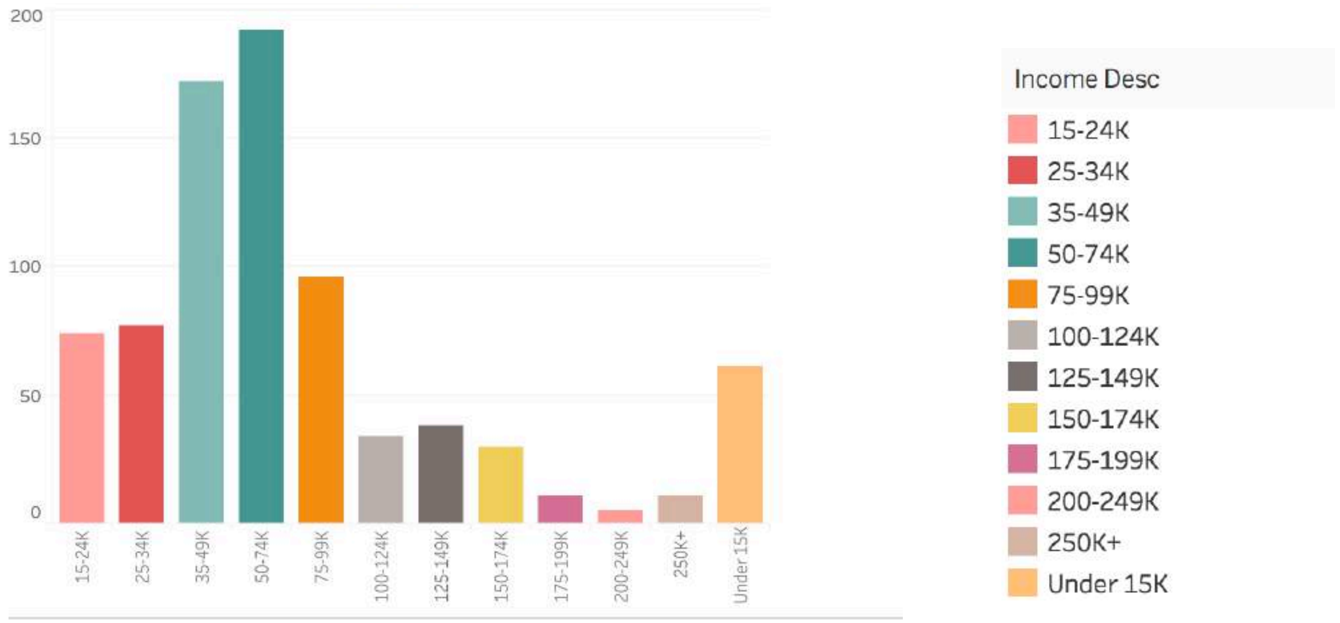
Mean : 44 years old

MARITAL STATUS



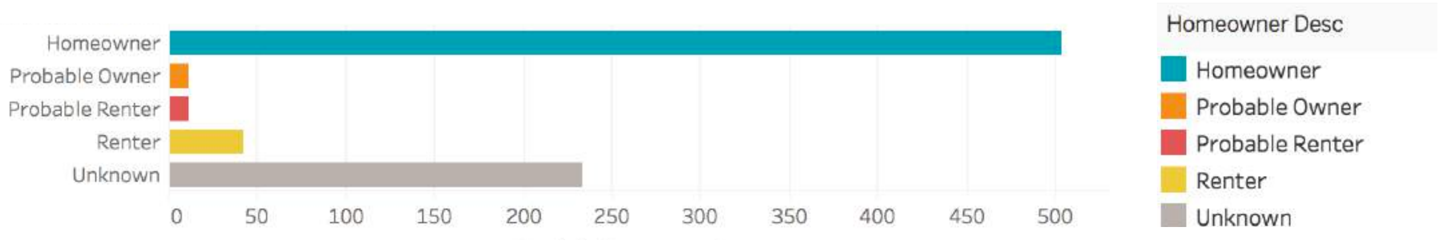
Most of people are married but there is a lot of unknow data so we won't use this variable anymore.

