

UNIVERSITÉ JEAN MONNET

MACHINE LEARNING & DATA MINING

Data Mining For Big Data: Study of the Groupama datasets

Valentin Benozillo
Josselin Marnat
Mathieu Viola
Rémi Viola



January 22, 2018

Contents

1	Introduction	3
2	Description of the dataset	3
3	Loading the dataset	3
4	Analysis of the different satisfaction surveys	4
4.1	How do we have to handle the dataset?	4
4.2	Satisfaction according to the ‘Typologie’	4
4.3	Satisfaction according to other criteria	4
4.3.1	Nature Personne	4
4.3.2	Segmentation Distributive	4
4.3.3	Tranche age	5
4.3.4	Type of surveys	5
4.4	Evolution of the level of satisfaction	5
4.4.1	Global evolution	5
4.4.2	According to other criteria	6
5	Analysis of the customers reclamations and terminations	6
5.1	Reclamations	6
5.1.1	Type	6
5.1.2	Typologie	6
5.1.3	Marche PSO	6
5.1.4	Departement	6
5.2	Termination	6
6	Comments analysis in satisfaction surveys	7
6.1	Analysis using term-frequencies of n-grams	7
6.2	Solutions to improve customers satisfaction	7
7	Conclusion	7
A	Levels of satisfaction	9
A.1	Global	9
A.2	Nature Personne	11
A.3	Segmentation Distributive	12
A.4	Type Survey	13
A.5	Tranche Age	14
A.6	Typologie	15
B	Evolution of the satisfaction	16
B.1	According to the previous mark	16
B.2	According to Marche CSP	19
B.3	According to Nature	21
B.4	According to Segmentation Distributive	21
B.5	According to Tranche d’age	23
B.6	According to Typologie	24
C	Reclamation & termination	26
D	Term frequency on comments	34

1 Introduction

This project aims at studying a dataset given by the University, in collaboration with Groupama, a French insurance group. As we will see in the next section, the dataset is composed of relational tables, and a set of surveys answers from customers. The goal is to find the actors – words, sequence of words – of (un)satisfaction according to the grades the customers gave in the surveys ; analysis the complaints in function of the customer category, profile, area ; and study the evolution of the customers satisfaction.

Section 2 presents the given dataset, section 3 explains how we loaded and the issues we had to face, section 4,5 and 6 presents the analysis of the customers satisfaction surveys according to various variables, the reclamations reasons, and the comments analysis in satisfaction surveys.

All the code used in this project is available in the folder `code/*.R`. You will have to change the path to the folder `big_data_project_confidential` at the beginning of each script inside the command `setwd(path_here)`.

All the figures and tables are located in the appendix at the end of the report.

2 Description of the dataset

This dataset the following tables, split in two subsets:

1. Relational Database (6 tables):

- 1.1. `BASE_Donnees_Clients` : informations about the customers (ID, age, living area, ...) ;
- 1.2. `BASE_Structure_Commerciale` : informations about the company's employees (agency, region, ...) ;
- 1.3. `BASE_Demandes_clients_hors_reclamations` : various requests from the customers ;
- 1.4. `BASE_Actions_rattachees_demandes` : actions that has be done for a request ;
- 1.5. `BASE_Reclamations_clients` : customers complaints ;
- 1.6. `BASE_Avantages_clients` : customers advantages ;

2. Satisfaction Surveys (16 tables) :

- `SATISFACTION_*` : tables containing satisfactions surveys done after a complaint, or randomly sent to customers. The main analysis will be done on these surveys.

All the tables that concerns customers are linked together with a customer ID (`ID_GRC`) Of course, this data is strictly confidential, and has been anonymized beforehand.

3 Loading the dataset

The dataset is a set of tables formatted in CSV (comma separated values), easily readable and loadable in most programming languages. It has the following form:

- the first row describes the names of the variables (IDs, dates, questions, ...);
- then, each row represents the data (either clients, an answer to a survey, a request, ...).

The CSV format works well when the semi-colons within a cell are escaped (*e.g.* `';` \rightarrow `\;`), but it wasn't the case, which is a real problem if we don't fix this issue. Since the number of columns is increased by how much there is unescaped semi-colons in a row, we decided to remove all these row using this property. It's a loss of data, but treating these rows would have been to much work, and it would be the work of the data pre-processor to escape the semi-colons efficiently.

Once these rows removed, we obtained a set of dataframes.

4 Analysis of the different satisfaction surveys

In this section, we will firstly try to see where are the more and the less satisfied customers of the society. After that, to complete our study, we will also analyze the level of satisfaction of the customers according to others criteria. Finally, we will analyze the evolution of the level of satisfaction of customers.

4.1 How do we have to handle the dataset?

We started by merging all the satisfaction surveys to get an overview of the satisfaction of all the customers. In a second step, we computed the average of the satisfaction for each customers who have completed several surveys. For this customers, we named the type of survey 'Average' to see that it is a mix of several survey. In this new base, we could compute the mean and the standard deviation of the level of satisfaction of the customers. In this case, the mean is 7.94/10 and the standard deviation is 2.33. With this 2 values, we decided to set thresholds to define 3 categories of customers. The most satisfied customers are those who have a level above the mean. The less satisfied customers are those who have a level under the mean minus the standard deviation. Between these 2 group, there is what we call the neutral part. After that we plotted several diagrams to answer the question. To see the code and reproduce our results, please edit the script `q1.R` by changing the first line and put the path where the survey are.

4.2 Satisfaction according to the 'Typologie'

To see the level of satisfaction according to the criteria 'Typologie', we plot the figure 5. This diagram is the aggregation of the number of dissatisfied customers (the left part) with the number of neutral one (the central part) and the number of satisfied one (the right part), according to their typology.

In this figure, the first thing which we can see it is that the majority of the customers come from the agricultural world. This is due to the history of the company. That is why, if we just compute the number of dissatisfied customers, we will see that is for 'rural dynamique' but it is due to the number of customers in this category. To have a better overview of the satisfaction level, we also plot the pies 11 according to each category of 'Typologie' to see the one where the percentage of dissatisfied customers is the most important.

In this set of figures, we can see that the proportion of satisfied and dissatisfied customers are more or less the same with just little differences. The most satisfied group is the 'Hors Territoire' one and by decreasing order, we find the 'Hyper Centre', the 'Peri Urbain', the 'Grande Périphérie Aisée', the 'Rural Dynamique' and the 'Rural Age' group which is the most dissatisfied one.

4.3 Satisfaction according to other criteria

To complete our study of the satisfaction of the customers, we have plotted the same diagrams for the other interesting criteria given in `BASE_Donnees_Clients.csv`. All these figure are available in the appendix A.

4.3.1 Nature Personne

As expected, for this criterion, the number of person is higher than the number of PM which represent associations, companies and others. It is very hard to see the difference of proportion in this situation. But, with the pies 6, we can see that the customers coming from companies are globally less satisfied than simple person.

4.3.2 Segmentation Distributive

For this criterion 2, we have to remember that :

- N = Nouveau,
- S1 = 'A laisser venir',

- S2 = 'A fidéliser',
- S3 = 'A redécouvrir et multi-équiper',
- S4 = 'A développer et fidéliser'.

The problem is that this field is not always complete in the file `BASE_Donnees_Clients`. Sometimes there is nothing, sometimes just a dot, and sometimes null and for all these situations, it seems there is no correlation with other criteria. The rest of the comparison is done without these values.

The most satisfied group is all new customers. After that, it is the sets S4, then S3, S2 and finally S1. The company have to work on these 2 last groups to improve its image.

4.3.3 Tranche age

The diagrams 3 and 10 for this criterion show that the number of young customers is very low but this is the group the most satisfied in proportion. The group of active is the biggest one and the most dissatisfied in average. It is worth noticing that the null set corresponds to the associations and companies set in this case.

4.3.4 Type of surveys

To be complete, when we have merged all the surveys we have kept the name of each ones. So we can analyze which one have the best marks and the worst. As a reminder, a big part of the dataset corresponds to the average of multiple satisfaction marks. It corresponds to the first bar of 4. After that, the most important parts of the dataset correspond to the field 'degats vehicule hors collision' and 'autres evenements ou dommages'.

All the pies of figure 8 correspond to customers having filled only one satisfaction survey. As we have say, the most importants are 'degats vehicule hors collision' and 'autres evenements ou dommages'. The second one correspond to one where the level of satisfaction is the highest with 'bris de glace(auto)'. The first one is not the worst. To find the worst, we have to look at 'demande' and mainly 'evenement entre deux vehicules' where there is more or less the same number of satisfied customers and dissatisfied ones.

Concerning the average computed diagram 9, we can see that the level of satisfaction is not bad. The proportion of dissatisfied customers is the second smallest after 'autres evenements ou dommages'. There is just a big neutral part in this diagram.

4.4 Evolution of the level of satisfaction

The third question of the company was to know the evolution of the level of satisfaction of its customers. To answer this question, we only kept the customers who filled several survey. After this selection, we built a database which contains the initial level of satisfaction and the difference between this mark and the next. If a customer filled more than 2 surveys, we only computed the difference between 2 consecutive surveys (With 3 surveys, the difference between the first and the second and between the second and the third). According to different criteria, we plotted several diagrams to analyze the problem. To see the code and reproduce our results, please edit the script `q3_1.R` by changing the first line and put the path where the survey are.

4.4.1 Global evolution

In the figure 12, we can see that most of the time the mark evolves of no more than one or two points in the decrease or in the increase. It often remains stable.

We also plotted several diagrams to see the evolution according to the previous mark. Thanks to the figures 17 and 18, we can say that it is after a 8, a 9 or a 10 that the evolution is the most frequent. It is also because it is the most frequent given marks.

4.4.2 According to other criteria

We also have computed the evolution of the level of satisfaction according to the 'Typologie', the 'Marche CSP', the 'Tranche d'age', the 'Segmentation distributive' and the 'Nature'. Because of a lack of time, we do not have to look farther in this direction. The different figure are available in the appendix B

5 Analysis of the custommers reclamations and terminations

The goal of this section is to better understood who are the client who do a reclamation or terminate their contract. To reproduce our result please edit the script `q2_1.R` by changing the first line and put the path where the table are.

5.1 Reclamations

5.1.1 Type

Firstly we simply compute the proportion of each possible type of reclamation 31 (sinistre, gestion contrat, cotisation, resiliation, souscription, contrat, commerciale, encaissement). As we can see on the pie chart the most hot topic is "sinistre" that's not very remarkable, but the second one is "gestion contrat" and the number of reclamations with this type is more than two times the number of reclamations with "cotisation". If we suppose that the real meaning of "gestion contrat" is : the client claim information about his contract, then maybe the custommers are not enough aware about their contract, so it's will be interesting to analysis witch are the real topic of their reclamation.

5.1.2 Typologie

Thanks to the "client" table we are able to know what is the "typologie" of a client for each reclamation. In a first hand we can simply count the number of reclamations for a given "typologie" 32. But it exist a bias, by looking this chart we can deduce that custommers "rural dynamique" has more "reclamation" than the others, but they also are the most representative categories in the "client" table. So if we look in term of proportion, by dividing our result by the number of "client" with this "typologie", we got a second chart 33, and as we can see the "typologie" as no impact on the proportion of reclamation.

5.1.3 Marche PSO

We do the same thing as before but with the characteristic "MARCHE_PSO" (agricole, acps, particulier, collectivites) fig. 34,35. But in this case we can see the class "agricole" have more reclamation (in proportion) than the other class. We compute their "TYPE" of reclamation fig. 37.

5.1.4 Departement

Thanks to the "COD_INSEE" in the table "client" we are able to know their department. So we compute the proportion of reclamation per department fig. 38,39.

5.2 Termination

We do the same study for the "resiliation" table.

- Typologie: fig. 41,42
- MARCHE_PSO: fig. 43,44
- Departement: fig. 45

6 Comments analysis in satisfaction surveys

6.1 Analysis using term-frequencies of n -grams

The goal of this section was to understand the reasons for customer satisfaction and dissatisfaction. To reproduce our results please edit the script `q2_keywords.R` by changing the first and put the path where the table are.

At first, we gathered all the comments from the satisfaction according to the satisfaction rating: one set for the good ratings – the highest 30%) and the bad ratings. Then, we removed the French stop-words: prepositions, articles, pronouns ; as well as numbers. And retrieved the roots of words by stemming, in order to remove conjugations, feminine or plural forms. This allowed us to group the words together more efficiently. Finally, we computed the n -grams for $n \in [1..5]$. For example, the sentence ‘I like your car’ has the following 2-grams: ‘I like’, ‘like your’, and ‘your car’ ; and we extracted the 20 most frequent terms on this n -grams dataset. In order to keep only the meaning-full n -grams, we removed from the satisfied and unsatisfied tables the n -grams who appears in both tables – *i.e.* we removed the intersection. Thus, the n -grams are exclusively attributed to one set of comments.

We obtained what is shown in tables 1 to 5. You can see for each set of two tables: 1) on the left: the 20 most frequent n -grams according to satisfied customers and 2) on the right: according to the unsatisfied clients.

6.2 Solutions to improve customers satisfaction

According to the previously mention tables, we can observe that satisfaction comes along terms like ‘rapid(e)’, ‘bon accueil’, ‘bon(s) conseil(s)’, ‘répons(e) rapid(e)’, ‘très bien renseign(é)’, and unsatisfaction with ‘trop long’, ‘trop cher’, ‘del(ai) trop long’, ‘sinistr(e) non respons(able)’, ‘toujour(s) rien reçu’. We can deduce the following improvements:

- keep welcoming people the same way ;
- treat the clients request more quickly or explain why it’s long ;
- explain clearly every details of the insurances ;
- discuss more with the customer for the responsibility of a sinister.

7 Conclusion

We have seen several approaches to analyze the given dataset: data preprocess, satisfaction analysis, term frequency on n -grams, evolution of the grades. We get graphical results on the level of satisfaction that reveals the general thoughts of the customers according to various parameters, we observed that generally the grade are not changing by much over time for each customer, and finally retrieved the lexicon of satisfied / unsatisfied customers in order to propose solutions to increase their satisfaction.

Finally, we hope that these results will be useful to the company Groupama, who provided the data.

Thank you for reading.

