Group 20 - Theo Lemonnier and Brieuc Mayer – Quantitative Economics - Data Management project

Introduction

We were given sixteen datasets, including a main one called 'dataset', which rows correspond to individuals. In this dataset, each of the 10000 French individuals is linked to personal information such as its name, age at time of the data collection and sex. Other information are linked to the individuals via codes, which in other datasets correspond to city of residence, occupation (depending on INSEE categories), household type…etc. All those data were provided by the INSEE, which is a pretty reliable source. We have missing information about how and when the data were collected, but they are not necessary for the study we want to make. The datafiles linking the individuals to their information via codes being available, the individuals might be identified, and the dataset can indeed be used, if not in a really dangerous way, at least with a commercial aim. The fact that the information was divided in multiple dataset is not enough to protect extortion of information if we are able to obtain them all.

The last variable of the 'dataset', at the core of our study, is the outcome of a marketing campaign (if it was successful or not), as a Boolean value. We do not have any information about what the marketing campaign object was, how a success is define, and how the INSEE got those results, but we can still derive information, and maybe knowledge, from the analysis of its results.

As a warning before entering in the analysis of the data, we remind that we did not compute statistical tests for the significance of the information we computed, so most of our analysis are partly given with our intuition. Also, we used our intuition to choose which statistics we wanted to compute, since a fully exhaustive exercise would be quasi infinitely long (as we can nearly compute as many cross-variables and conditional correlations statistics as we want. We still believe that our analysis describes the most important information that we could find, and that this analysis really gives robust conclusions to many questions we might ask ourselves. Sometimes, we did not obtain statistics to describe what we observed in the graphs (especially for the grouped analysis part), but we think our analysis of the graphs are robust enough to be kept.

We think our comments on the code explain well what were the technical points that we needed to describe, but we kept the actual economic analysis for the present paper, by pasting on several results that our code directly prints when it runs (which will be pasted in blue in this report).

Data loading and representation

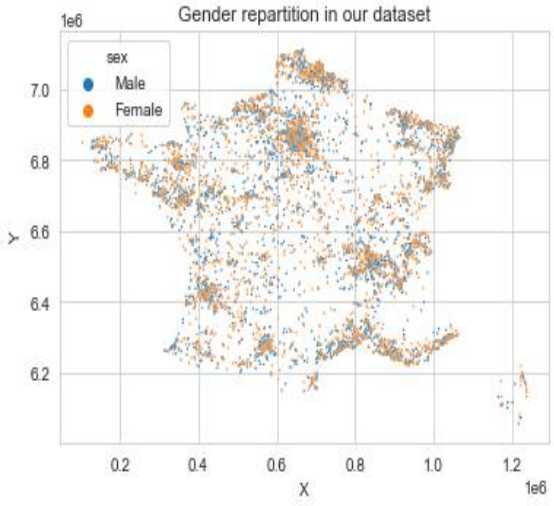
The firsts steps that we realized was to reorganize the data in the most useful way for us : we considered that we do not need to keep separated datasets, so we put all the variables together in the "dataset" dataframe under python. We then saw that each code in a column of the dataset corresponded to a value for the variable in the corresponding dataset, which gave us an overall dataset with only few missing values, and this makes it really easy for us to deal with those extremely scarce missing values. For the majority of our analysis, we did not even needed to deal with it : for most of the variables, we did not considered the missing values as a sign of mistaken information, since we cannot distinguish one from another since we do not have information about how the data was collected. For example, only 774 people had a club, but be kept the 10000 persons, merging every individual in our dataset, including those that were not related to a club.

For the sake of our analysis, we kept in our dataframe several variables that have the same signification (or perfectly correlated signification), such as a column 'Success' which values are Booleans, a column 'success' which values are dummies, and a column 'failure', which values are dummies. Since we do not want to make an overall regression, having repeated columns as such is just a way of helping ourselves, especially since we were two working at the same time in different ways on the same

dataset, and even if it might be a redundant. In our columns, we chose to keep the whole name of the variables (which could be long for the household type or last degree obtained for example), in order to be able to understand perfectly every value quickly at each time. We modified as categorical a list of values that could be, in order to use all the possibilities python offers us, and others where transformed into Booleans. We did not modify fully the dataframe at the beginning, since we had to go deeper in the analysis to know what was the best way to modify the dataframe. The most efficient dataframe is indeed the one we finish with at the end of our analysis (containing cuts of the ages in ranges, cuts of the location in areas depending on the latitude and longitude…etc.). We dropped columns that we did not need anymore, such as the codes, many redundant columns and so on.

Before proceeding to the actual analysis of the paper, we wanted to get an insight of what composed our dataset, so we computed several statistics about it. We paste here what we obtained : all the information given here are rates of what composes our dataset.

Gender

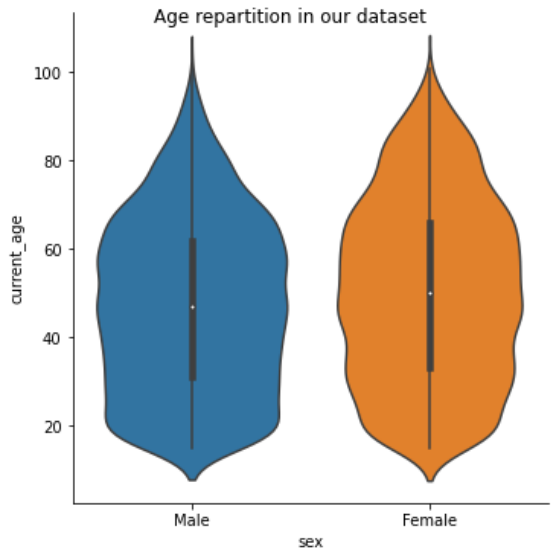


The gender repartition is :
Female 0.5279
Male 0.4721

There are a bit more women the men in our dataset (52,79% women).

We see that they seem to be equally spread on the territory (and it also gives us a good idea of where our individuals come from : the more aggregated the points are the more people come from a region).