INCOME DESCRIPTION



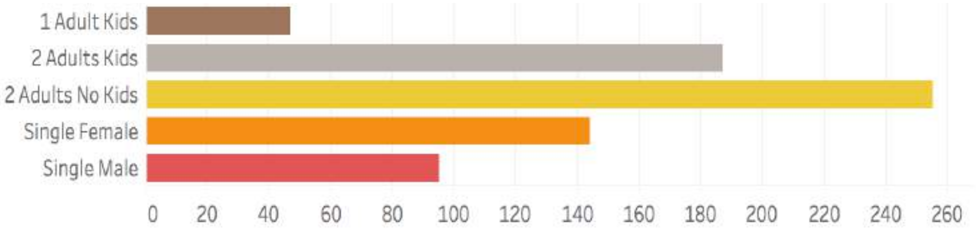
Mean : 64 500 dollars

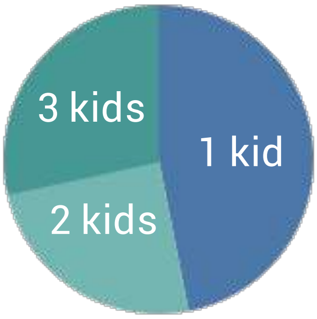
HOMEOWNER DESCRIPTION



A majority of homeowner

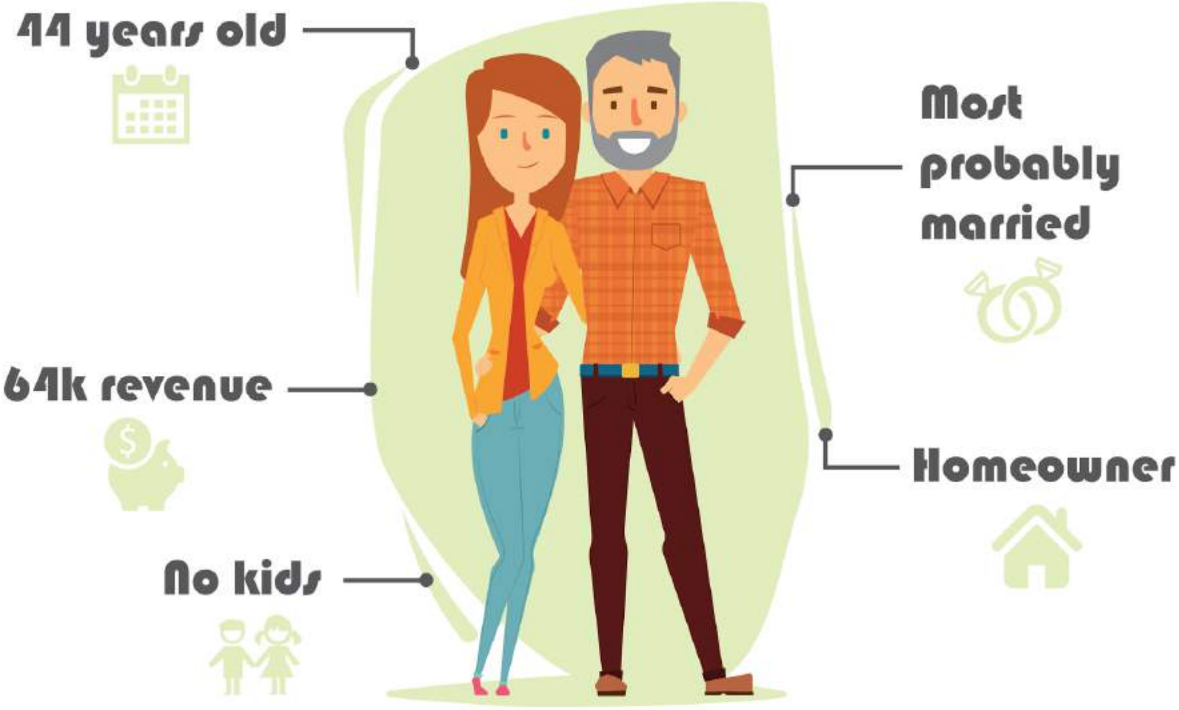
HOUSEHOLD COMPOSITION





Most of households don't have any kids and when they have, it is mostly one kid.