Appendix

List of Figures

1	Level of satisfaction according to NATURE PERSONNE	9
2	Level of satisfaction according to SEGMENTATION DISTRIBUTIVE	9
3	Level of satisfaction according to TRANCHE AGE	10
4	Level of satisfaction according to TYPE OF SURVEY	10
5	Level of satisfaction according to TYPOLOGIE	11
6	Level of satisfaction according to NATURE PERSONNE	11
7	Level of satisfaction according to SEGMENTATION DISTRIBUTIVE	12
8	Level of satisfaction according to TYPE OF SURVEY	13
9	Level of satisfaction according to TYPE OF SURVEY - Computed Average	14
10	Level of satisfaction according to TRANCHE AGE	14
11	Level of satisfaction according to TYPOLOGIE	15
12	Evolution of the grade between two satisfaction surveys - Global	16
13	Evolution of the grade after a 0 and a 1	16
14	Evolution of the grade after a 2 and a 3	17
15	Evolution of the grade after a 4 and a 5	17
16	Evolution of the grade after a 6 and a 7	18
17	Evolution of the grade after a 8 and a 9	18
18	Evolution of the grade after a 10	19
19	Evolution of the grade for 'marche CSP' TNS and ACPS	19
20	Evolution of the grade for 'marche CSP' AGRICOLE and PARTICULIER	20
21	Evolution of the grade for 'marche CSP' RETRAITE	20
22	Evolution of the grade for 'nature' PM and PP	21
23	Evolution of the grade for 'segmentation' N	21
24	Evolution of the grade for 'segmentation' S1 and S2	22
25	Evolution of the grade for 'segmentation' S3 and S4	22
26	Evolution of the grade for 'tranche d'age': 20-44 and 45-64 ANS	23
27	Evolution of the grade for 'tranche d'age': 65 AND ET PLUS and NULL	23
28	Evolution of the grade for typologie - RURAL AGE and RURAL DYNAMIQUE	24
29	Evolution of the grade for typologie - PERIURBAIN and GRANDE PERIPHERIE AISEE	24
30	Evolution of the grade for typologie - HYPER CENTRE and HORS TERRITOIRE	25
31	Number of 'reclamations' according to their TYPE	26
32	Number of 'reclamations' according to the client TYPOLOGIE	26
33	Number of 'reclamations' according to the client TYPOLOGIE in proportion of the client of this categorie	27
34	Number of 'reclamations' according to the client MARCHE_PSO	27
35	Number of 'reclamations' according to the client MARCHE_PSO in proportion of the client of this category	28
36	Number of 'reclamations' according to the client MARCHE_PSO in proportion of the client of this category	28
37	Number of "agricole" reclamation according to the TYPE of reclamation	29
38	Number of 'reclamations' according to the client department	29
39	Number of 'reclamations' according to the client department	30
40	Number of 'reclamations' according to the client department in proportion of the client of this category	30
41	Number of 'resiliation' according to the client TYPOLOGIE	31
42	Number of 'resiliation' according to the client TYPOLOGIE in proportion of the client of this category	31
43	Number of 'resiliation' according to the client MARCHE_PSO	32
44	Number of 'resiliation' according to the client MARCHE_PSO in proportion of the client of this category	32
45	Number of 'resiliation' according to the client department	33

List of Tables

1	1-grams frequencies for satisfied and unsatisfied	34
2	2-grams frequencies for satisfied and unsatisfied	34
3	3-grams frequencies for satisfied and unsatisfied	35
4	4-grams frequencies for satisfied and unsatisfied	35
5	5-grams frequencies for satisfied and unsatisfied	36