Age

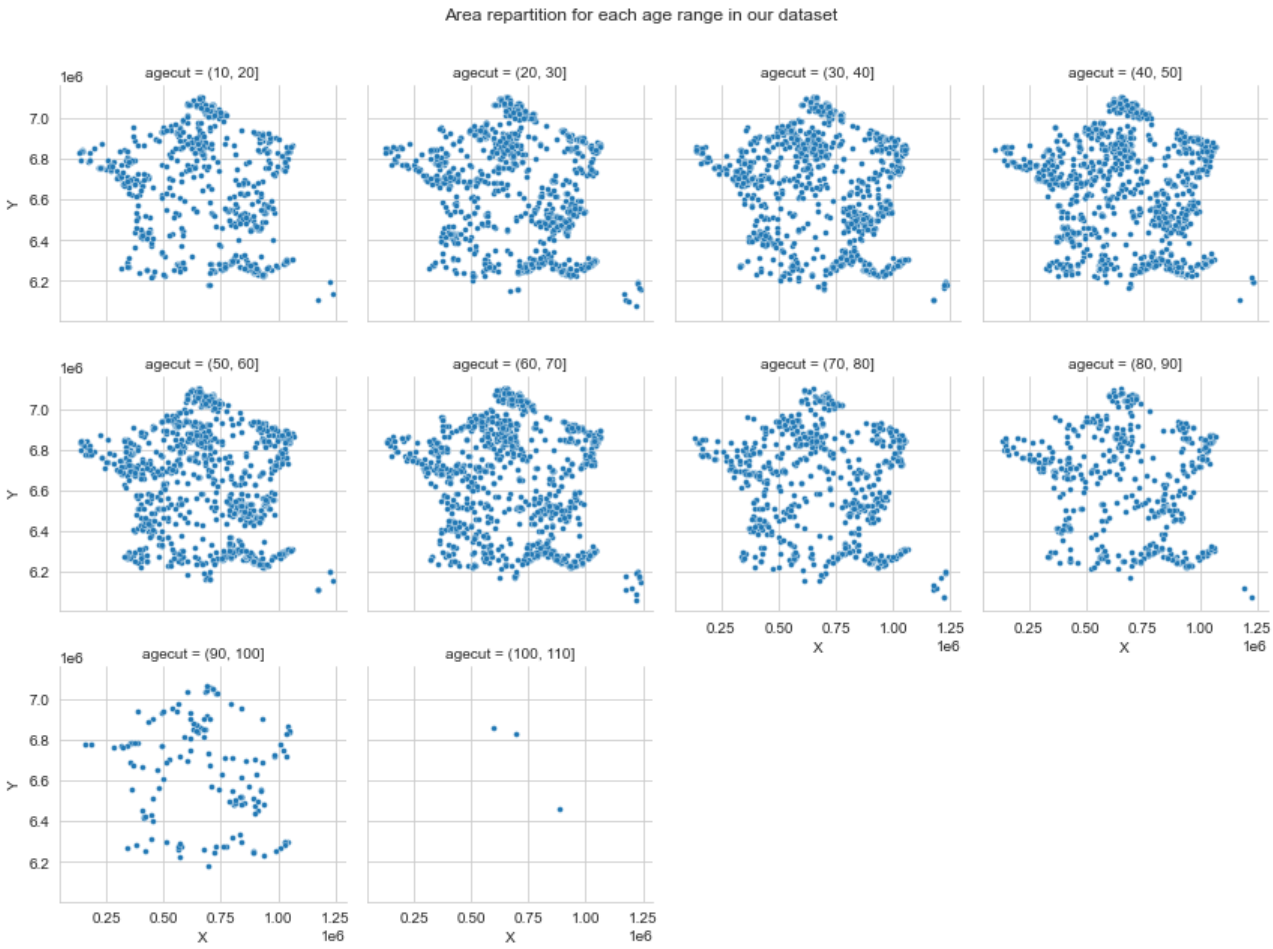


The average age is : 49.1046
The first quartile age is : 32.0
The median age is : 49.0
The third quartile age is : 65.0

There is only one quarter of the population under 32 years old, and half of the population older than 49 years old. This seem older than the real French population.

We see that there is an older population in the women population, but both have the same shape, with a few really young people, a quite uniform repartition of the population between 20 and 60 years old and a decrease after 60 years old, which is stronger for the men then for the women.
It is not representative of the French population, especially for the young ages.

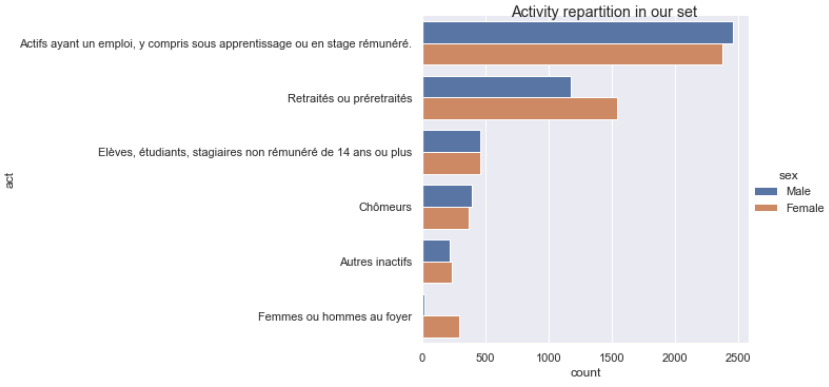
The with the age of the individuals does not seem to be correlated to the regional repartition of the population, as we can see on the following graph.



Occupation / Activity

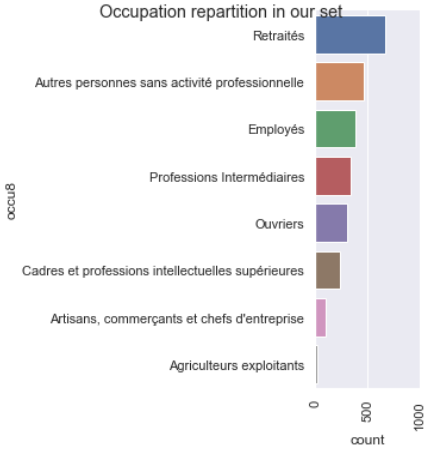
The studying rate is : 0.1157

There is only 11,57% of our population that studies, according to the 'Studying' variable.



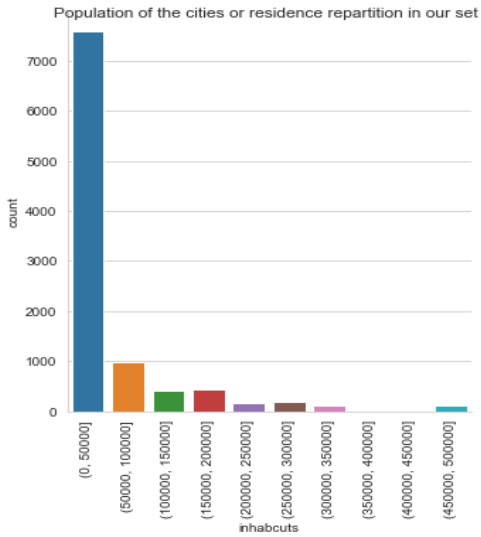
The activity repartition is :

Actifs ayant un emploi, y compris sous apprentissage ou en stage rémunéré.	0.4838
Retraités ou préretraités	0.2718
Elèves, étudiants, stagiaires non rémunéré de 14 ans ou plus	0.0922
Chômeurs	0.0758
Autres inactifs	0.0454
Femmes hommes au foyer	0.0310



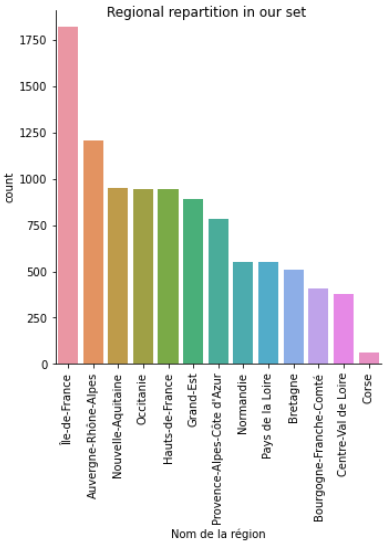
Nearly half of the population is working (similarly for women and men), while students considered as student by their INSEE code account for only 9,22% of our population (so the last 'Studying' information was obtained in another way, maybe with a self-declaration ?). 27% are retired (with a bit more women than men), and others are unemployed (7,58%), inactive (4,54%) or housepersons (mostly women).

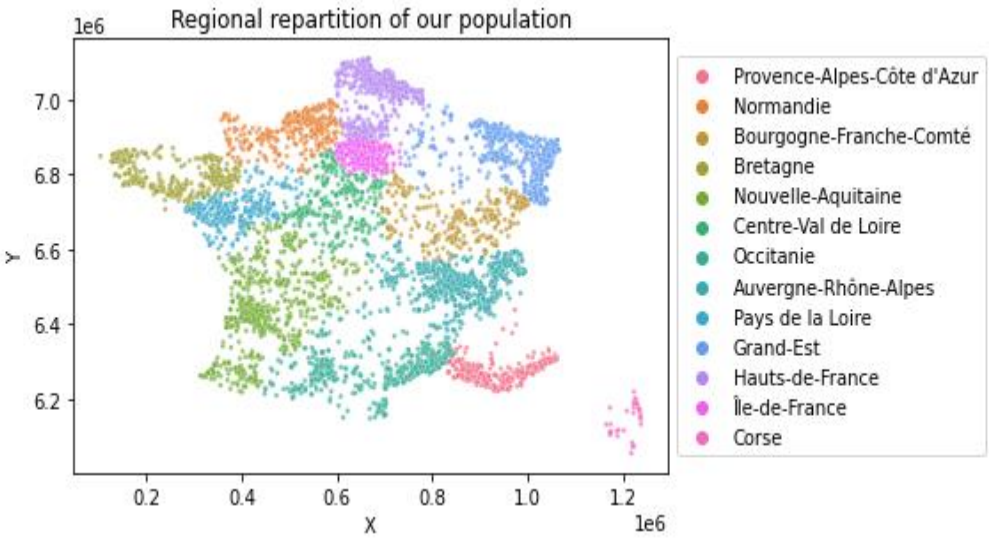
Location of our individuals



The average city population of people of this dataframe is : 46072.9583
The first quartile city population is : 2780.75
The median city population is : 11581.0
The third quartile city population is : 47890.75

Individuals are overall living in small cities (less then 11581 people for half of them). More than 75% of the population lives in cities of less than 50000 inhabitants, which is about 10 times less than the biggest city of the dataframe.



