Department Analysis

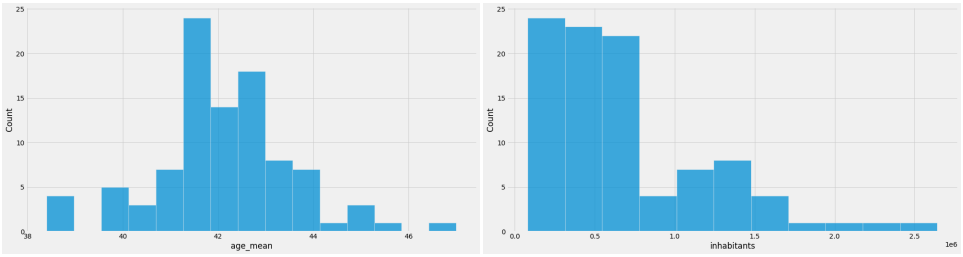
Sacha Cohen & Loup Theuret

Abstract. This report aims at providing an exploratory analysis on French departments and more specifically, to study their potential heterogeneity. The analysis will rely on descriptive analysis and clustering algorithms.

1 Part 1 : Exploratory analysis

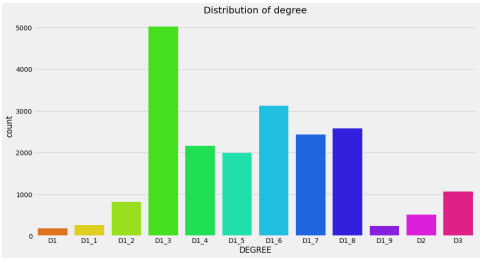
1.1 Data set 1 : descriptive statistics

The clustering analysis is based on two data sets : the first one characterises the departments by their inhabitants while the second one describes the departments using the people who work in it.
The first data set is composed of age, degree, household, inhabitants and gender. The following figures give the distribution of these variables in the data set. The distribution of Inhabitants is right-skewed. There is a concentration of departments with population less than 1 million inhabitants. As far as age is concerned, one observes that the majority of people in the data set is in the 40's.

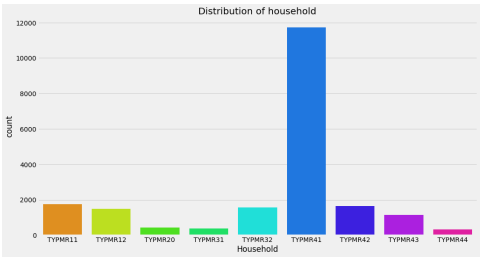


(a) Age distribution in Dataset (b) Inhabitants distribution in Dataset

There is a high proportion of TYPMR41 type of household which are families composed of a working couple, compared to the other categories that are less represented (see annex for all labels). There are many people with a $D1_3$ degree (CAP/BEP) and few people with a D1, $D1_1$, $D1_2$, $D1_9$, D2 and D3 degree (cf annex). Gender is well balanced in this data set.

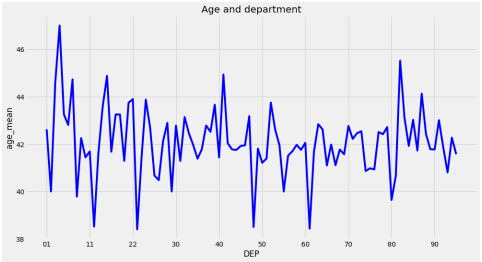


(a) Degree distribution in the data set

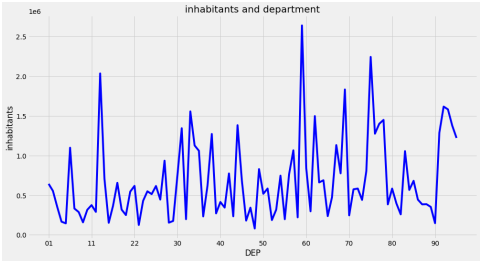


(b) Household distribution in the data set

Now, let us confront the characteristics with departments. One expects that huge variations across departments of one or several categories may indicate a cluster. For instance, some departments with mean age above 44 could constitute a cluster and the ones with mean age below 38 another cluster. This important volatility is also present for inhabitants but not for all diplomas. The standard deviation could be an interesting proxy to give an insight into a characteristic as a cluster.