A Levels of satisfaction

A.1 Global

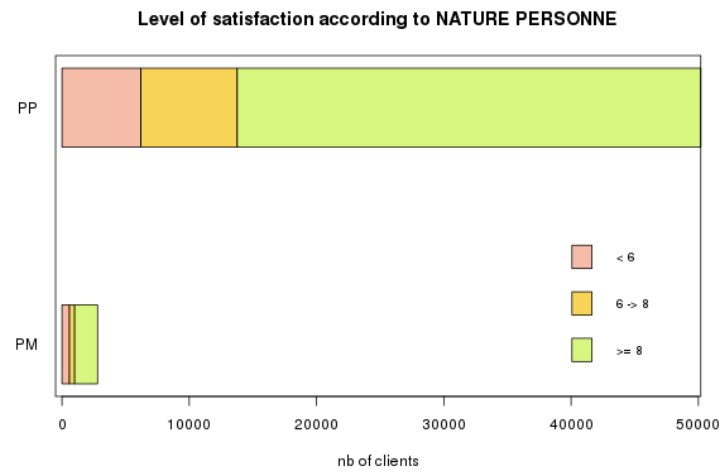


Figure 1: Level of satisfaction according to NATURE PERSONNE

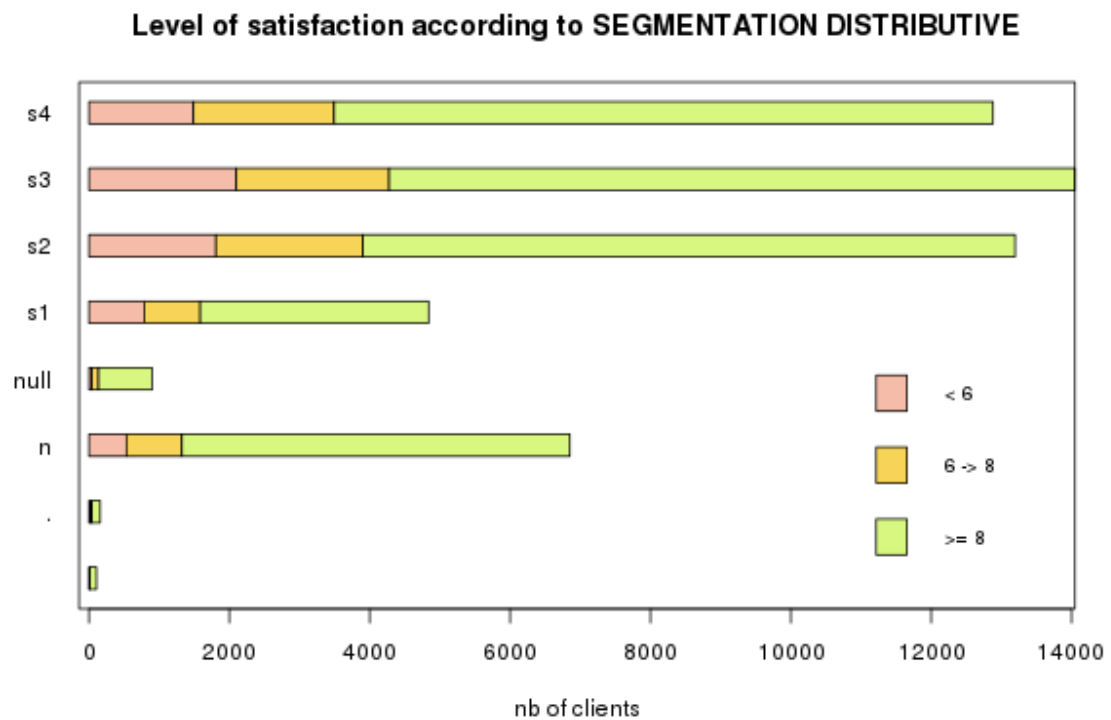


Figure 2: Level of satisfaction according to SEGMENTATION DISTRIBUTIVE

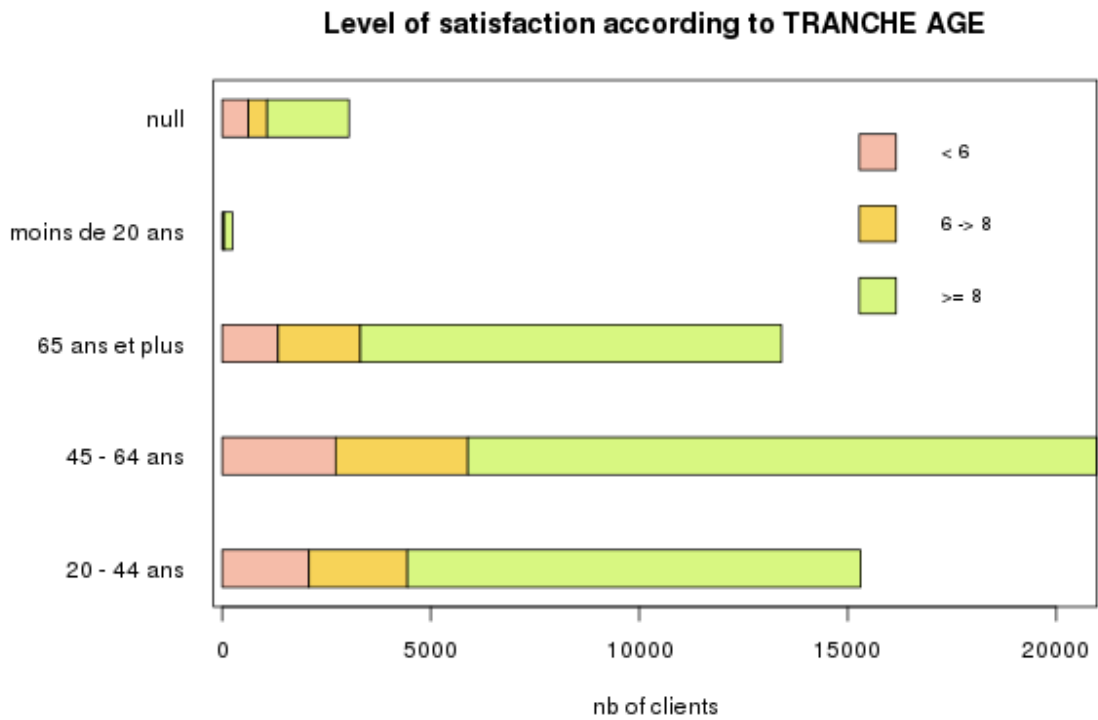


Figure 3: Level of satisfaction according to TRANCHE AGE

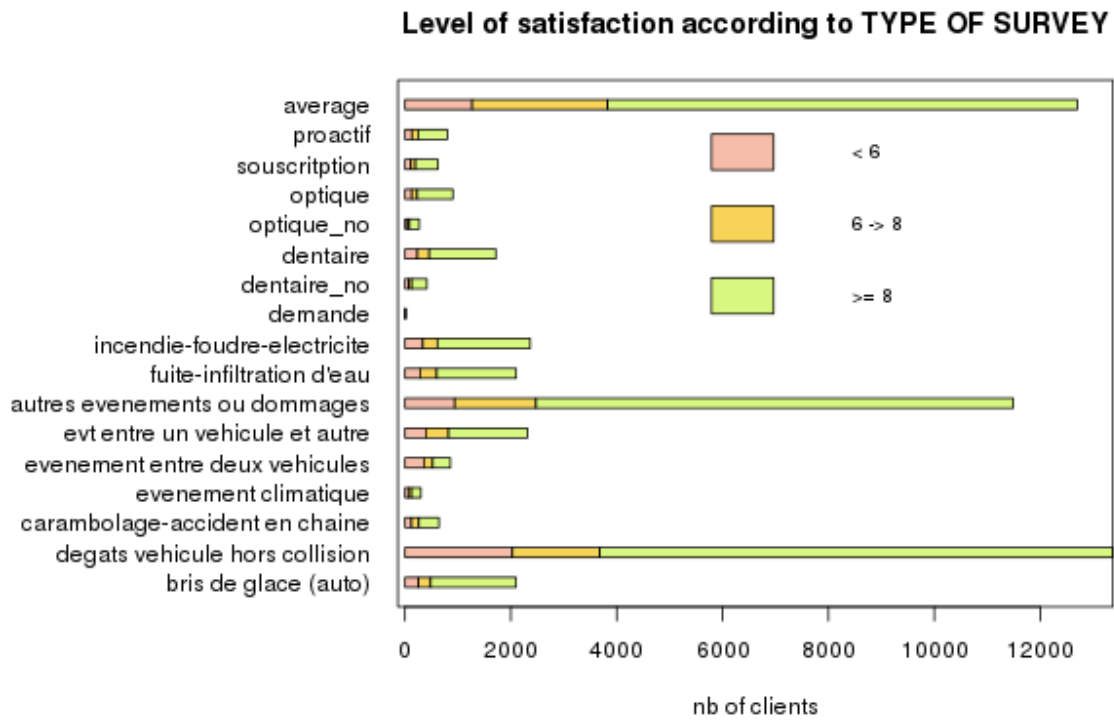


Figure 4: Level of satisfaction according to TYPE OF SURVEY

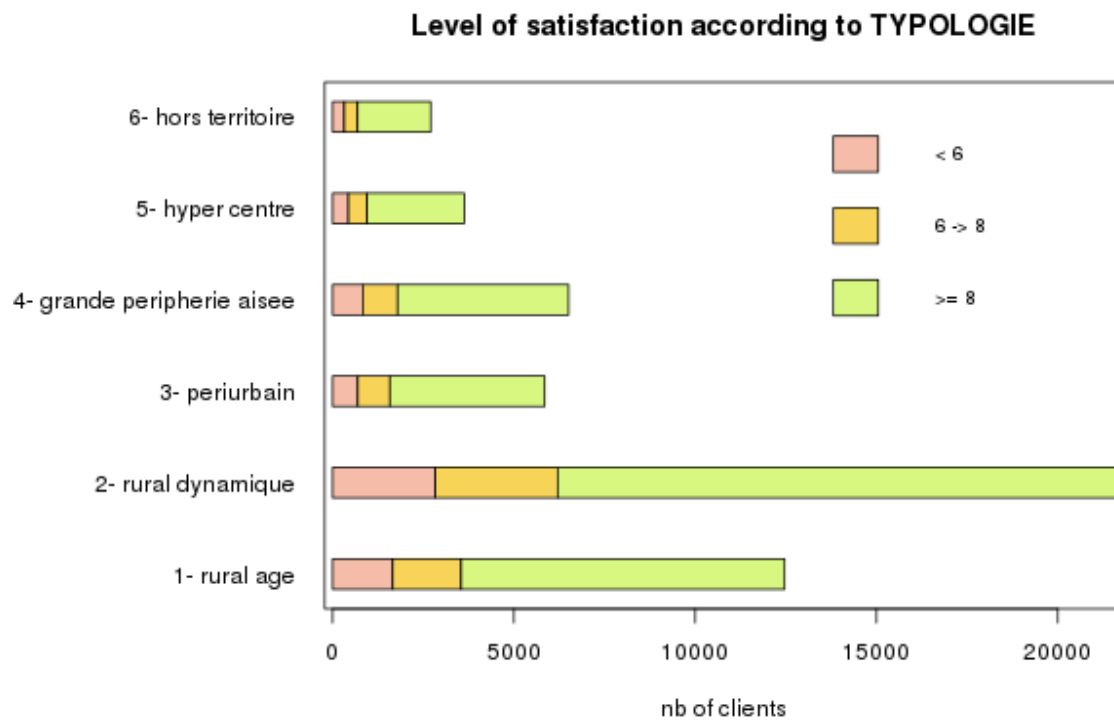


Figure 5: Level of satisfaction according to TYPOLOGIE

A.2 Nature Personne

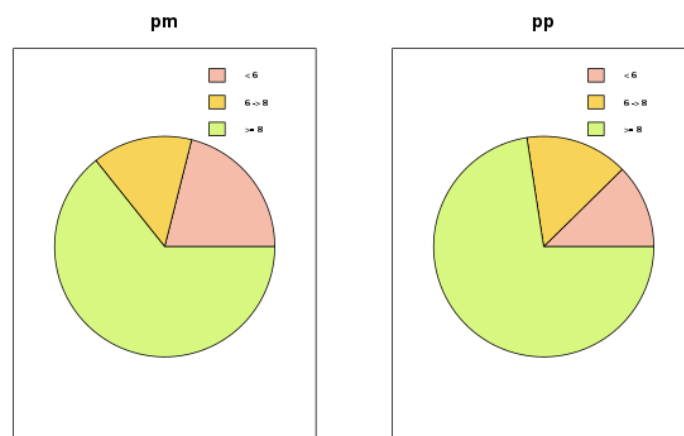


Figure 6: Level of satisfaction according to NATURE PERSONNE

A.3 Segmentation Distributive



Figure 7: Level of satisfaction according to SEGMENTATION DISTRIBUTIVE

A.4 Type Survey

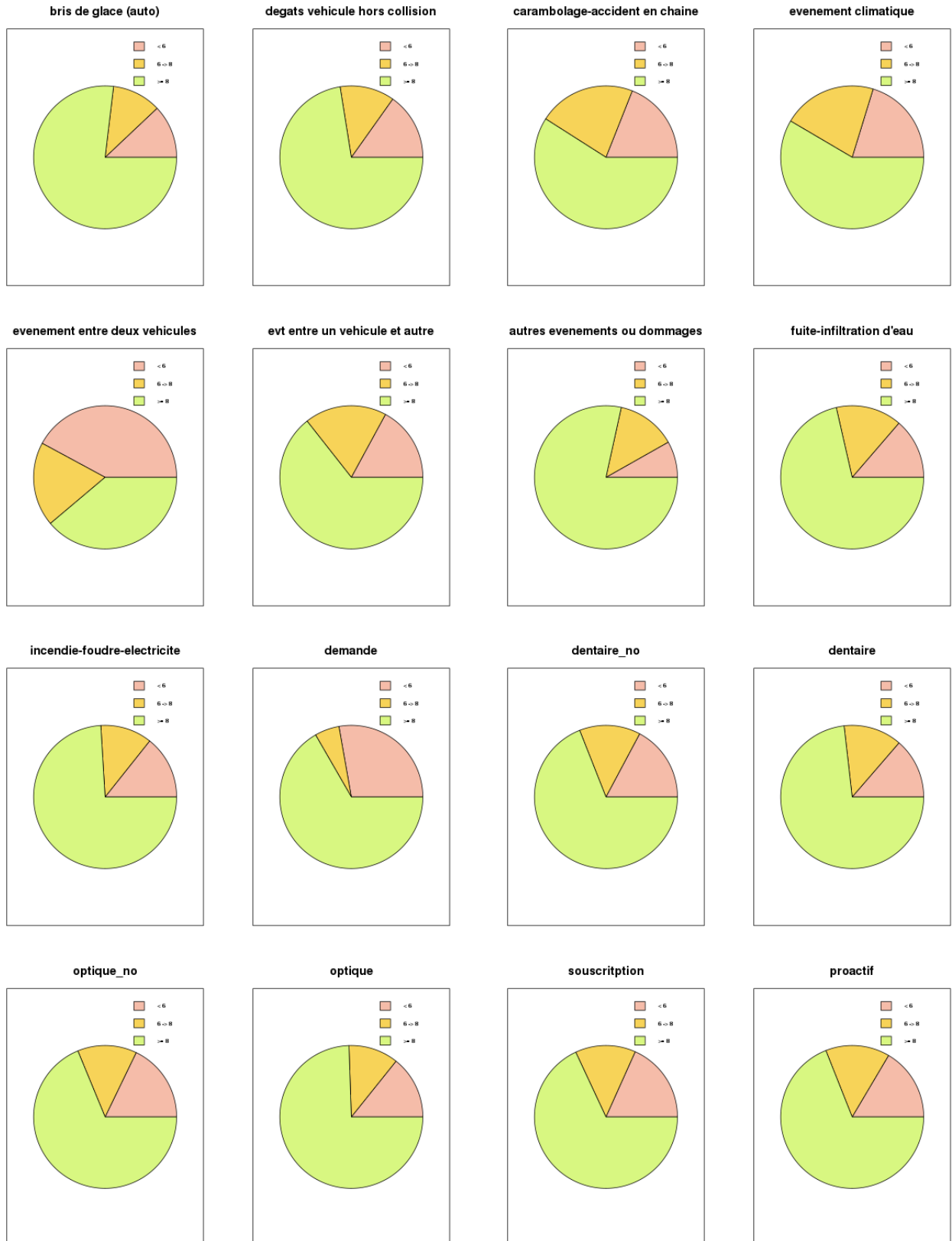


Figure 8: Level of satisfaction according to TYPE OF SURVEY

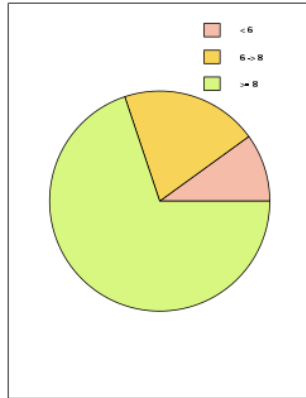


Figure 9: Level of satisfaction according to TYPE OF SURVEY - Computed Average