Typical Household



BASKET ANALYSIS

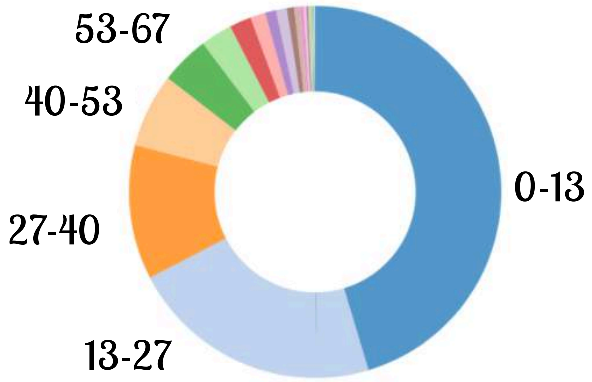
METHODOLOGY

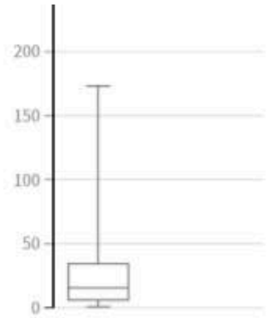
- ❖ Delete baskets with quantity 0
- ❖ Construction of new columns for the GROUP in the transaction data :
- ❖ Creation of a Sub_commodity_prime column :
  - 1 = product without gazoline
  - 0 = gazoline
- ❖ Creation of a Sub_commodity_quantity prime column :
  - If Sub_commodity_prime = 1 then put the "QUANTITY" variable
  - If Sub_commodity_prime = 0 then put 1. Gazoline will value 1 as 1 checkout
- ❖ Creation of a Sub_commodity_value_sales_prime column :
  - If Sub_commodity_prime = 1 then put "SALES VALUE"
  - If Sub_commodity_prime = 0 then put also "SALES VALUE". We now have always the exact value of a basket with gazoline

First group : group per basket ID in order to have the quantity, the value…

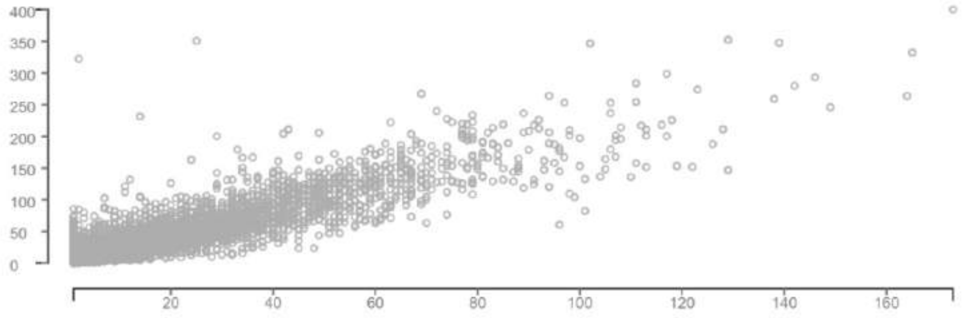
2nd group: group per commodity in order to have the product with the most important quantity sales value per basket