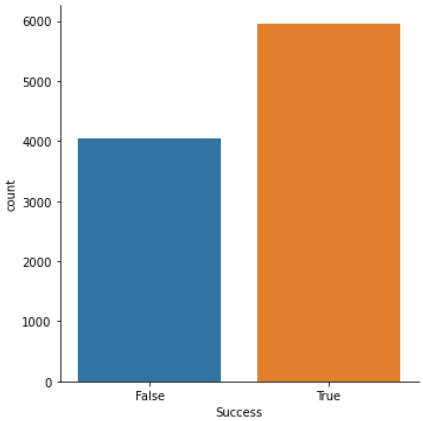


The largest part is in Ile-de-France (about 1800 individuals), and then it is quite uniform for 6 regions, with around 1000 individuals each. 5 regions have about 500 inhabitants interviewed, and Corsica has a lot less.

Clubs

The rate of people registered in clubs is : 0.1537

Outcome of the marketing campaign



The success rate is : 0.596

The main information for now is that the global success rate of the marketing campaign is of 59,6%. We should now compute statistics to have a more precise view over what the data we have can teach us.

Person level analysis

As a first part of our analysis, we will begin with a person level analysis : variable by variable, we should be able to obtain both a descriptive view of the success set when compared to the failure set (by this we mean the set of people upon which the marketing campaign succeeded, and the one upon which it failed), and a predictive point of view, describing which characteristics of an individual correspond to a significantly higher probability of success of the marketing campaign on this individual than the average 59,6%, and at the opposite, we could also describe which characteristics correspond to a significantly lower probability than 59,6%.

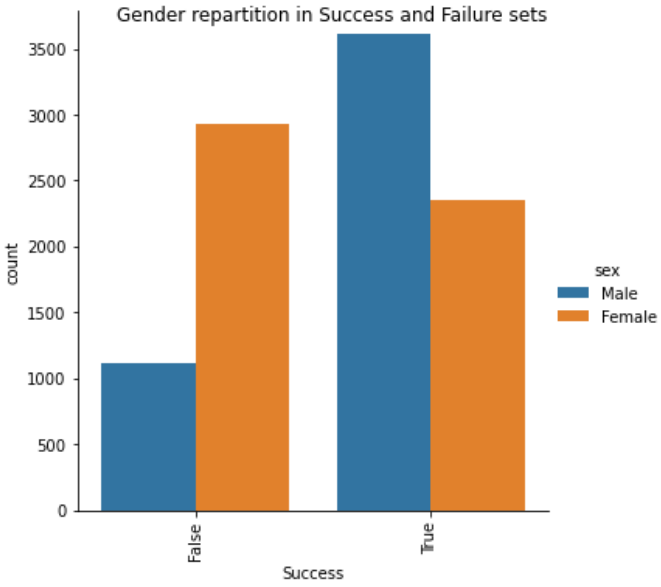
Gender

Descriptive view

The gender repartition in the success and failure set is :

Success	False	True
Female	0.724257	0.394799
Male	0.275743	0.605201

Among the success set, only 39,5% of the population are women while among the failure set, 72,4% are women. This is clearly significative : women are a smaller part of the success set than from the failure set. We indeed can understand here that they are less attracted by the campaign then men are.

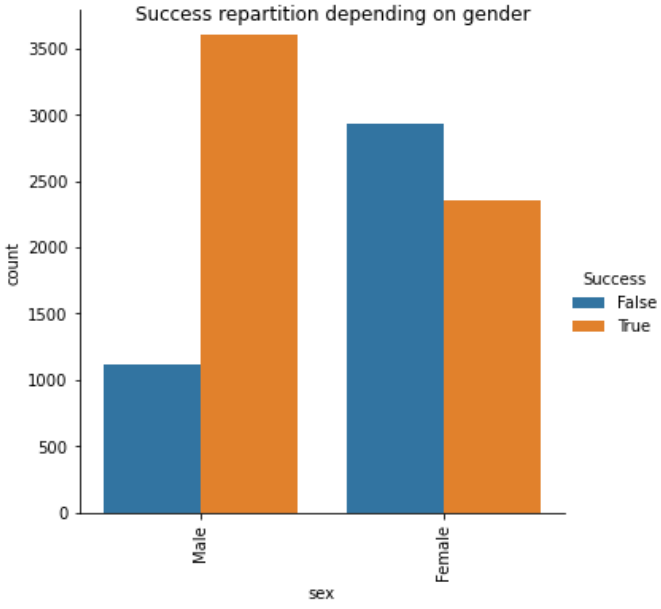


Predictive view

The success rates in the men and women population are :

sex	Female	Male
Success		
False	0.554272	0.235967
True	0.445728	0.764033

44,5% of all women were attracted by the campaign. It is significantly less than the 76,4% of the men population that was attracted.



Age

Descriptive view

The mean of ages in both sets are :

Success	False	48.069554
	True	49.806208

The success set is on average 2 years older than the failure one.

The first quartile of ages in both sets are :

Success	False	31.0	True	33.0

The median of ages in both sets are :

Success	False	47.0	True	49.0

The third quartile of ages in both sets are :

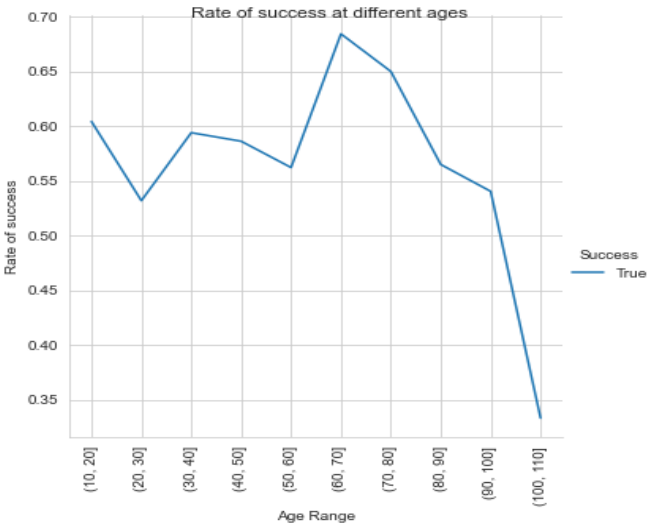
Success	False	62.0	True	66.0

The first quartile and the median of the success set are two years older than the failure set. The third quartile of the success set is three year older than the failure set.

We observe a small aging specificity of the success set compared to the failure set : it is a generally, older despite the fact that it is mostly composed of men, and that the men are globally younger.

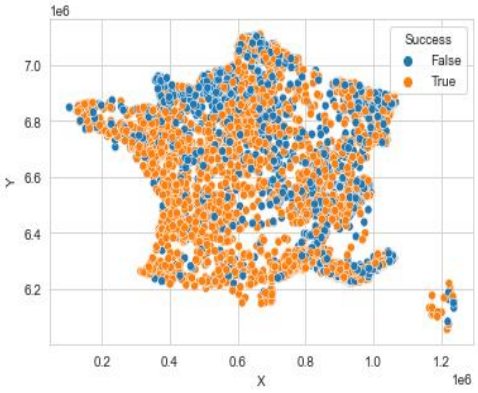
Predictive view

We observe that the success rate is between 53% and 60% for six categories of age out of 10. For people between 60 and 80 years old, the success rate is of 65% and 68%, significantly higher than for the others.
For people over 90 years old, that represent a nearly insignificant part of the population, the marketing campaign success rates fall.