A.5 Tranche Age



Figure 10: Level of satisfaction according to TRANCHE AGE

A.6 Typologie

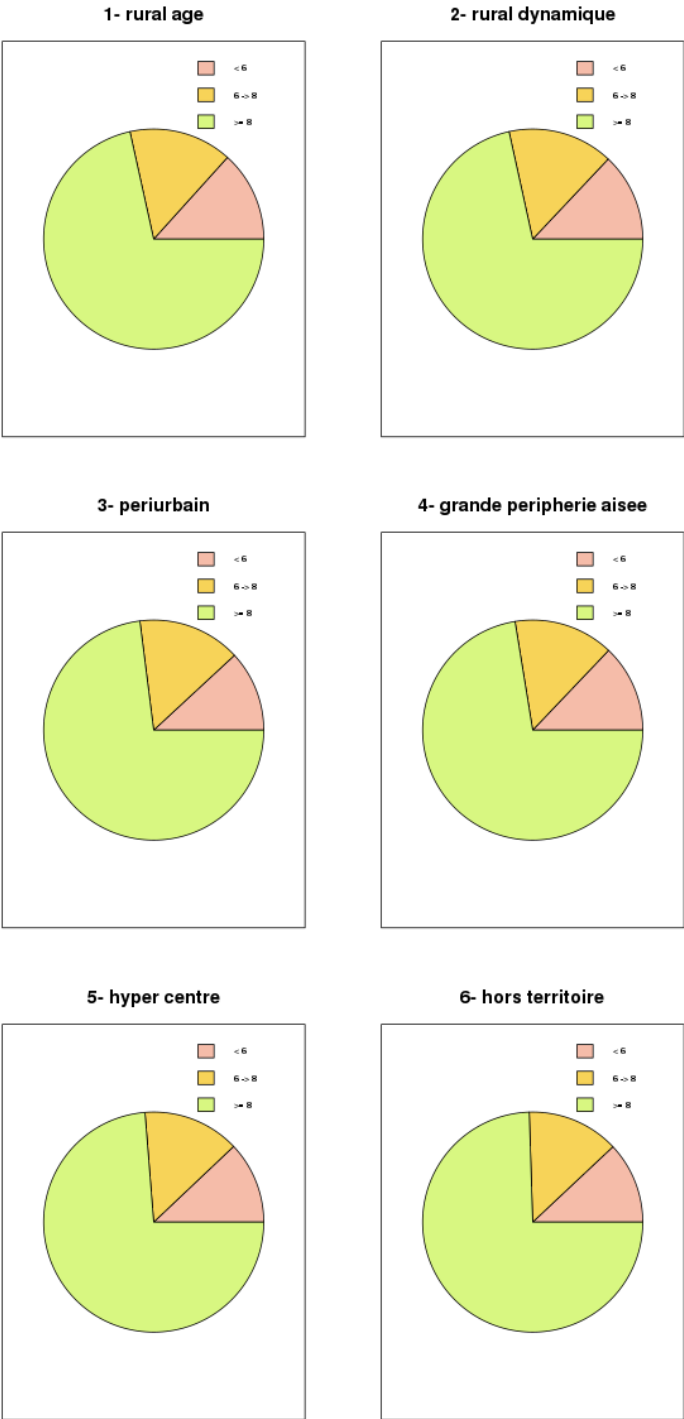


Figure 11: Level of satisfaction according to TYPOLOGIE

B Evolution of the satisfaction

B.1 According to the previous mark

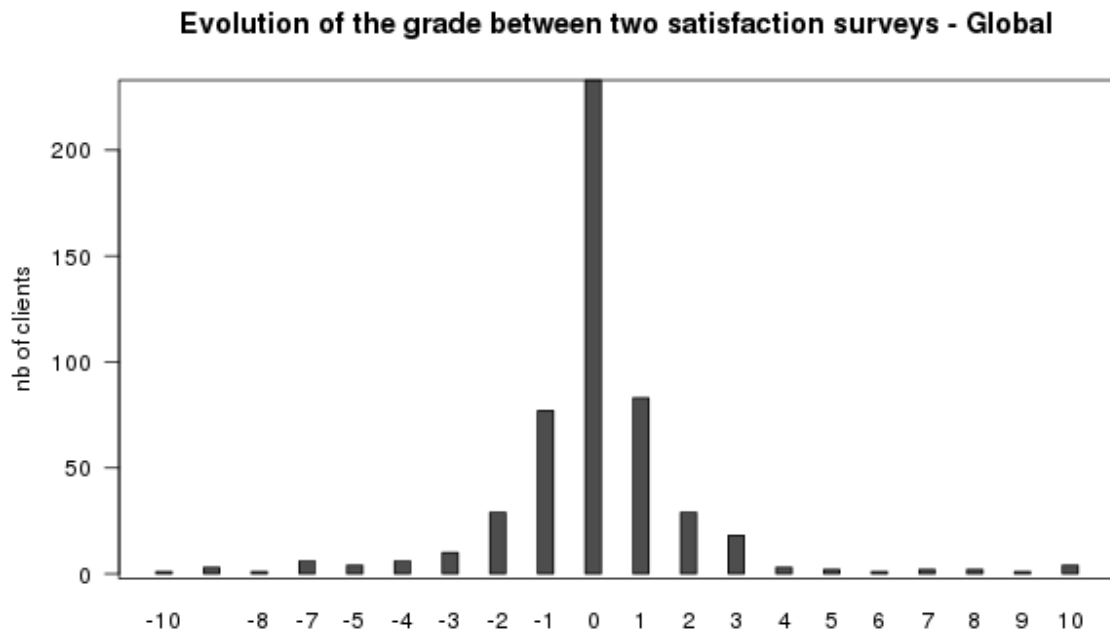


Figure 12: Evolution of the grade between two satisfaction surveys - Global

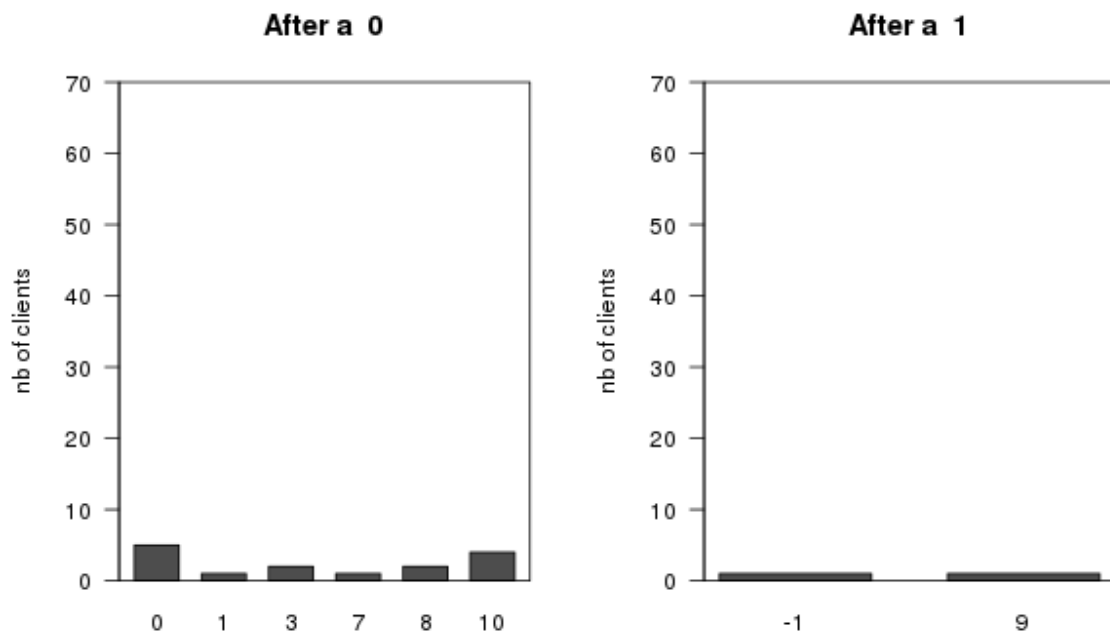


Figure 13: Evolution of the grade after a 0 and a 1

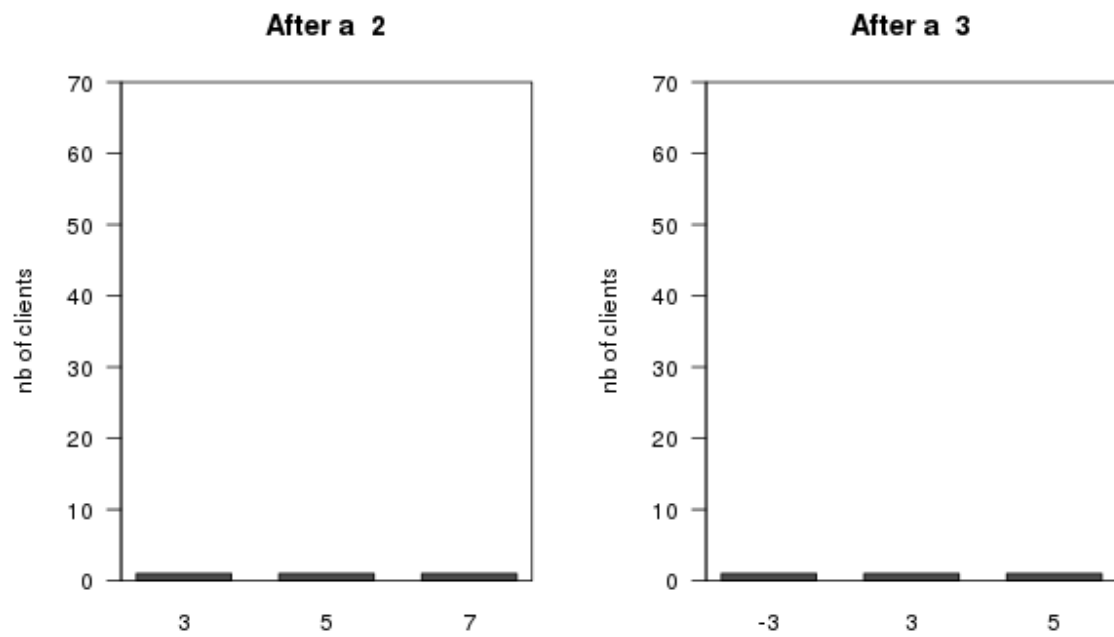


Figure 14: Evolution of the grade after a 2 and a 3

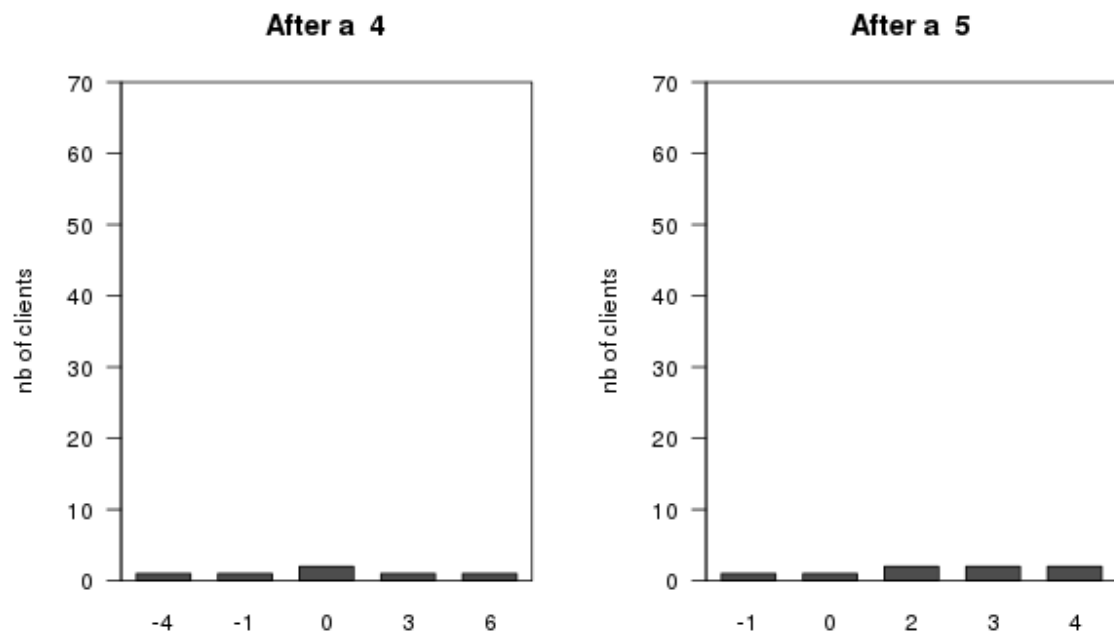


Figure 15: Evolution of the grade after a 4 and a 5

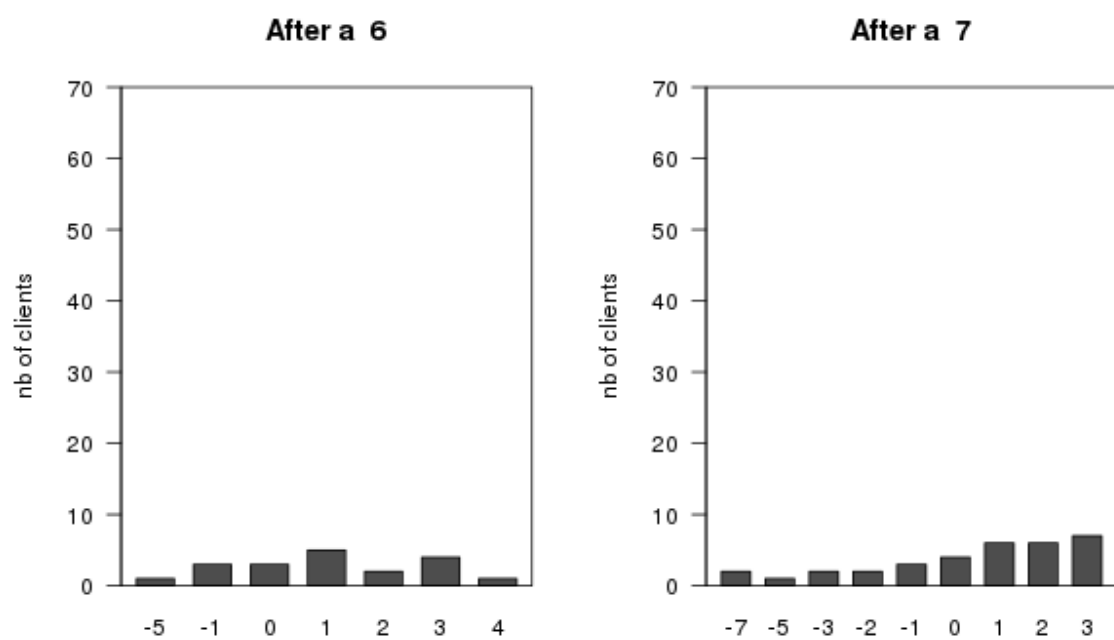


Figure 16: Evolution of the grade after a 6 and a 7

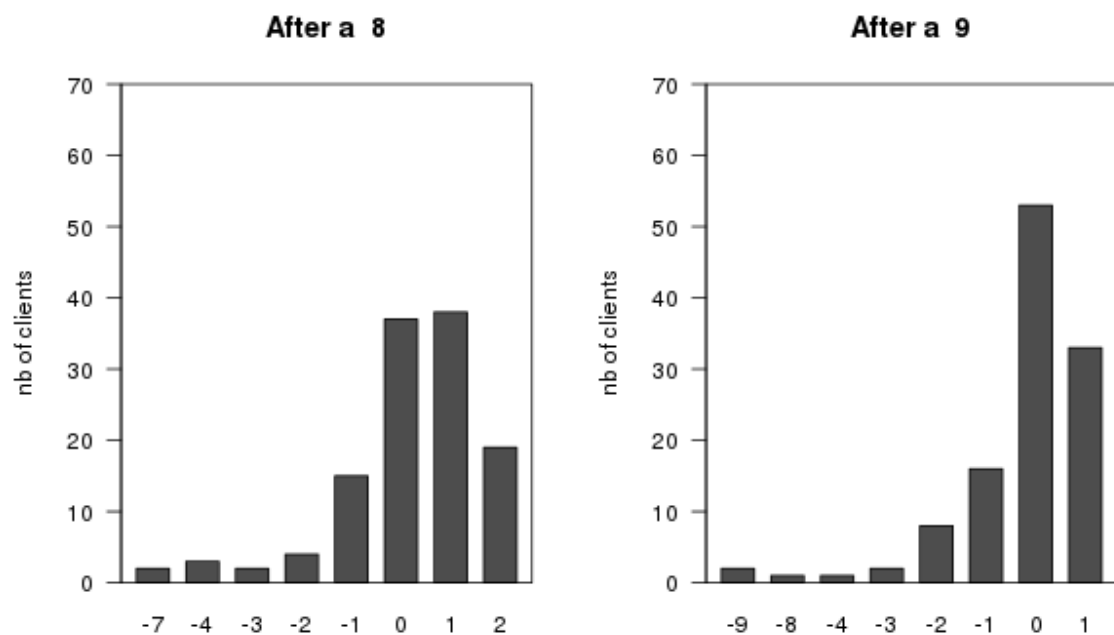


Figure 17: Evolution of the grade after a 8 and a 9

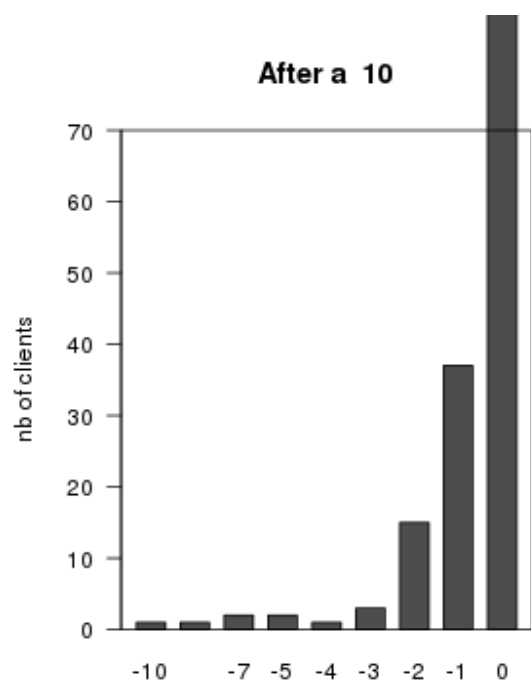


Figure 18: Evolution of the grade after a 10

B.2 According to Marche CSP

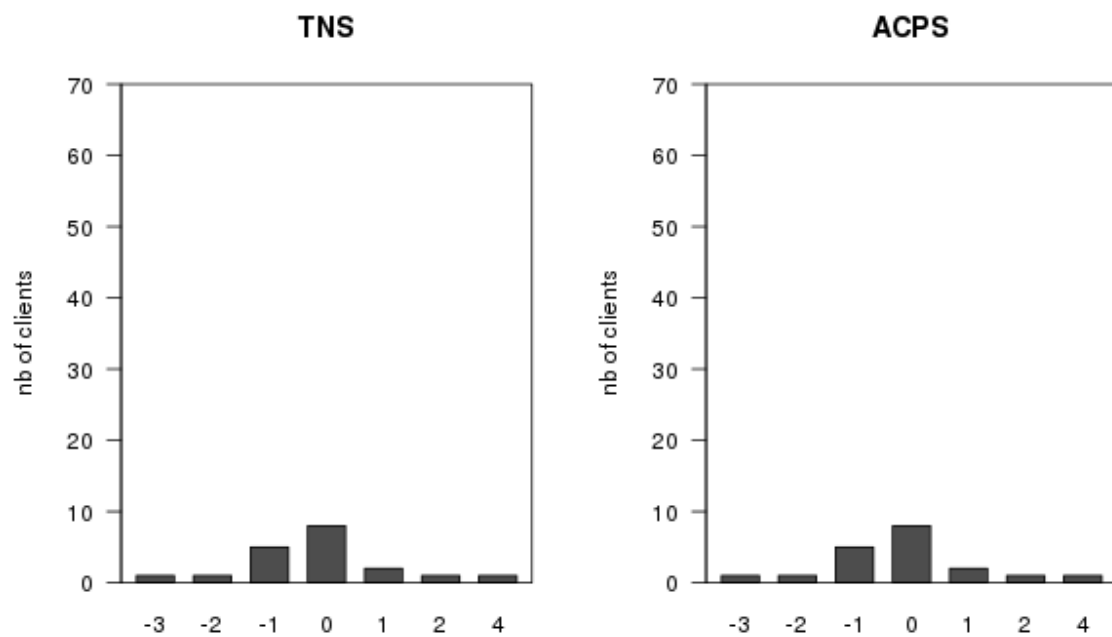