SALES VALUE PER BASKET

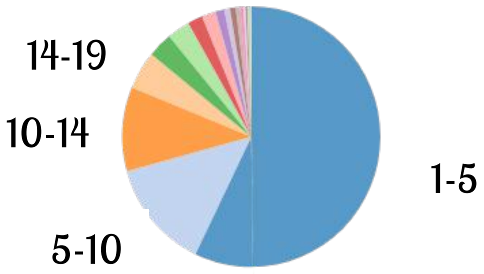


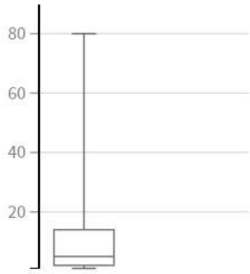


Median = 15
Mean = 27,8
Standard deviation = 35,6



NUMBER OF PRODUCTS PER BASKETS

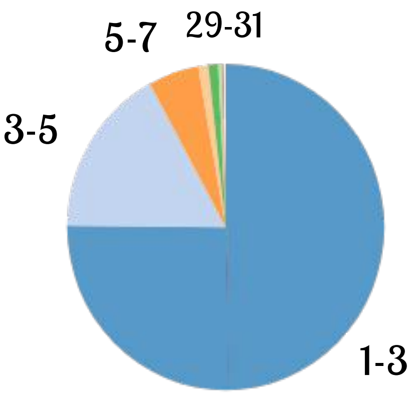




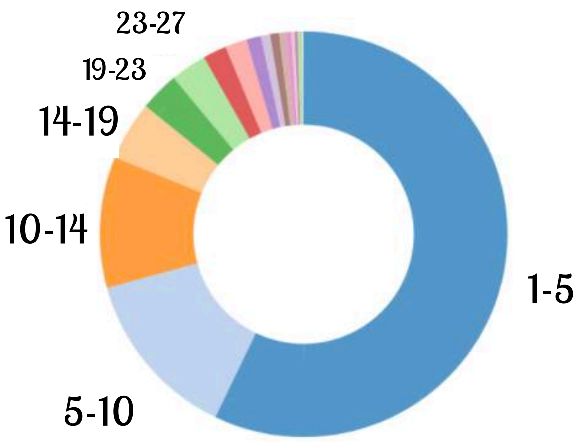
Median = 5
Mean = 11
Standard deviation = 16

BASKET ANALYSIS

NUMBER OF MAXIMUM PRODUCT PER BASKET :



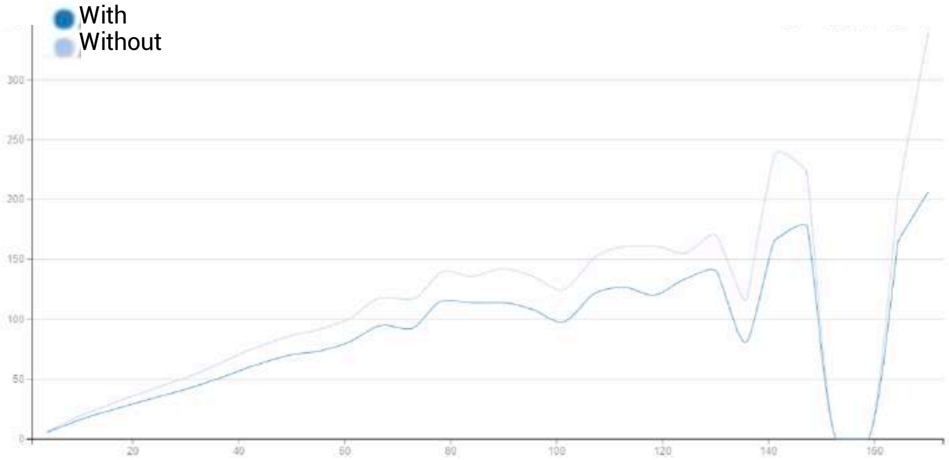
NUMBER OF DISTINCT PRODUCT PER BASKET :



DIFFERENCE OF PRICE :

	Non loyalty card price	Loyalty card price
Min	0	-9
Max	338	249
Mean	21	17
Median	10,6	8,9
Standard deviation	30	25,035

QUANTITY OF PRODUCTS PER BASKET WITH AND WITHOUT LOYALTY CARD :



People buy more quantity with loyalty card

Typical Basket

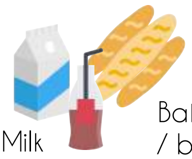
Basket more important with a loyalty card