Regions

Descriptive view

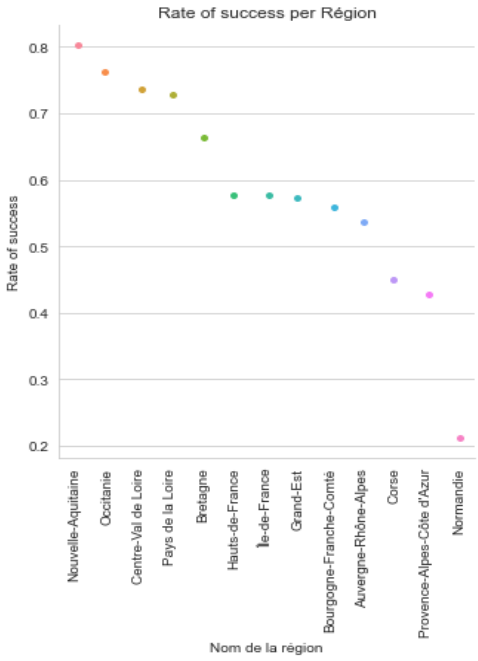


The regional repartition in the success and failure set is :

Success	False	True
Nom de la région		
Auvergne-Rhône-Alpes	0.138614	0.108725
Bourgogne-Franche-Comté	0.044307	0.038087
Bretagne	0.042574	0.056879
Centre-Val de Loire	0.024505	0.046477
Corse	0.008168	0.004530
Grand-Est	0.094059	0.085403
Hauts-de-France	0.098762	0.091443
Normandie	0.107673	0.019631
Nouvelle-Aquitaine	0.046535	0.128356
Occitanie	0.055693	0.120638
Pays de la Loire	0.036881	0.067114
Provence-Alpes-Côte d'Azur	0.111386	0.056376
Île-de-France	0.190842	0.176342

It is very different in both set : for example, Normandie represents 2% of the success set and 10% of the failure one. We understand that the success does depend on the region at which we look at .

Predictive view



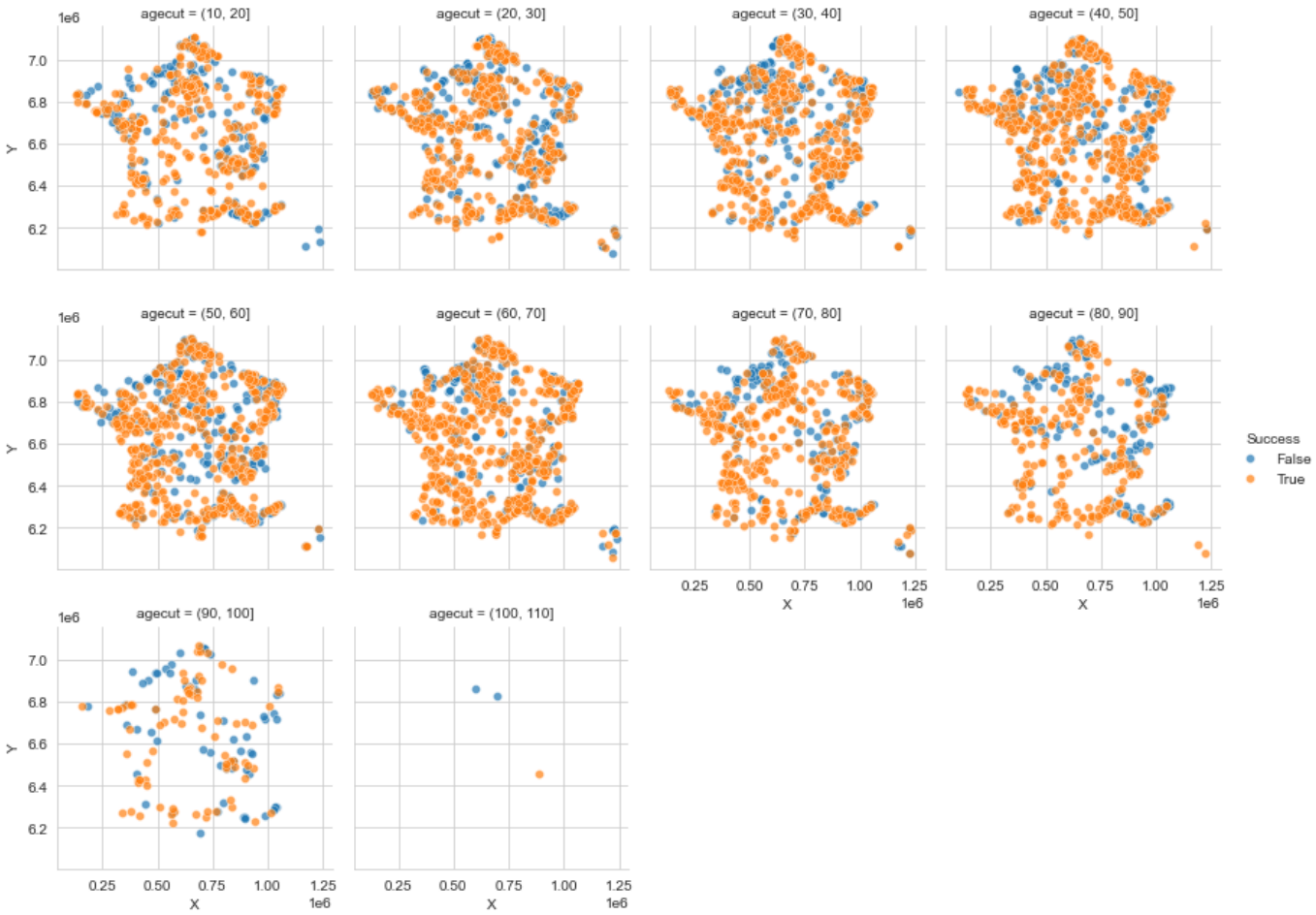
We obtain that Bretagne, Pays de la Loire, Occitanie, Nouvelle Aquitaine and Val de Loire have success rates between 66% and 80%, while Normandie success rate is only 21%.
The other regions have success rates between 42 and 57%. Those are huge differences : the success rate on the Atlantic coast is really higher than in other parts of France, and Normandie is an exception with an exceptionally low success rate.
Looking at the following graph, we see that the regional interpretation that we just had seem to be globally robust for any range of age that we might analyze.

If we decide to divide France in two according to the Atlantic coast against the rest of France, we obtain that there is a true effect based on this separation.

Area		Success		
Atlantic Coast	False	0.253973	True	0.746027
Rest of France	False	0.472863	True	0.527137

This separation is significant : the success rate goes from 74,6% among the Atlantic coast region to 52,7% among the rest of France.

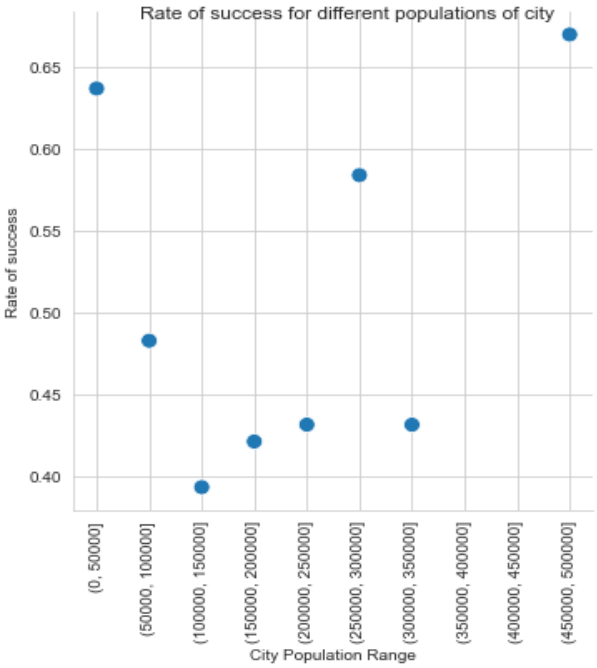
Looking at the following graph, we see that the regional interpretation that we just had seem to be globally robust for any range of age that we might analyze.



Cities population

Descriptive view

The mean of the population of the city of the individuals in both sets are :
Success False 56792.21 True 38806.885
The first quartile of the population of the city of the individuals in both sets are :
Success False 2239.00 True 3186.25
The median of the population of the city of the individuals in both sets are :
Success False 22549.0 True 8301.0
The third quartile of the population of the city of the individuals in both sets are :
Success False 70251.0 True 33609.0
We observe that there are more people from little cities and less people from big cities in the success set then in the failure set : the mean, median, first and third quartile of the number of habitants for the people in the success set are way smaller than those for the people in the failure set.



Predictive view

We obtain differences quite significative : 64% of success rate in cities of less than 50000 inhabitants is high, and we know that more than 75% of the population that we have lives in those smallest cities.
For the cities of more than 4000000 habitants, the success rate is also high, at 66%.