Figure 19: Evolution of the grade for 'marche CSP' TNS and ACPS

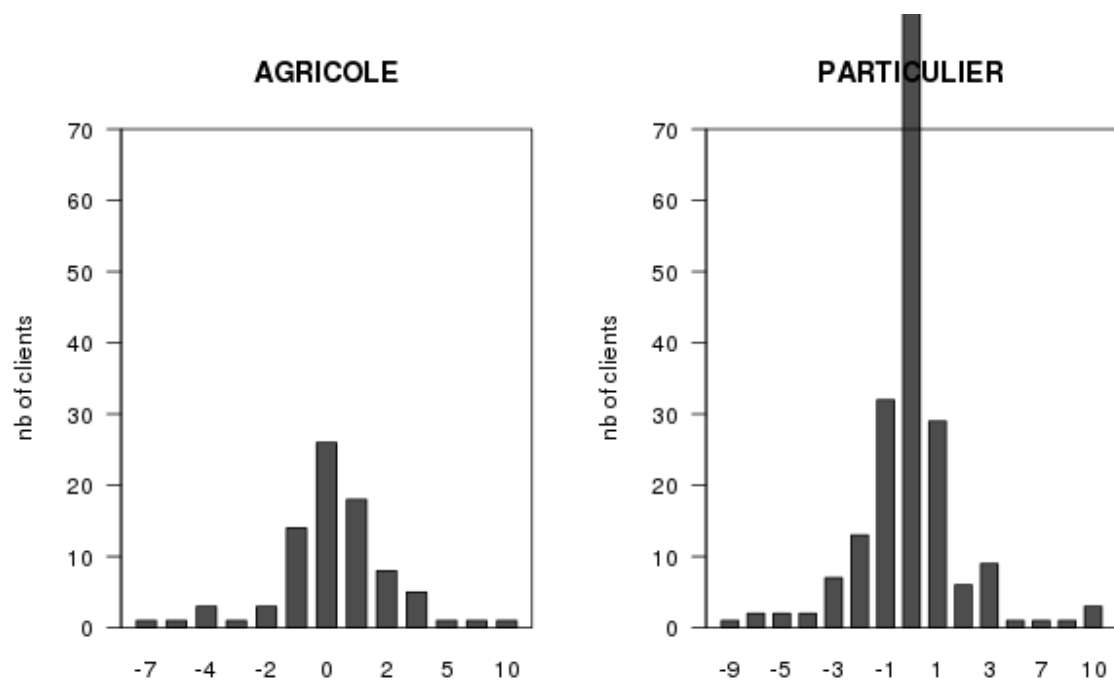


Figure 20: Evolution of the grade for 'marche CSP' AGRICOLE and PARTICULIER

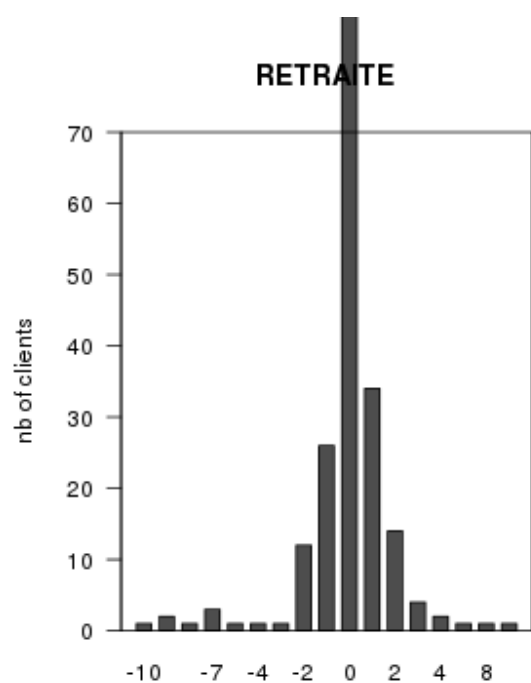


Figure 21: Evolution of the grade for 'marche CSP' RETRAITE

B.3 According to Nature

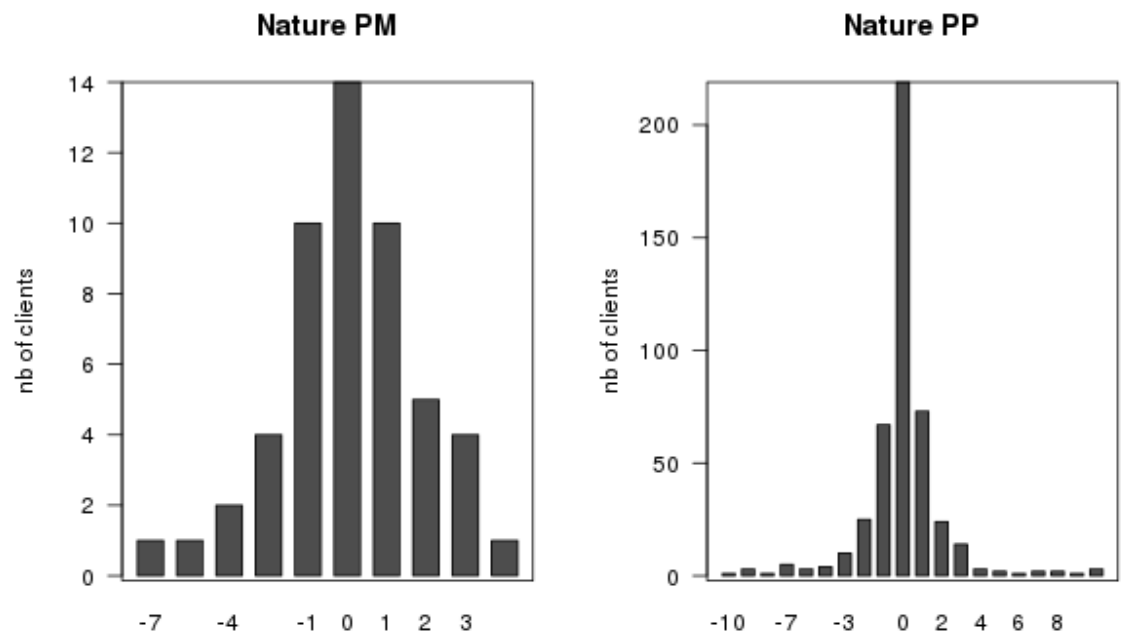


Figure 22: Evolution of the grade for 'nature' PM and PP

B.4 According to Segmentation Distributive

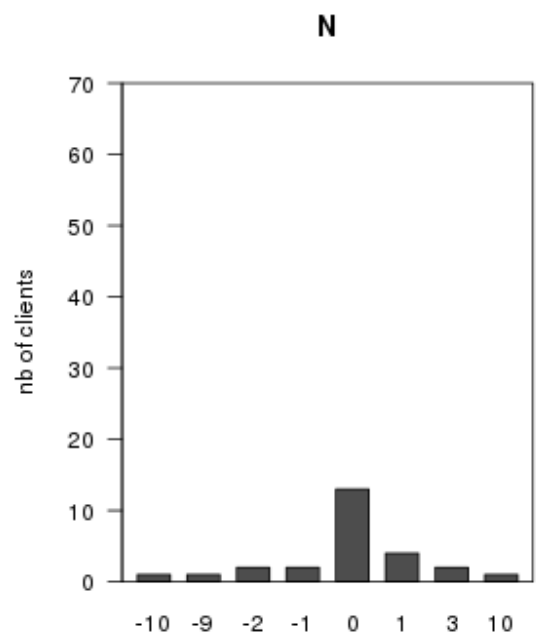


Figure 23: Evolution of the grade for 'segmentation' N

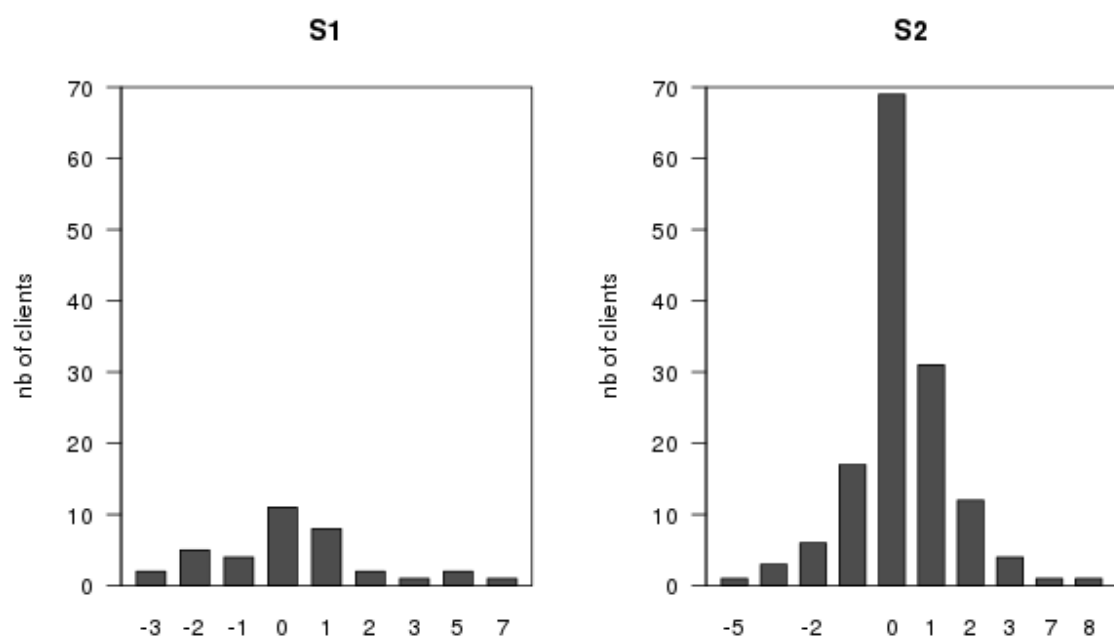


Figure 24: Evolution of the grade for 'segmentation' S1 and S2

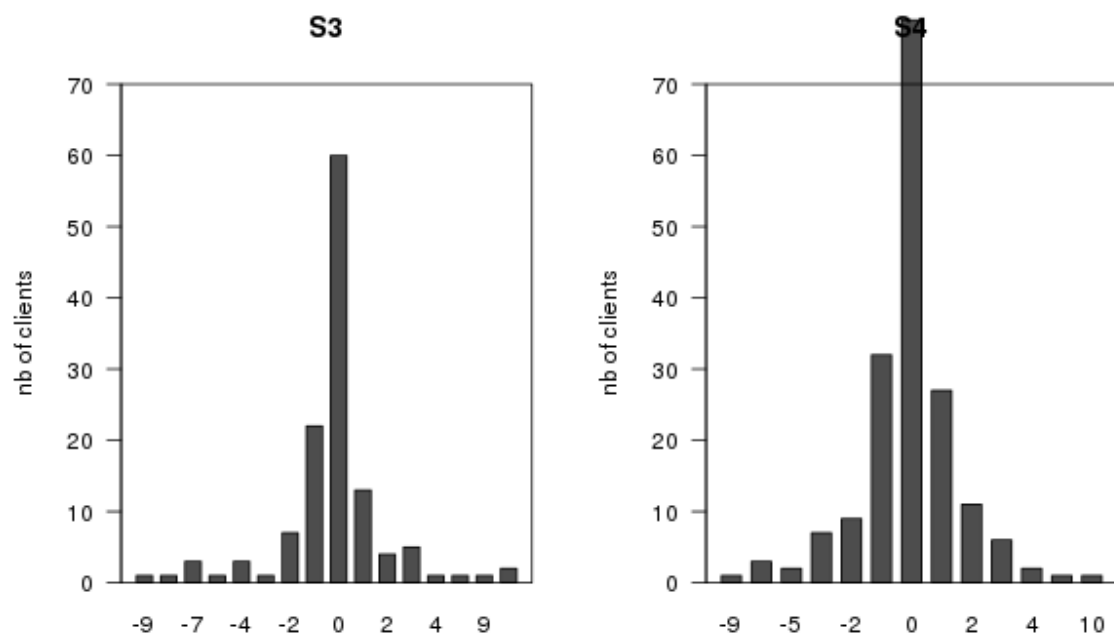


Figure 25: Evolution of the grade for 'segmentation' S3 and S4

B.5 According to Tranche d'age

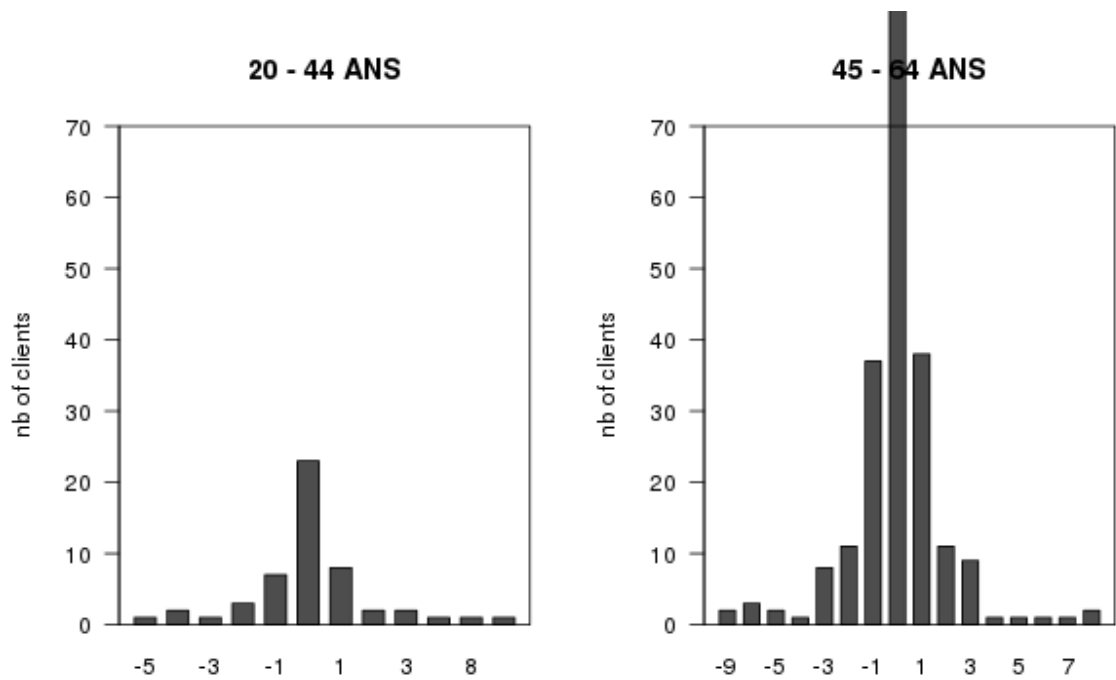


Figure 26: Evolution of the grade for 'tranche d'age': 20-44 and 45-64 ANS

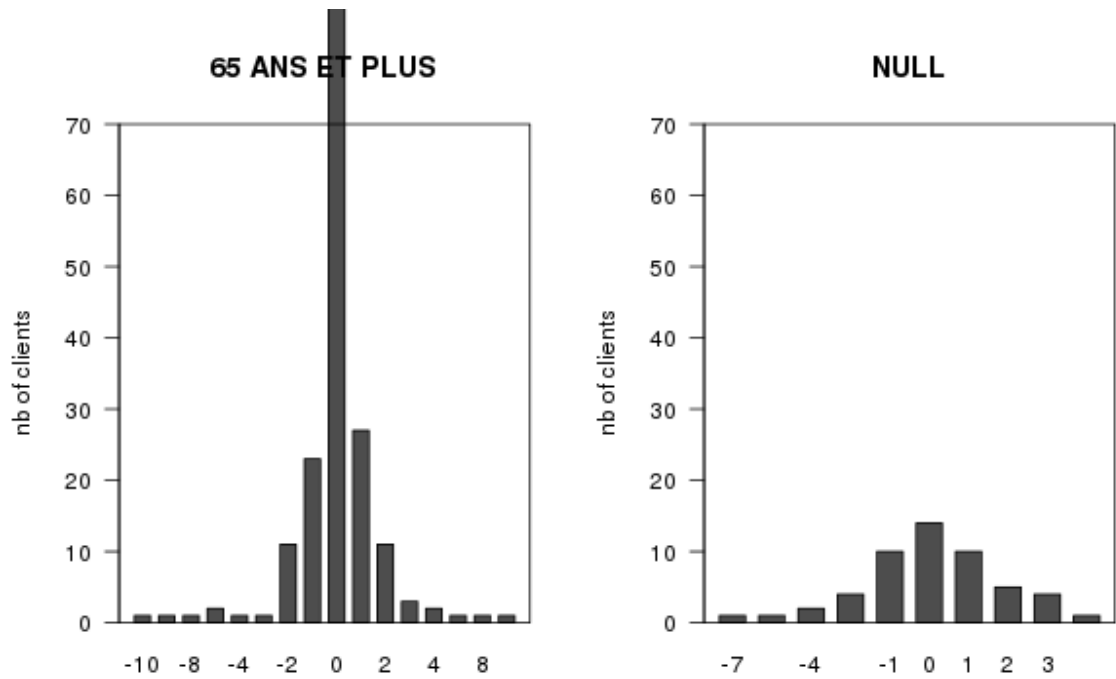


Figure 27: Evolution of the grade for 'tranche d'age': 65 AND ET PLUS and NULL