Mostly bought in grocery stores

1 – 10 items

1 - 27€

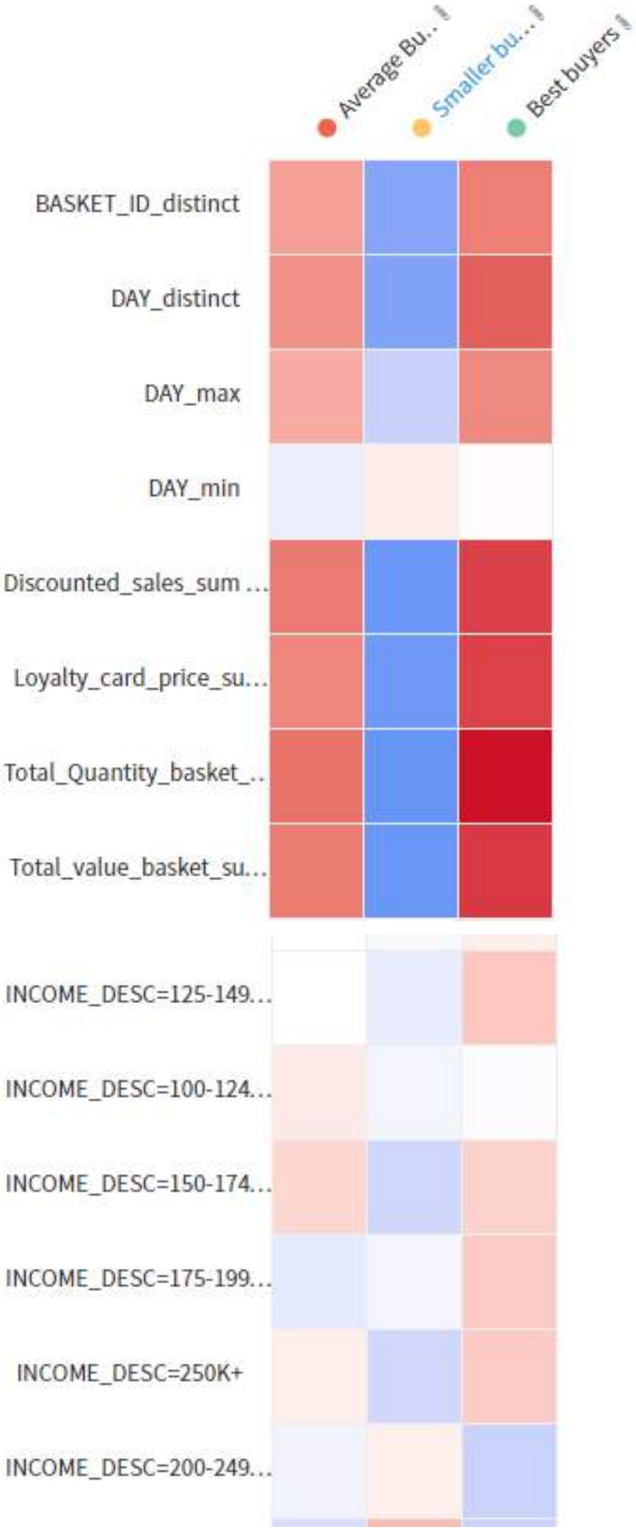
Top 3 products



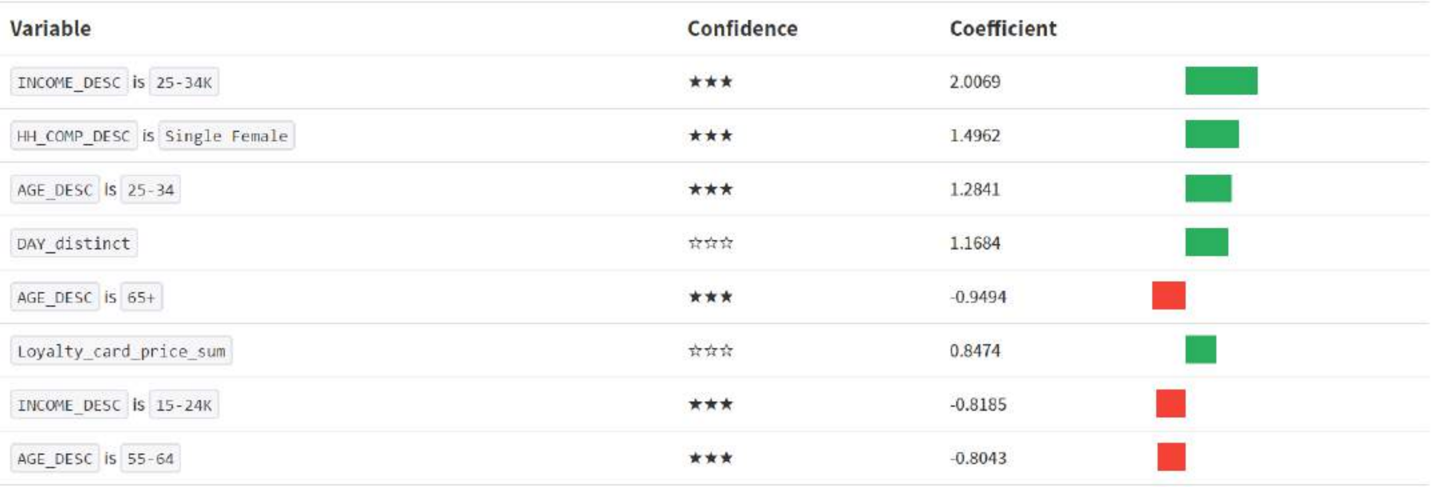
Cluster Analysis

In order to have a better grasp on the typical household and basket that we had in this dataset, we used a cluster, segmenting the population in 3 groups. The first group was composed of customers who bought in average 40% less in volume and in value, buying 40% less discounted products compared to an average customer. The second group was composed of customers buying 28% more in volume and price, with 28% more discounted products than an average client. Meanwhile, the third cluster was composed of the best clients, buying in average 135% more in quantity and 143% more in value, while buying 144% more discounted products.

Thanks to the analysis of the Heatmap and the Cluster Profiles, we could determine our most loyal customers, who also are the biggest buyers. They have visited the store on average 278 times, almost 95% more than the average person. Meanwhile, the cluster of less loyal customers only visited the store 100 times, and average loyal customers visited the store 171 times out of 700 