Clubs

Descriptive view

The part of the success and failure sets having clubs are :

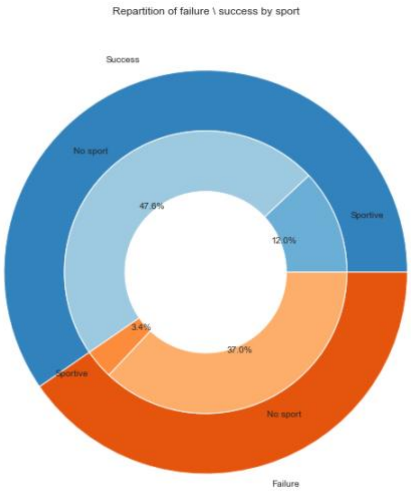
Success	False	True
Has no club	0.916089	0.798993
Has Club	0.083911	0.201007

Predictive view

The success rates depending on having a club or not are :

		Has no club	Has Club
Success	False	0.437315	0.22056
	True	0.562685	0.77944

The failure set has a slightly higher proportion of people not registered in clubs then the success set. The success rate is quite different : from 56% for not registered people to 78% for registered people.



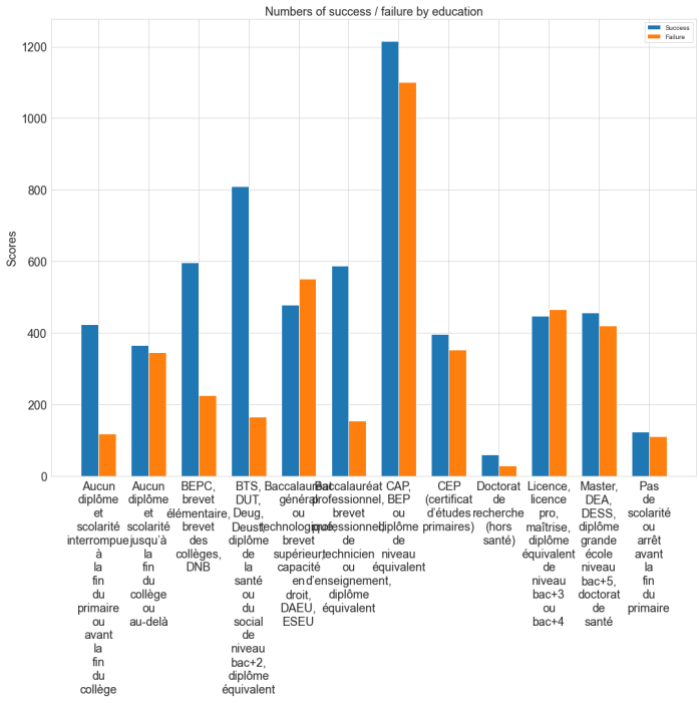
Education

Descriptive view

We can observe two major differences between both sets here : People whose highest degree is brevet des colleges or BEPC are a part nearly two times more important of the success set than of the failure set. Those who's highest degree is BTS or DUT are a part more than three times more important of the success set than of the failure set. Vice versa, people whose scholarship stopped before the end of college are a part nearly three times less important of the success set than of the failure set.

Predictive view

We also observe here strong disparities that can be significative : the success rate is between 46% and 53% for seven categories. It grows to between 67% and 82% on the other five categories (BTS-DUT/Bac pro/Arrêt avant fin du college/BEPC-brevet des collèges/Doctorat), which could generally be targeted.



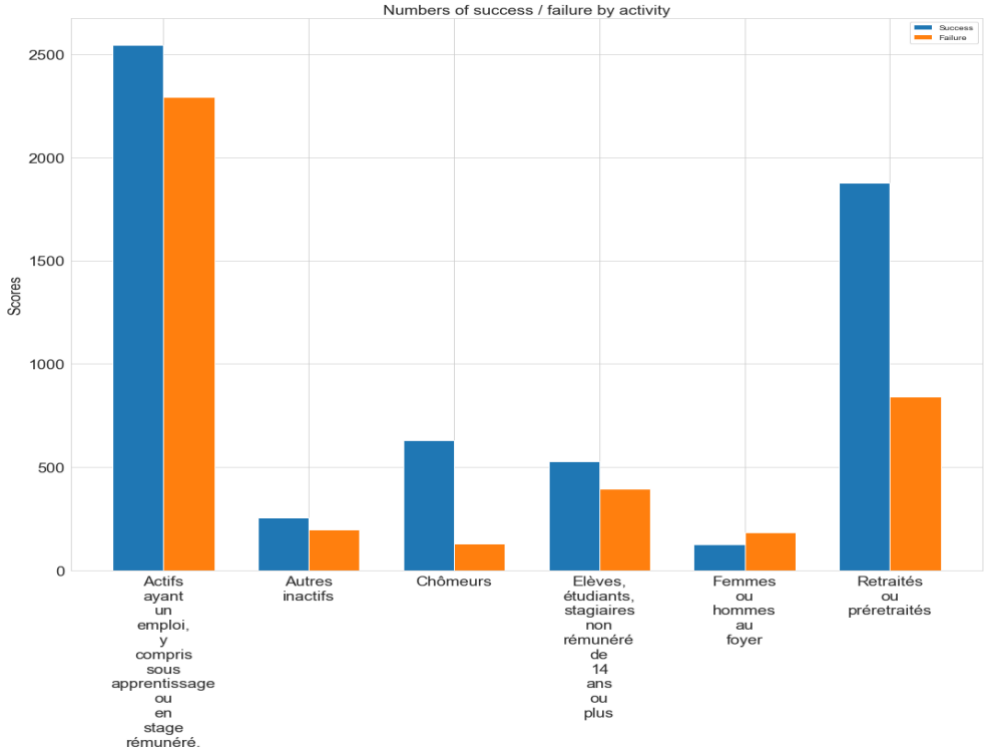
Activity

Descriptive view

The major difference is that the success set contains a part three times more important of unemployed people than the failure set.

Predictive view

There is a huge difference depending on the activity practiced : the campaign succeeded at a rate of 82% over unemployed people and 69% over retired people. It only succeeded at a rate of 40% over men and women not working.



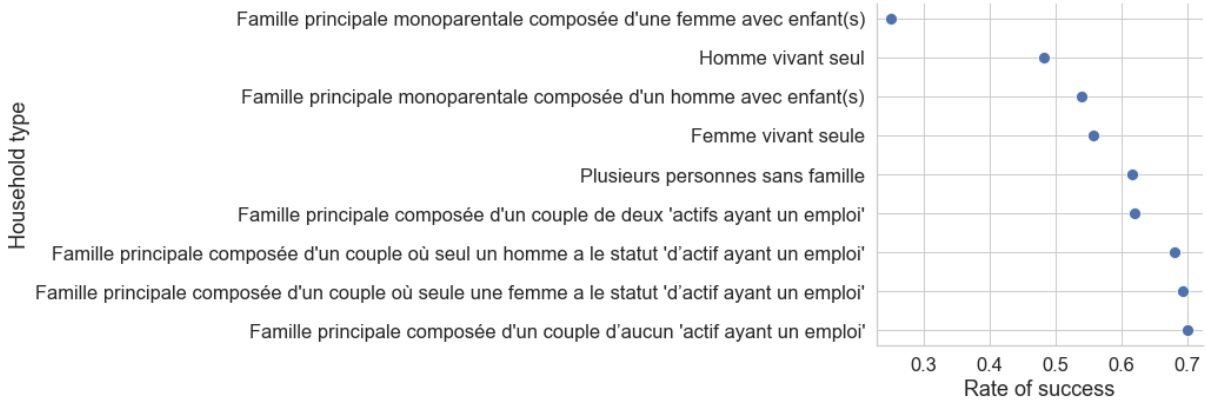
Household types

Descriptive view

The major difference between both set is that families with a single mother and families with a couple of unemployed parents are around two time more important in proportion to the others in the success set than in the failure set.

Predictive view

For families with a single parent or none working, success rates are between 68% and 70%. This is a significantly high level, compared to all other categories that are between 48 and 62%, except for the families with a single mother, over who's the success rate is 25%.