B.6 According to Typologie

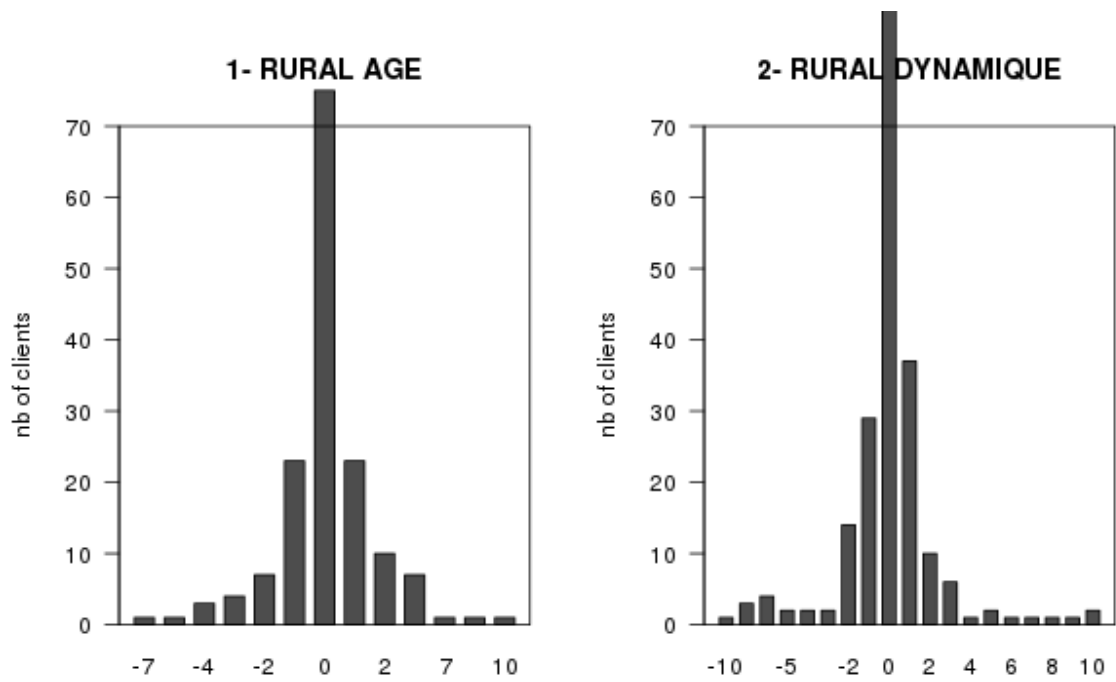


Figure 28: Evolution of the grade for typologie - RURAL AGE and RURAL DYNAMIQUE

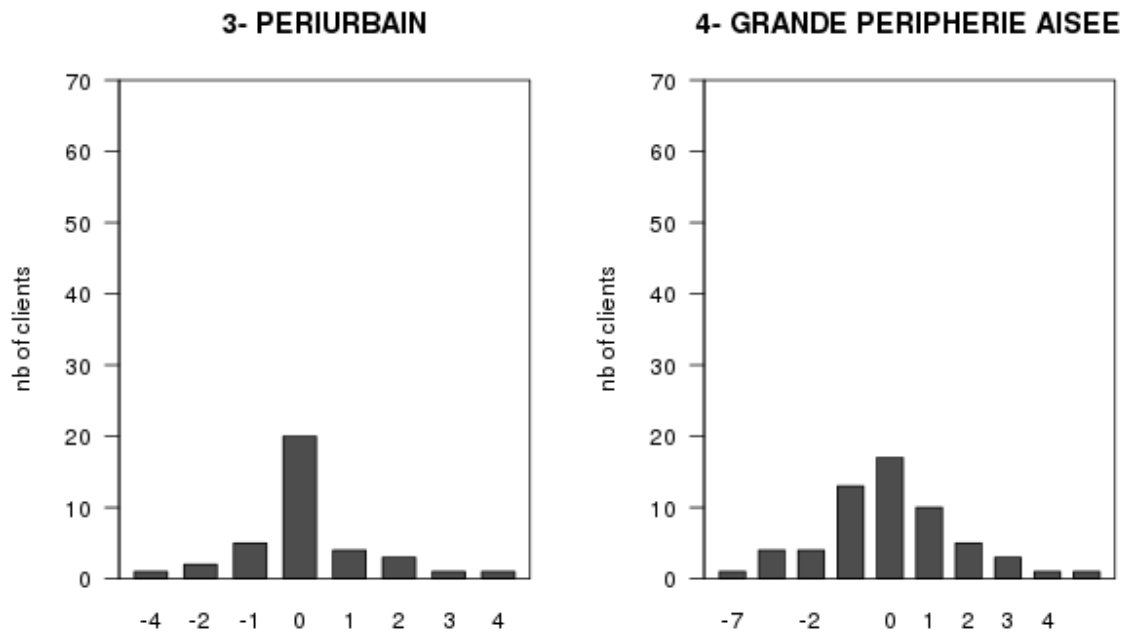


Figure 29: Evolution of the grade for typologie - PERIURBAIN and GRANDE PERIPHERIE AISEE

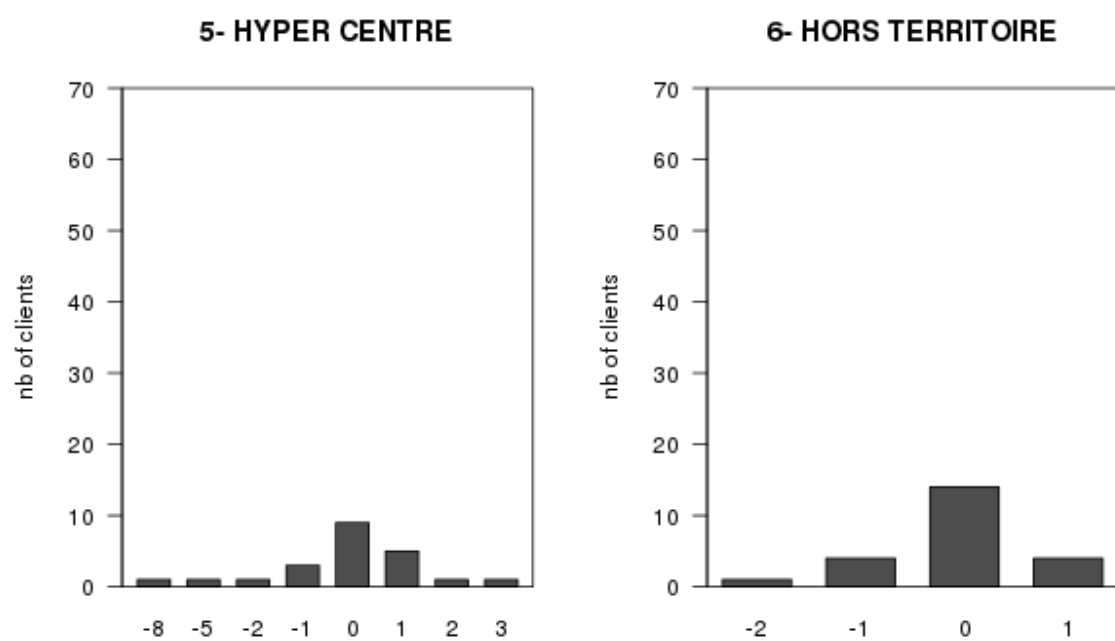


Figure 30: Evolution of the grade for typologie - HYPER CENTRE and HORS TERRITOIRE

C Reclamation & termination

Number of reclamation in fuction of the TYPE

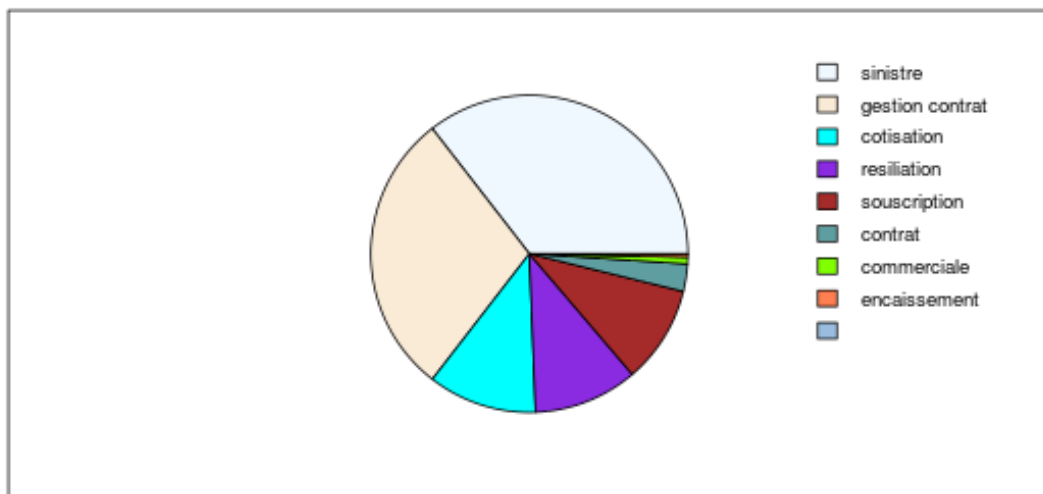


Figure 31: Number of 'reclamations' according to their TYPE

Number of reclamation in fuction of the client typologie

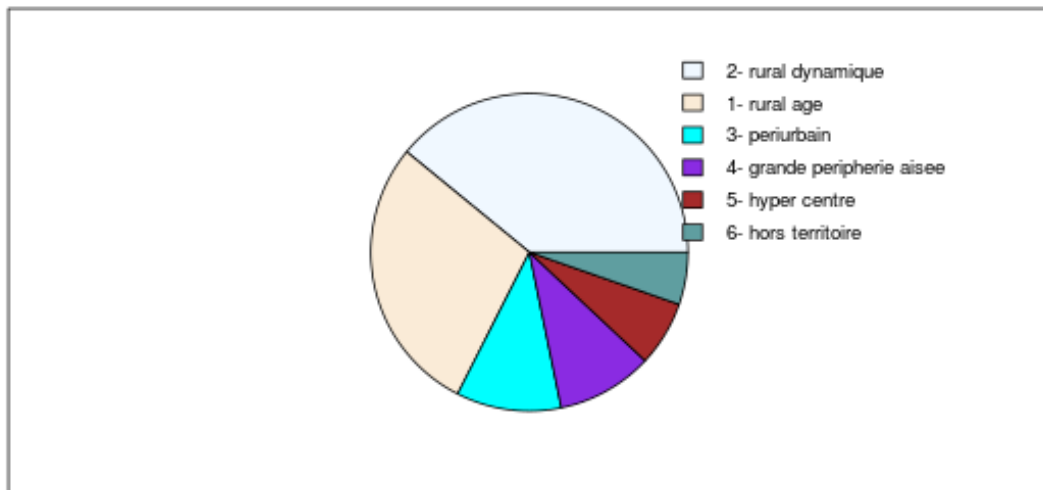


Figure 32: Number of 'reclamations' according to the client TYPOLOGIE

**Number of reclamation in fuction of the client typologie
in proportion of client of this categorie**

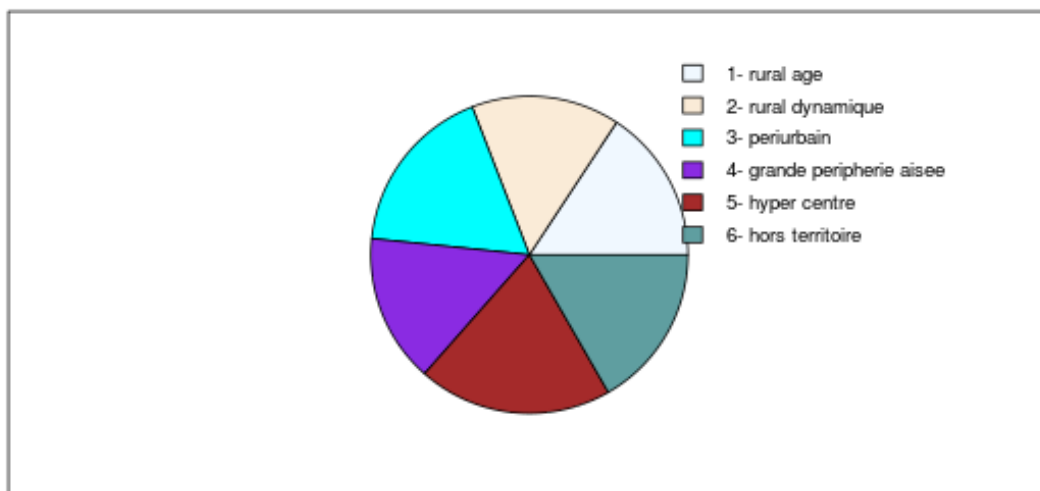


Figure 33: Number of ‘reclamations’ according to the client TYPOLOGIE in proportion of the client of this categorie