days. Most loyal customers have a total basket value and quantity almost 140% higher than the average. Their age is evenly split between 45-54 and 35-44, each accounting for 38% and 39% of customers in that cluster. We can also get some insights on their marital status, as 35% of people in this cluster are 2 adults with kids, and are mostly part of a family of 3 people. It is all the more important to care about these customers since 71% of them are homeowners and will most likely stay in the area and consume in the same shop if they are treated well.25% of the best buyers cluster have an income between 50-74k, but the most representative part of this cluster is the higher amount of people having a revenue of over 100K (over 30%), which could explain the higher amount of expenses they make in stores.



Ridge regression model of beer :



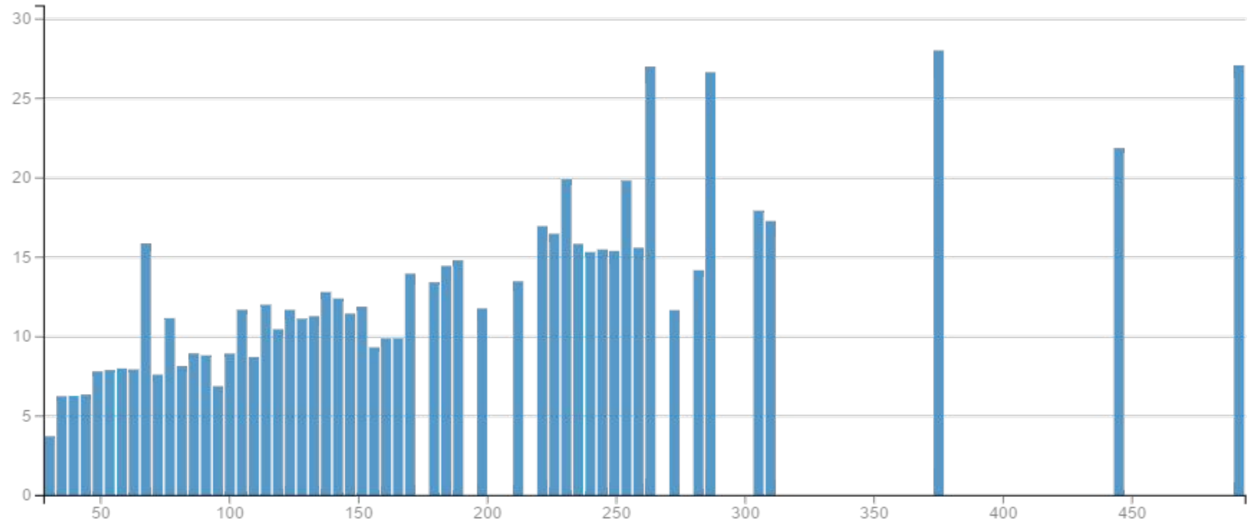
Variable	Confidence	Coefficient	
INCOME_DESC is 25-34K	★★★	2.0069	🟩
HH_COMP_DESC is Single Female	★★★	1.4962	🟩
AGE_DESC is 25-34	★★★	1.2841	🟩
DAY_distinct	☆☆☆	1.1684	🟩
AGE_DESC is 65+	★★★	-0.9494	🟥
Loyalty_card_price_sum	☆☆☆	0.8474	🟩
INCOME_DESC is 15-24K	★★★	-0.8185	🟥
AGE_DESC is 55-64	★★★	-0.8043	🟥