Conclusion for the person level analysis

As a remark, we could use all the descriptive view statistics that we obtain here to understand who are the clients that the marketing campaign succeeded in attracting. We then have an interesting global view of what our success set is composed of, when it is compared to the failure set.
We can conclude that the success set of people is different from the failure set of people on several points :

- There is significantly a bigger proportion of men in the success set then in the failure set.
- The success set is composed of older people on average then the failure set.
- It is composed of a bigger part of people living on the Atlantic coast then the failure set.
- It is composed of a bigger part of people registered in clubs then the failure set.
- It is composed of a bigger part of people from small cities then the failure set.
- We can derive also some differences in the average education level, in the most represented household types and on the occupations of the people of both sets

As a warning, we should consider the fact that some correlations exist in those variables (in the education level and occupation for example), so we should pay attention when we want to use those information in practice. Also, those conclusions cannot be directly used to forecast which client a similar marketing campaign would attract, since it would require to have a targeted population which is perfectly statistically similar to the one that we just obtained.
To do such forecasting, we need to look at the predictive point of view, which is precise on individuals probabilities for the marketing campaign to be successful, but at the contrary does not give a full picture of the people we successfully attracted, since it does not fully take into account the statistical properties of our initial set of targeted people. Before trying to observe the interdependence between variables in the grouped analysis part, and even if our analysis is not exhaustive and does not include real statistical tests, we can determine (naively for the moment) that the following characteristics of targeted peoples do increase the probabilities of success of the campaign (which we can estimate with the success rate observed here, which is on average of 59.6%) :

- A man is more probably attracted by the campaign then a woman (76,4% vs 44,5%)
- The campaign has significantly higher success rates on people between 60 and 80 years old.
- Bretagne, Pays de la Loire, Occitanie, Nouvelle Aquitaine and Val de Loire have success rates between 66% and 80%, way higher than the 59,6% on average
- The success rate is exceptionally high (78%) for people registered in clubs
- The success rate is higher (64-66%) for people in cities of less than 50000 or more than 4000000 inhabitants.
- For families with a single parent or none working, success rates are really high (68% to 70%).
- the campaign succeeded at a rate of 82% over unemployed people and 69% over retired people, which is also really higher than the average success rate.
- Some level of education correspond to higher success rates.

But to be more precise and even more accurate, we should check if those hypothesis do not depend on specific variables : we will indeed realize grouped analysis. The first one will be based on gender, to see if being a man or a woman leads to different conclusions, then we will apply the same kind of reasoning on department location and coordinates.

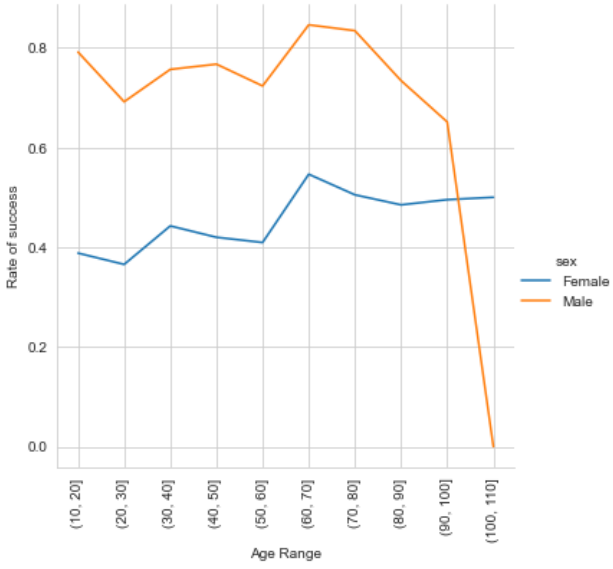
Grouped analysis

Grouping on the gender :

We then turn our interest on the gender types of the respondents, trying to find out if our type provokes some change in the probability of success. We remind that we registered 4721 male and 5279 female in the sample, then representing respectively 47,21% and 52,79% of the population. After dissociated each group, we underline that 76,40% of the male population gives success and only 44,57% for female. This gives rise to the restrictive assumption that male are more disposed to answer the marketing campaign than female. We therefore look for intrinsic individual 's components that possibly explained this gap between both gender groups. The leading goal of this grouped analysis is to emphasize dominant intrinsic variables that campaigning staff should know about in order to maximize the likelihood success. We will then try to estimate crossing effects between variables and so determine an ideal profile for both gender group.

Age effect :

We remind that we previously discovered that being older had a significantly positive impact in the probability of success of the marketing campaign. We also remind that women are older than men.



The mean age of the success and failure groups depending on the sex is :

Success	False	True
Female	49.010595	52.761156
Male	45.597846	47.878569

The median age of the success and failure group depending on the sex is :

Success	False	True
Female	48	53
Male	45	48

Women are 3 years older than men in average, and the median of their age is also 3 years higher than the median of men age.

Between both gender, there is a difference in the importance of the age in the repartition of the success. For men, the mean age of the success group is two years superior to the one of the failure group. For women, the mean age of the success group is nearly four years superior to the one of the success group. In median, the difference is of 3 years for men and of 5 years for women : this mean that the age has a more important impact in the results of the marketing campaign for women then for men : we can see on the graph that the success rates falls after 80 years old for the men, while it stays still for the women. In consequence, having an older individual is even more important when the target is a woman then when it is a man.

City population effect :

We remind that we had previously discovered that living in a city of less than 50000 inhabitants or more than 4000000 inhabitants had a significantly positive impact in the probability of success of the marketing campaign.

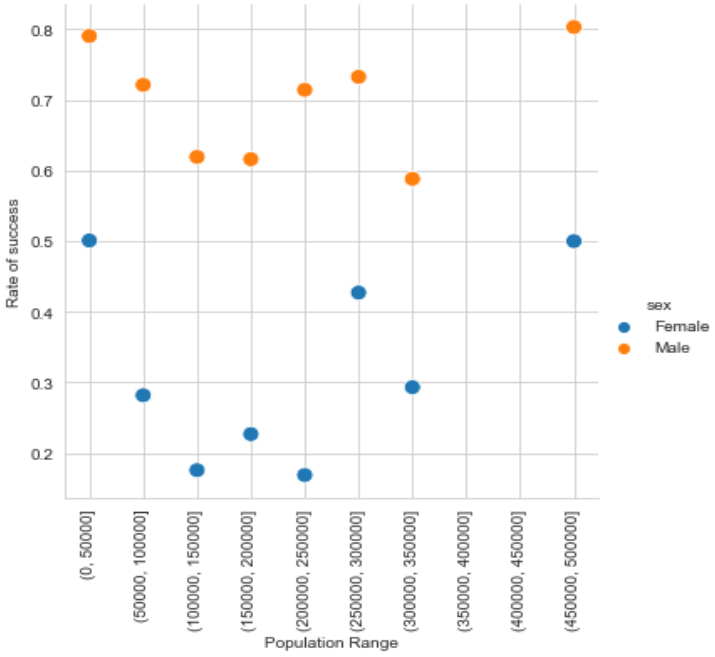
The mean population of city of the people depending on the sex is :

Female	44306.828755
Male	48047.836052

The median population of city of the people depending on the sex is :

Female	10762
Male	12399

Men live in cities with 4000 inhabitants more than women in average. The median of the population of their city is 2000 inhabitants higher than the one of women. This is significant since the median city population is globally 10000.



The mean population of city of the people of the success and failure groups depending on the sex is :

Success	False	True
Female	55107.738209	30875.693583
Male	61216.611311	43980.739950

The median population of city of people in the success and failure group depending on the sex is :