Number of reclamation in fuction of the client MARCHE_PSO

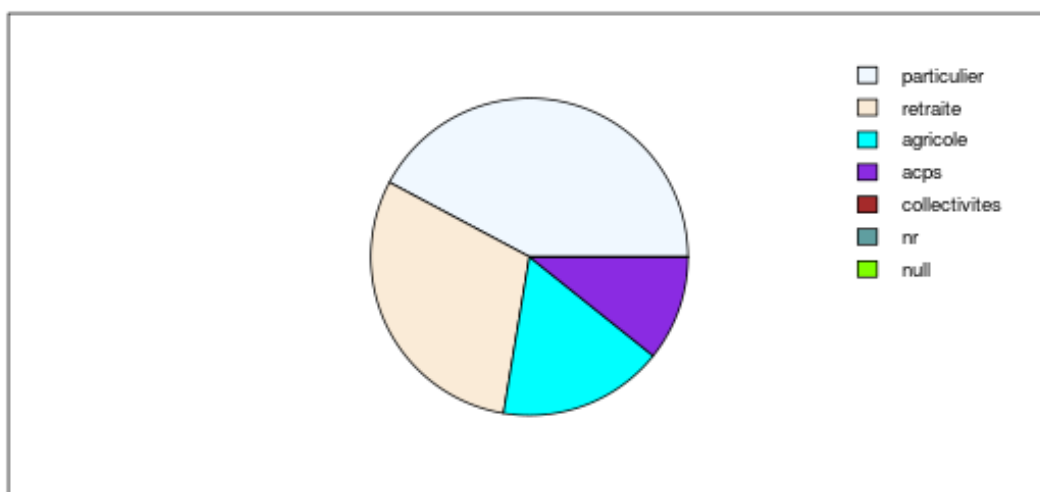


Figure 34: Number of ‘reclamations’ according to the client MARCHE_PSO

**Number of reclamation in fuction of the client MARCHE_PSO
in proportion of client of this categorie**

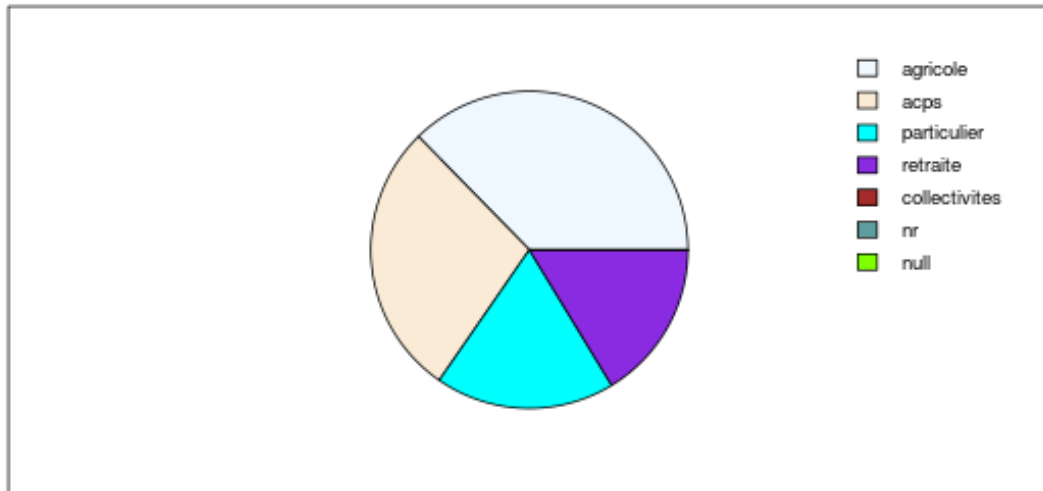


Figure 35: Number of ‘reclamations’ according to the client MARCHE_PSO in proportion of the client of this category

**Number of reclamation in fuction of the client MARCHE_PSO
in proportion of client of this categorie**

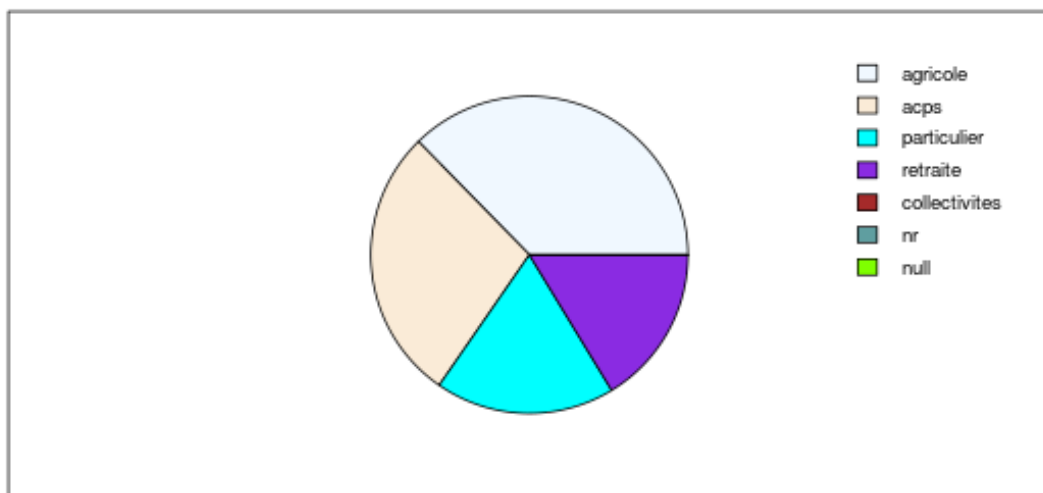


Figure 36: Number of ‘reclamations’ according to the client MARCHE_PSO in proportion of the client of this category

Number of "agricole" reclamation in fuction of the TYPE

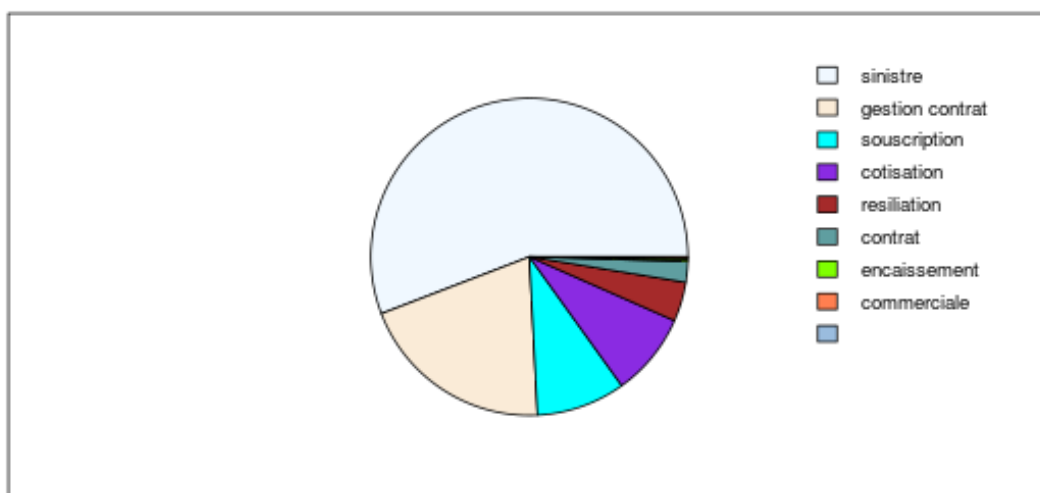


Figure 37: Number of "agricole" reclamation according to the TYPE of reclamation

Number of reclamation in fuction of the client 'departement'

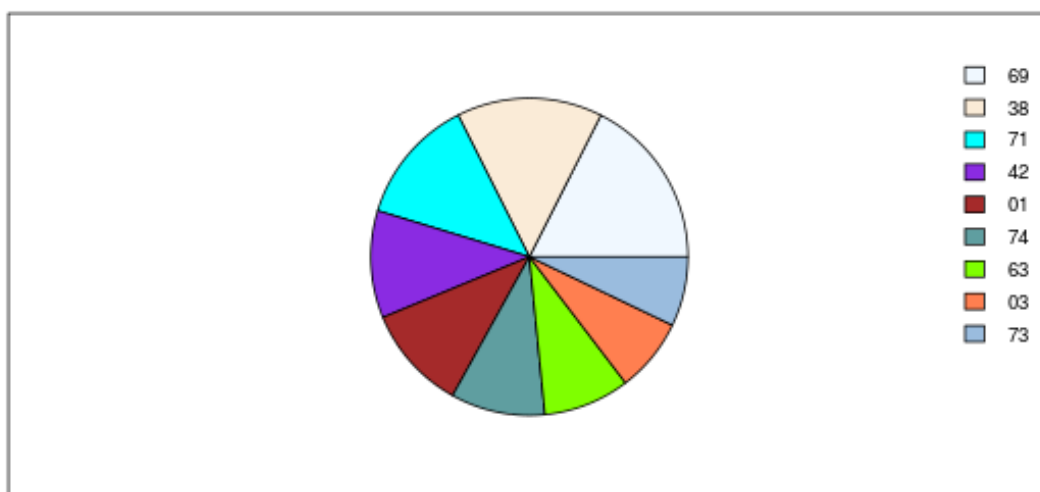


Figure 38: Number of 'reclamations' according to the client department

Number of reclamation in fuction of the client 'departement'

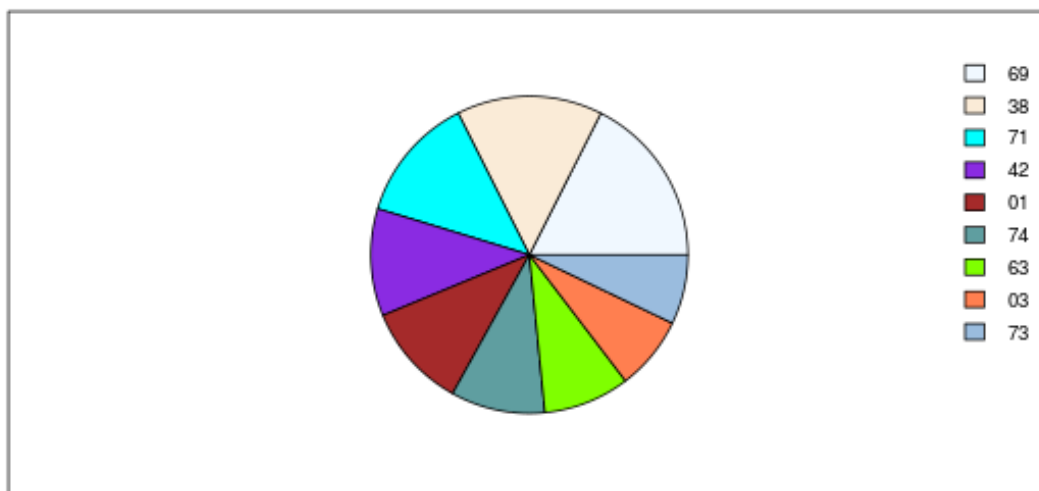


Figure 39: Number of 'reclamations' according to the client department

**Number of reclamation in fuction of the client 'departement'
in proportion of client of this categorie**

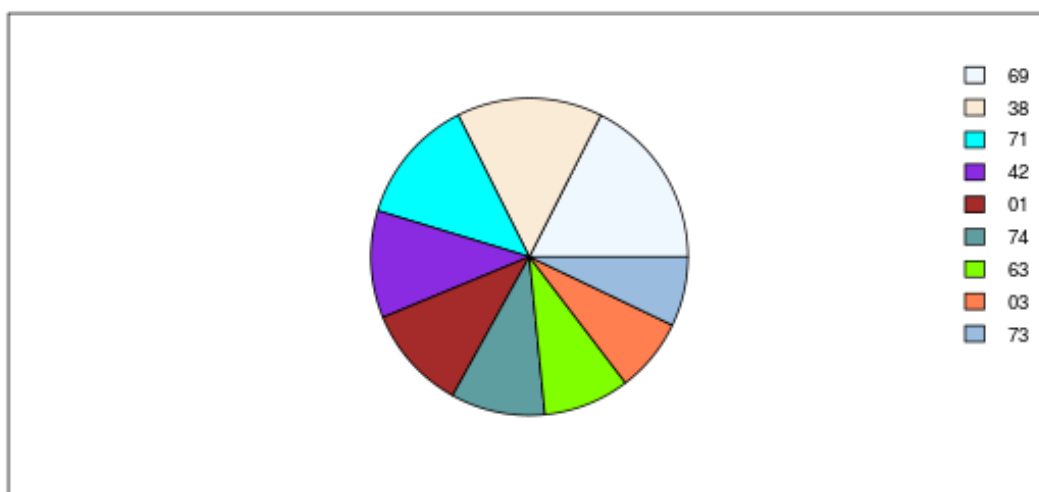


Figure 40: Number of 'reclamations' according to the client department in proportion of the client of this category

Number of resiliation in fuction of the client typologie

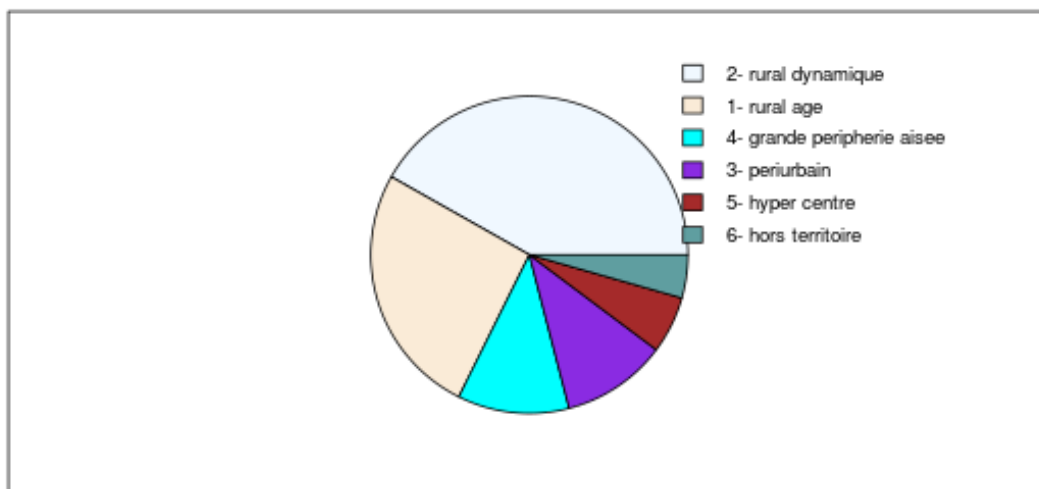


Figure 41: Number of 'resiliation' according to the client TYPOLOGIE

Number of resiliation in fuction of the client typologie in proportion of client of this categorie