Random Forest R^2= 10%
Ordinary least squares R^2= 16%
Ridge regression R^2 = 10%

The R^2 is very low so **the model don't really represent the reality**. We tried to be more precise with the variables creating for example columns with 'kids or not kids" "married or not" and rejecting some variables but it didn't improved the model.

With this model it is possible to predict the beer consumption:

Avg. of prediction by DAY_distinct



If we take this model as representation of the reality we would have a consumer population caracterised by :

Typical Buyer

25-34
years old

Loyalty card

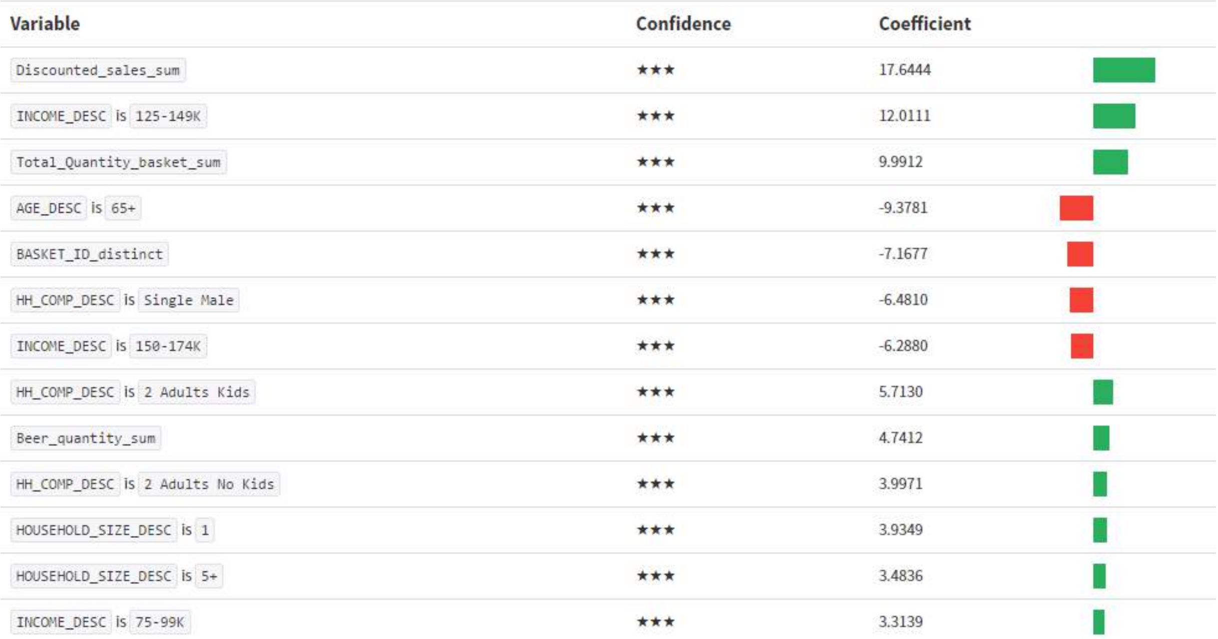
Also cheese consumer

Income : 25-34K

Single Female

CHEESE ANALYSIS

Ridge regression model of cheese :