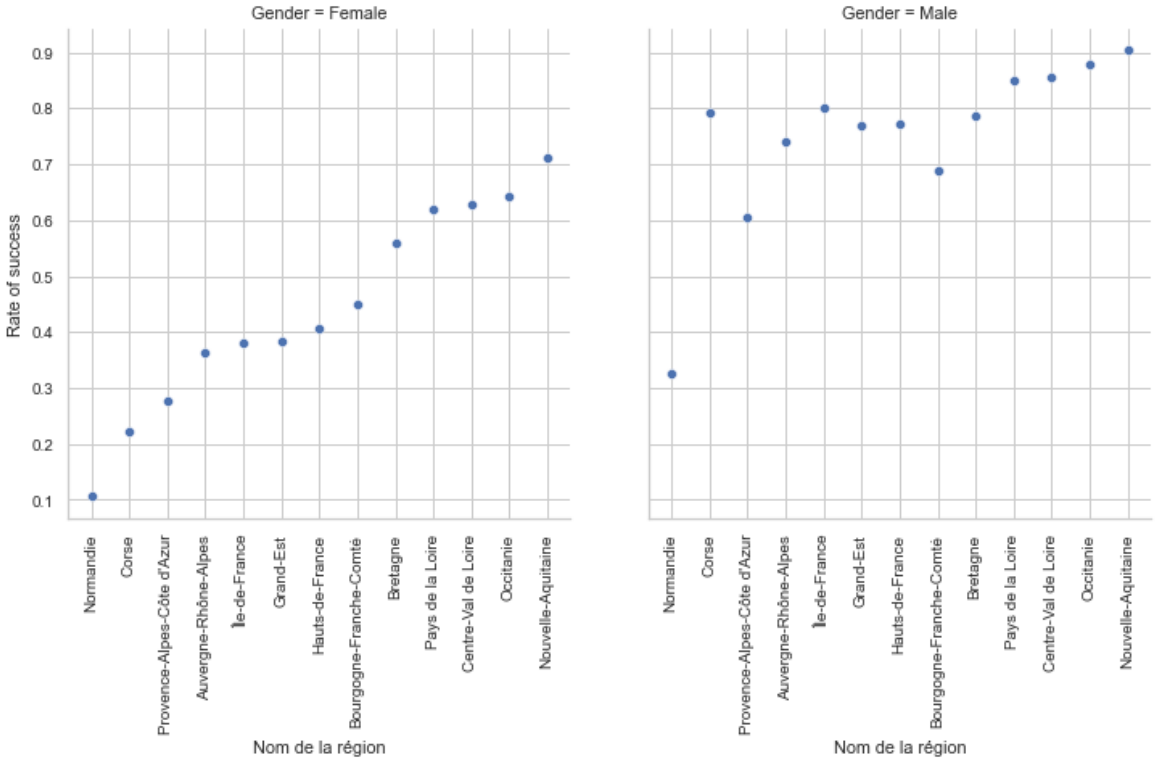
Success	False	True
Female	22445.5	6859.0
Male	22982.0	10033.0

For men, the mean of the population of their cities is 16000 higher in the failure group then in the success group. For women, the mean of the population of their cities is 24000 higher in the failure group then in the success group.

In median, the difference is of 13000 for men and 16000 for women. Those results show that the size of the city of the individuals also has a more important impact in the results of the marketing campaign for women then for men. We can clearly see in on the graph : little city population is way more receptive to the campaign than middle cities in the case of women, and just a bit more receptive in the case of men. It is also the case for cities of more than 4000000 inhabitants.

Region effect :

We remark directly on the following graph that the success rate is more conditional on women than on men : despite for Normandie and Provence, men have a success rate that does not importantly depend on the region, at the opposite of women, which success rate really changes gradually from 0.1 to 0.7 from the less successful region to the most successful region.



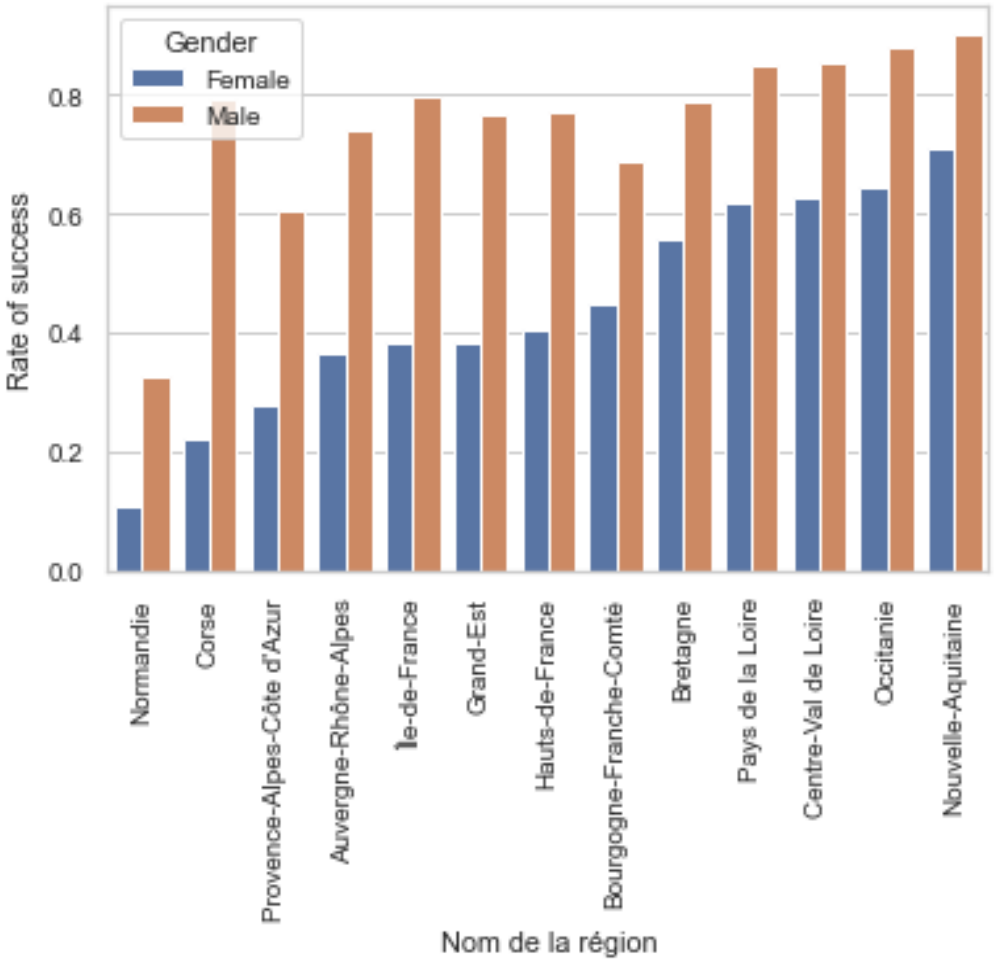
Indeed, we can conclude from this grouping that even if all our primary conclusions stand for both men and women, the effect of living in a certain region, living in a city of a given population or having a certain age is a more precise clue for a high success rate for the women than for the men.

Then, considering that the men population will probably be more targeted rather than the women population, the marketing staff could decide not to differentiate so much the analysis on the regions, apart from avoiding Normandie, on the age or on the size of the population of the city of the individual, since the success rates do not depend so much on those factor for men, contrarily to the general fact that what we first concluded. At the contrary, if the marketing staff decides to operate a more woman friendly campaign, and to target woman that were not so much attracted by the first campaign, all of those factor should be considered, since we remarked that they have a significative impact for women.

Grouping on the Region :

Gender effect :

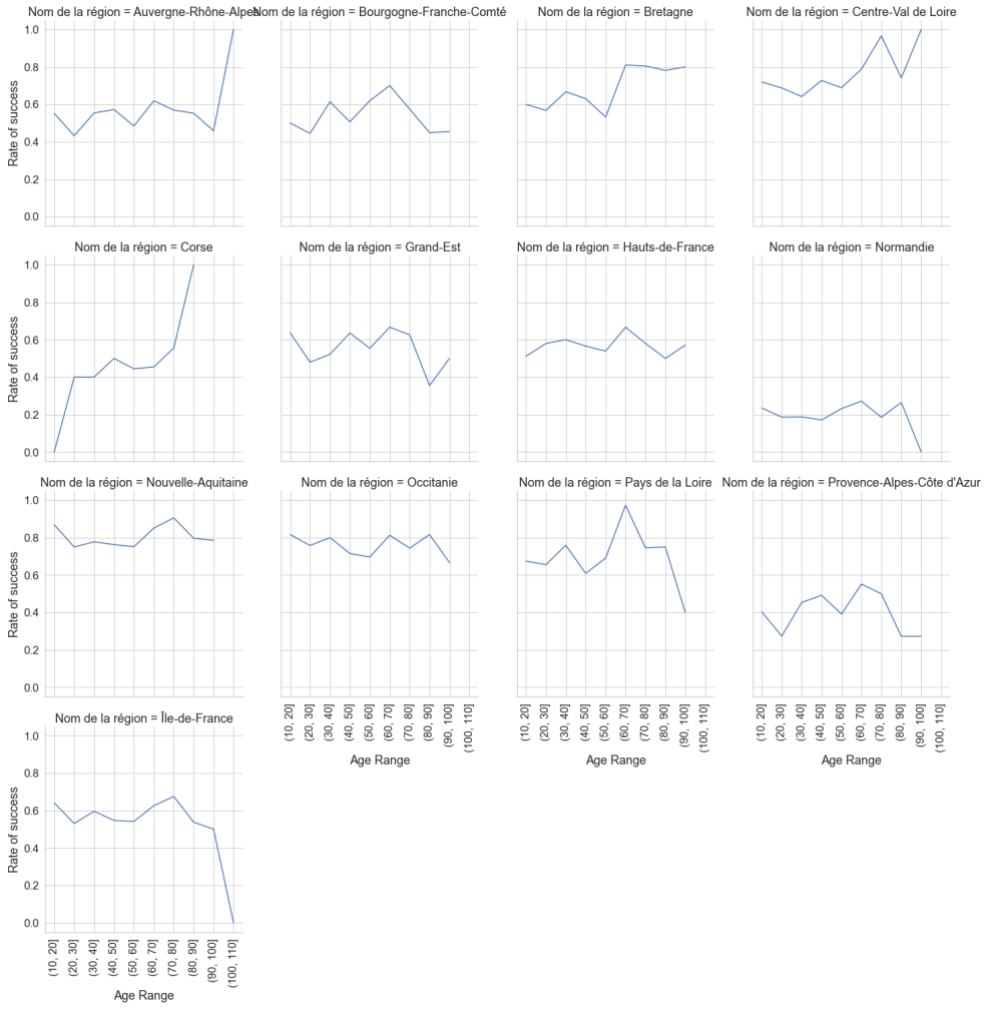
Looking at the following graph, we see that the fact of targeting a man implies a higher probability of success than targeting a woman in every region, but that there are regions where the difference is way higher than in others.