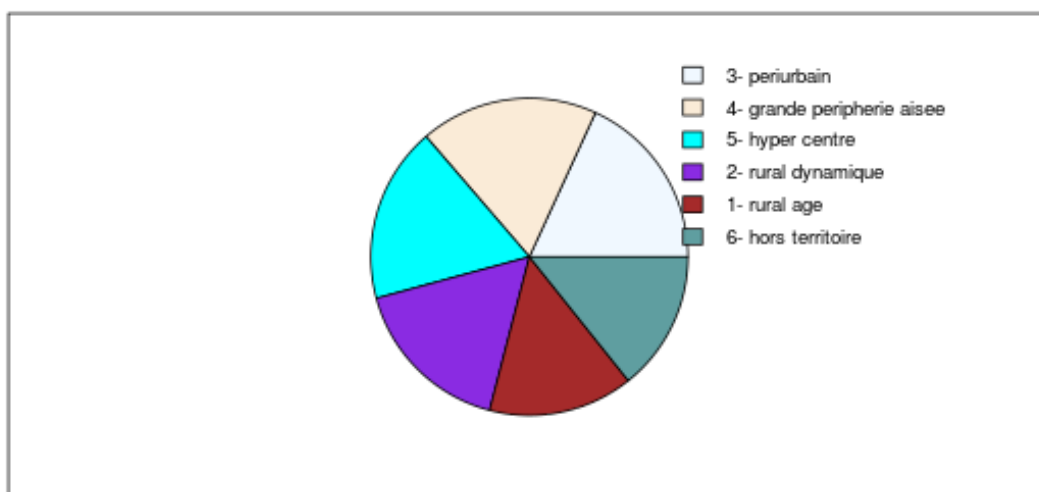


Figure 42: Number of 'resiliation' according to the client TYPOLOGIE in proportion of the client of this category

Number of resiliation in fuction of the client MARCHE_PSO

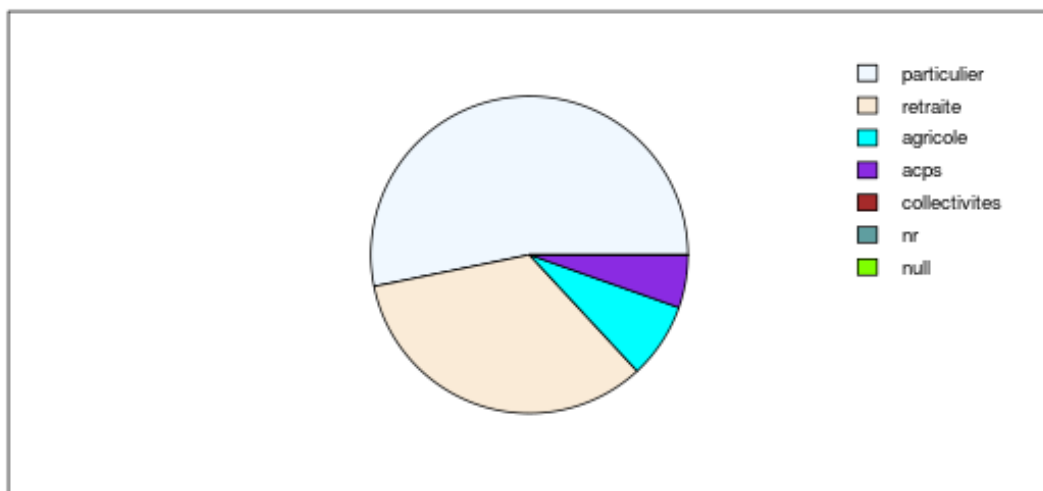


Figure 43: Number of 'resiliation' according to the client MARCHE_PSO

**Number of resiliation in fuction of the client MARCHE_PSO
in proportion of client of this categorie**

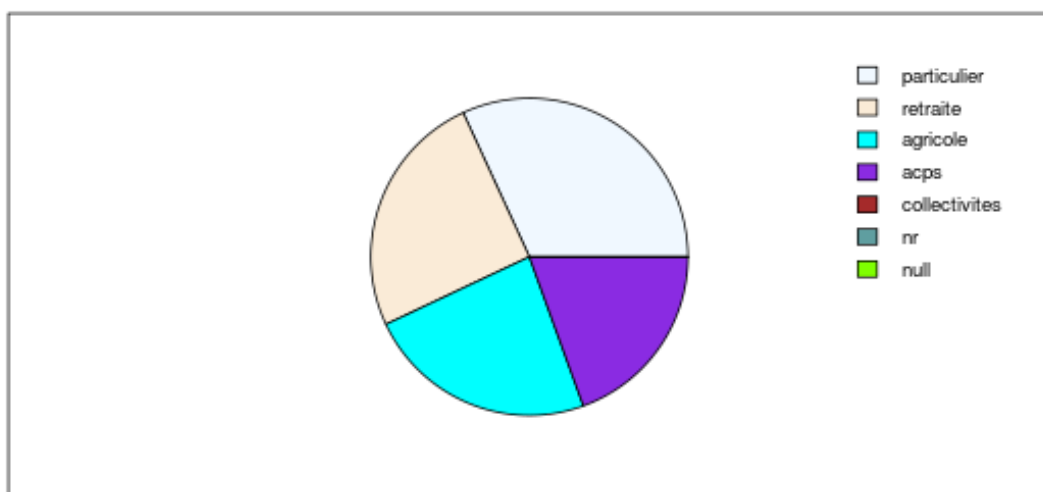


Figure 44: Number of 'resiliation' according to the client MARCHE_PSO in proportion of the client of this category

Number of resiliation in fuction of the client 'departement'

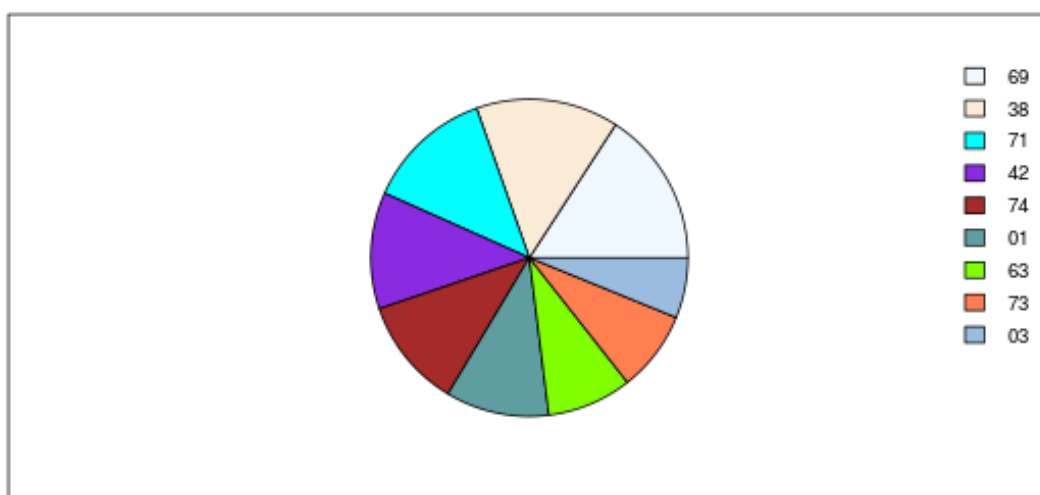


Figure 45: Number of 'resiliation' according to the client department

D Term frequency on comments

	ngrams	nappear		ngrams	nappear
1	rapid	8564	1	alor	720
2	accueil	5712	2	mois	692
3	efficac	3426	3	non	585
4	écout	2650	4	apres	569
5	satisf	2465	5	pai	563
6	tre	2136	6	expert	494
7	satisfait	2056	7	plusieur	483
8	réactiv	1643	8	envoi	456
9	compétent	1559	9	quand	432
10	expliqu	1470	10	moin	403
11	renseign	1348	11	tous	399
12	clair	1290	12	comm	391
13	qualit	1276	13	dit	376
14	agréabl	1242	14	cotis	369
15	attent	1140	15	beaucoup	358
16	question	1091	16	mutuel	358
17	personnel	1074	17	euros	355
18	not	968	18	deux	338
19	parf	832	19	nouvel	336
20	prestat	786	20	courri	332

Table 1: 1-grams frequencies for satisfied and unsatisfied

	ngrams	nappear		ngrams	nappear
1	bon accueil	2203	1	trop long	143
2	rapid efficac	1126	2	trop cher	130
3	bon conseil	826	3	contrat assur	125
4	tres rapid	761	4	moin cher	113
5	bien reçu	706	5	attend toujours	112
6	tres satisf	626	6	gest commercial	111
7	répons rapid	569	7	assur voitur	110
8	tres satisfait	535	8	cel fait	108
9	tre bon	526	9	plus cher	107
10	bon contact	484	10	résili contrat	107
11	bien conseil	445	11	nouveau contrat	101
12	tres agréabl	443	12	plusieur fois	100
13	tres professionnel	431	13	assur auto	97
14	bien renseign	403	14	tous contrat	93
15	bon servic	380	15	depuis plus	92
16	trait rapid	373	16	tres déçu	91
17	servic rapid	361	17	contrat chez	90
18	satisf servic	358	18	assur habit	88
19	efficac rapid	351	19	beaucoup trop	87
20	accueil bon	343	20	suit sinistr	85

Table 2: 2-grams frequencies for satisfied and unsatisfied

	ngrams	nappear		ngrams	nappear
1	tres bon accueil	961	1	client depuis an	29
2	bon accueil bon	300	2	del trop long	28
3	pris charg rapid	265	3	tous contrat chez	27
4	bon pris charg	240	4	contrat chez groupam	26
5	tres bon contact	213	5	tout assur chez	23
6	tres bon conseil	189	6	assur tous risqu	22
7	tre bon accueil	152	7	sinistr non respons	22
8	tout bien pass	139	8	toujour rien reçu	22
9	tres bien accueil	136	9	beaucoup trop long	21
10	bon accueil téléphon	134	10	appel plusieurs fois	20
11	tres bien conseil	129	11	jour plus tard	20
12	tres bien renseign	129	12	non pris compt	20
13	tres bon servic	126	13	aller voir ailleur	19
14	rapid pris charg	124	14	assur tout risqu	19
15	bon accueil agenc	120	15	chang par bris	19
16	accueil bon conseil	108	16	cel fait plus	18
17	rapid trait dossi	104	17	résili tous contrat	18
18	pris compt demand	100	18	mois plus tard	17
19	bon accueil tres	96	19	an assur chez	16
20	bon accueil expliqu	92	20	aucun gest commercial	16

Table 3: 3-grams frequencies for satisfied and unsatisfied

	ngrams	nappear		ngrams	nappear
1	bon accueil bon conseil	97	1	tous contrat chez groupam	11
2	tres bon accueil tres	78	2	chez depuis plus an	8
3	tres bon accueil bon	77	3	bonjour mis not car	7
4	tres bon accueil téléphon	74	4	ni plus ni moin	7
5	bon accueil tres bon	62	5	plus an assur chez	7
6	tres bon accueil agenc	60	6	quelqu jour plus tard	7
7	bon rapport qualit prix	56	7	assur depuis plus an	6
8	tres bon pris charg	48	8	client chez groupam depuis	6
9	bon accueil expliqu clair	46	9	suit sinistr non respons	6
10	tres bon accueil expliqu	44	10	toujour reçu cart vert	6
11	pris charg rapid efficac	37	11	trop cher rapport concurrent	6
12	pris charg tres rapid	35	12	cel fait plus mois	5
13	bon accueil bon expliqu	33	13	chez groupam depuis plus	5
14	tres bien reçu agenc	33	14	client groupam depuis an	5
15	bon accueil bon renseign	31	15	contact fair point situat	5
16	tres bon accueil répons	31	16	depuis plus an chez	5
17	bon accueil bon écout	29	17	jour toujours rien reçu	5
18	peut toujours fair mieux	28	18	rembours beaucoup trop long	5
19	tres bon accueil conseil	28	19	résili tous contrat chez	5
20	tres bien reçu conseil	26	20	agenc trop souvent ferm	4

Table 4: 4-grams frequencies for satisfied and unsatisfied

	ngrams	nappear		ngrams	nappear
1	tres bon accueil tres bon	54	1	client chez depuis plus an	3
2	tres bon accueil bon conseil	25	2	client chez groupam depuis plus	3
3	bon accueil tres bon conseil	21	3	goélet don silvano mor bihan	3
4	tres bon accueil expliqu clair	21	4	jour plus tard toujours rien	3
5	tres bien reçu tres bien	13	5	oblig aller chez partenair groupam	3
6	tres bon accueil bon expliqu	11	6	accueil cap larg abordag gourmet	2
7	tres bon accueil répons rapid	11	7	alor client depuis plus an	2
8	bon accueil expliqu tres clair	9	8	appel plusieurs fois avoir bon	2
9	mis not car tres bien	9	9	apres avoir vain tent joindr	2
10	tres bien accueil tres bien	9	10	apres plus an chez groupam	2
11	tres bon accueil bon écout	8	11	assur auto beaucoup trop cher	2
12	tres bon accueil renseign clair	8	12	assur auto trop cher rapport	2
13	tres bon accueil tres bien	8	13	assur chez depuis nombreux anné	2
14	bien accueil tres bien renseign	7	14	assur chez groupam depuis an	2
15	bien reçu tres bien conseil	7	15	assur depuis plus an chez	2
16	demand pris compt tres rapid	7	16	aupres conseil groupam plus proch	2
17	tres bon accueil téléphon rapid	7	17	auto trop cher rapport concurrent	2
18	tres bon accueil tres professionnel	7	18	cap larg abordag gourmet entrepris	2
19	tres bon rapport qualit prix	7	19	capabl fair gest commercial prendr	2
20	bon accueil tres bon écout	6	20	cel fait plus an chez	2

Table 5: 5-grams frequencies for satisfied and unsatisfied