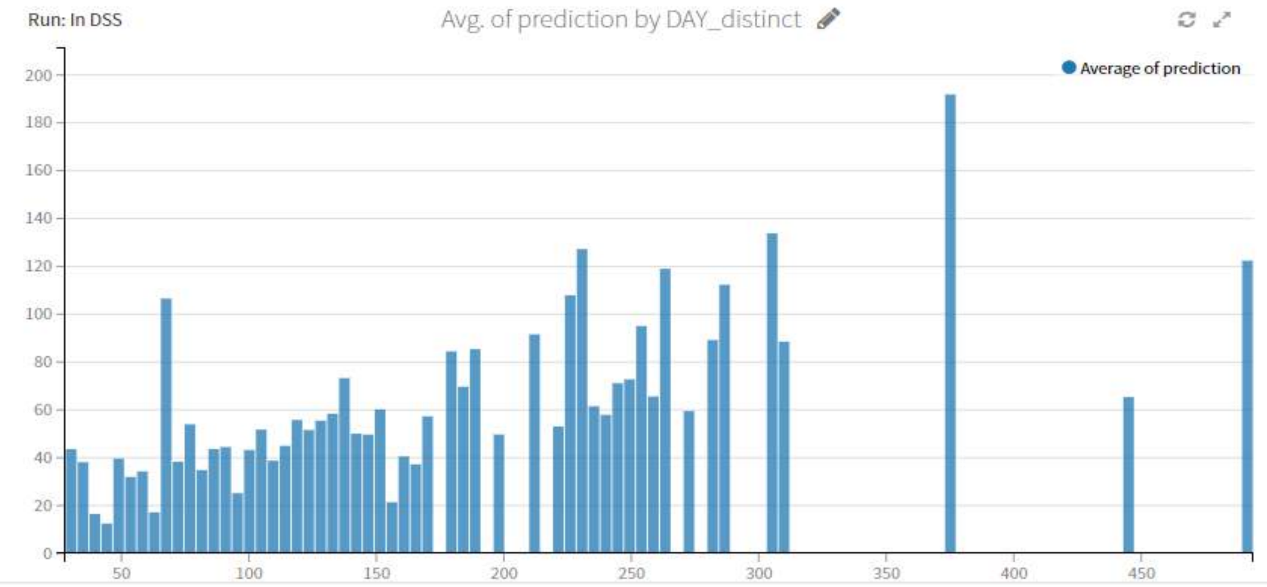


Variable	Confidence	Coefficient	
Discounted_sales_sum	★★★	17.6444	🟩
INCOME_DESC is 125-149K	★★★	12.0111	🟩
Total_Quantity_basket_sum	★★★	9.9912	🟩
AGE_DESC is 65+	★★★	-9.3781	🟥
BASKET_ID_distinct	★★★	-7.1677	🟥
HH_COMP_DESC is Single Male	★★★	-6.4810	🟥
INCOME_DESC is 150-174K	★★★	-6.2880	🟥
HH_COMP_DESC is 2 Adults Kids	★★★	5.7130	🟩
Beer_quantity_sum	★★★	4.7412	🟩
HH_COMP_DESC is 2 Adults No Kids	★★★	3.9971	🟩
HOUSEHOLD_SIZE_DESC is 1	★★★	3.9349	🟩
HOUSEHOLD_SIZE_DESC is 5+	★★★	3.4836	🟩
INCOME_DESC is 75-99K	★★★	3.3139	🟩

Ridge regression R^2 = 51,8 %

Quite good model becuase it explains half of the variations of cheese consumption.

With this model it is possible to predict the cheese consumption:



If we take this model as representation of the reality we would have a consumer population caracterised by :

Typical Buyer

2 adults with kids

Product discount

Income : 125-149K

Also beer consumers

Large basket buyers