The gender is a way more important clue (due to the higher difference in the rate of success between men and women) in Normandie, Corse, Provence, Auvergne, Ile de France, Grand-Est and Hauts de France. In Bretagne, Pays de la Loire, Centre, Occitanie and Nouvelle Aquitaine (generally where the success rates are higher), the success rate difference is lower between men and women, so it is maybe not necessary to differentiate the target by gender in those regions.

Age effect

At the contrary, on the following graph, we can see that the relation between rate of success and age differentiated by region is really not stable (and we might not have enough individuals to have significant results, since each success rate represented corresponds in mean to 10000 individuals divided by 13 regions and by 13 age ranges, so 75 people on average).



Since age and regions are not uniformly distributed, we should consider that the effects in Corsica is not significant, neither the success rates in the lowest and highest ranges of ages. Those considerations being taken, we see that the success rates in each region of all ranges of ages are close to the mean. The only effect of age that we can obtain with this graph is that for regions with generally high success rates (such as Centre-Val-de-Loire, Bretagne, Pays de la Loire or Nouvelle-Aquitaine), the success rates are higher than the average for the 60 to 80 years old people.

Indeed, if the marketing staff want to target region by region, in regions where the initial success rate was low (such as Normandie, Corse, Provence, Auvergne) , it could try to target the men rather than the women, since gender effect is high for those regions. This differentiation is not necessary in regions where the initial success rate was high, since the gender effect is low there.
At the opposite, it could use a age differentiation in the regions where the initial success rate was high (such as Centre-Val-de-Loire, Bretagne, Pays de la Loire or Nouvelle-Aquitaine), targeting people between 60 and 80 years old, on which the marketing campaign seems to have more effect. It is not necessary to do such a differentiation in the regions where the initial success rates were low.

Conclusion

As a recall, we could not be exhaustive so we decided to compute the information that really mattered according to our instinct. We also only showed in this analysis the variables for which there existed an impact of the marketing campaign (it was not the case of the studying variable for example). We believe that our visualizations are clear enough so that there is no need to look individually at each dataframe of statistics that we have obtained during the computation of this analysis, and we recall that most of the important statistics that our program prints were pasted as such in this report (in blue).

The initial population of our dataset was French, contained a bit more woman then men, generally living in small town. There are no children in our individuals so the population is a bit older than the average French population. Half of the people are working, around ten percent were students and a bit more than a quarter was retired. Over this population, the success rate of the campaign was 59,6%.

To estimate over which variables should the marketing focus, we did a personal analysis of our dataset:

- A descriptive view of the set of people over who the campaign succeeded told us that compared to the average population studied, they are more men, older, living closer to the Atlantic coast, in smaller cities and more registered in clubs. Again, when compared to the average population, the people on which the marketing campaign succeeded are also more unemployed, a bigger part of them has not finished high school, and if they household has two parents, it is most likely that one of them is not working.

- A predictive view is more precise in telling on which person would the campaign succeed more or less probably : men have a higher probability of being attracted than a woman (76,4% vs 44,5% of success rate here). There is a significantly higher success rates on people between 60 and 80 years old, but also on people living close to the Atlantic coast (75% vs 53% if not) or on people registered in clubs (78%of success rate. It is higher (64-66%) for people in cities of less than 50000 or more than 4000000 inhabitants, for families with a single parent or none working high (68% to 70%), and it also succeeded really well over unemployed people (82%) and retired people (69%). But all of those statistics can just help to divide the population in two groups variable by variable.

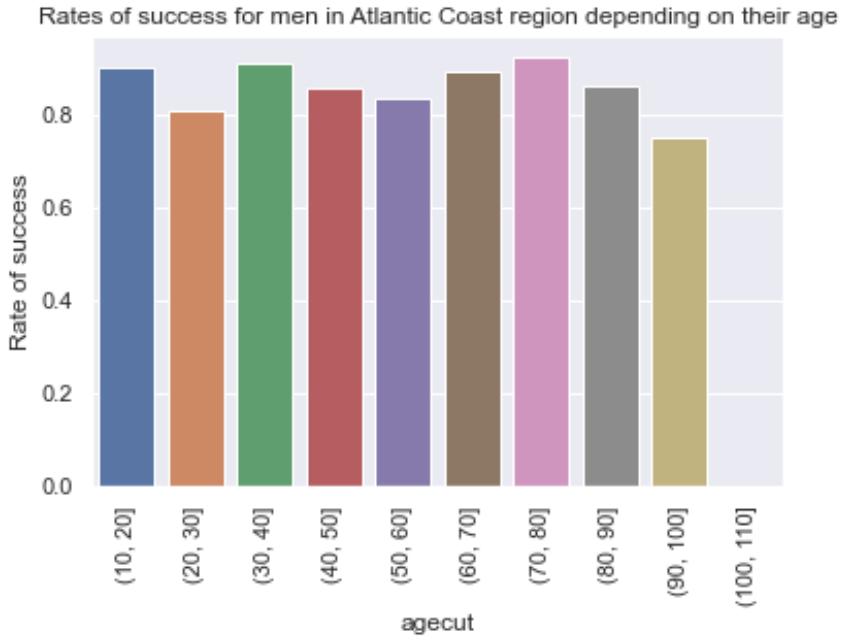
To study even more accurately the results of the marketing campaign, we realized grouped analysis :

- Our main results if we divide the population between men and women is that the results we obtain before about the age effect, the effect of the population of the city of the individual and the regional effect do not stand for the men anymore. Indeed, if we redo the same campaign and to target the men, it is not necessary to discriminate them between age ranges, regions or cities of different sizes. At the opposite, if the next campaign wants to attract women, it could use those factors to differentiate groups over which the campaign has a higher or lower probability of success.

- Our result if we divide the population region per region is that the gender has a more significant effect on regions where the average success rate is the lowest (Normandie, Provence and Auvergne) than on those where the average success rate is high (on the Atlantic Coast), but that in reverse, the age effect has a more significant effect on regions where the average success rate is high (such as Centre-Val-de-Loire, Bretagne, Pays de la Loire or Nouvelle-Aquitaine) than on the rest of France. If people targeted are on the Atlantic Coast, it could be successful to target the people between 60 and 80 years old, which is not necessary is the targeted people come from the rest of France.

Crossing the information that on average, the age effect over the outcome of the campaign is low over men, but that but that the age effect is high in the regions of the Atlantic coast, we decided that a final crossed-analysis might give us a really precise target if the age effect over the campaign is high for the men situated in the Atlantic coast region. As we can see in the following graph, it is not significantly the case : the success rates are independent of the age range for men on the Atlantic coast, and is very high (more then 80%), and significant, since there are 1503 men in this region. We should really consider targeting them if our aim is to target a population on which the precedent campaign succeeded.

There are 1503 men in the Atlantic Coast region we build.



Many more computation can be realized to detect even more link with the variable, but with our restricted population size, many analysis would not be significant anymore. Realizing this analysis, we understood the importance of having a huge initial population in order to be able to derived grouped analysis which would keep significance, and we realized the deep necessity to visualize clearly our statistics to be able to perfectly understand them.