(a) Mean age and departments



(b) Number of inhabitants in departments

Dispersion seems pretty high for the number of inhabitants (std : 514881.82). In comparison, the one of age seems less noteworthy (std : 1.45 if mean age is used). Regarding the four most frequent degrees, $D1_3$, $D1_6$, $D1_7$ and $D1_8$, the dispersion is quite low (std :0.09, 0.08, 0.06 and 0.07 respectively), with some outlier departments for each degree.

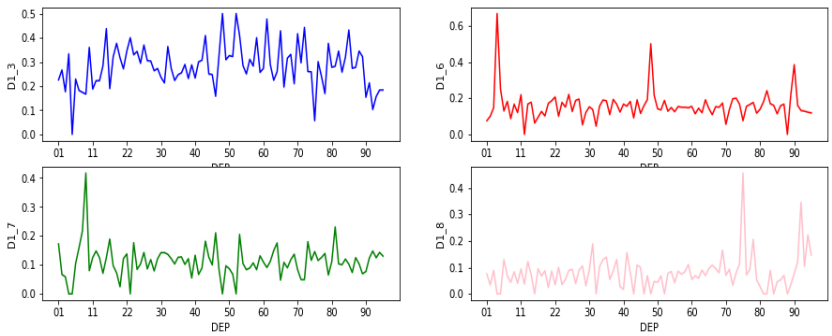
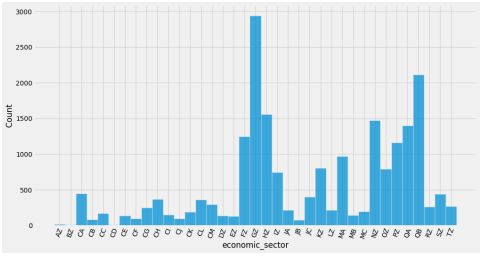


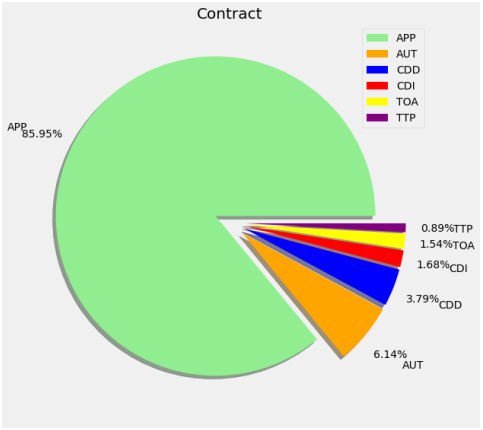
Figure 4: Degree and departments

1.2 Data set 2 : descriptive statistics

The same analysis is conducted for the second data set. The latter is composed of the type of contract, economic sector, job conditions, employer type, employee count, job category, working hours and emolument. One sees that almost all economic sectors are represented in the data set even though GZ (Commerce : répararation automobile et moto) and QB (hébergement médico-social...)are the most present in the data set. As for type of contract, there is a almost 90% of APP contracts (contrats d'apprentissage).



(a) Economic sector



(b) Type of contract

ct_9 (Autres sociétés privées) is the most frequent employer type with 13559 occurrences in the data set against 3259 for ct_8 and 1265 for ct_7 (the other being less than 1000). Finally, there are 348 people that are employed in a corporation with 0 employee. But one can identify them as outliers since the number of people in all the other categories of "employee count" is comprised between 2220 (20-50 jobs) and 4643 (1-10 jobs).

```
C    15896
P     3898
D      241
F      206
K       95
Y       65
Name: Job_condition, dtype: int64
O    19573
X      505
A      323
Name: JOB_CATEGORY, dtype: int64
```

Figure 6: Job condition and job category

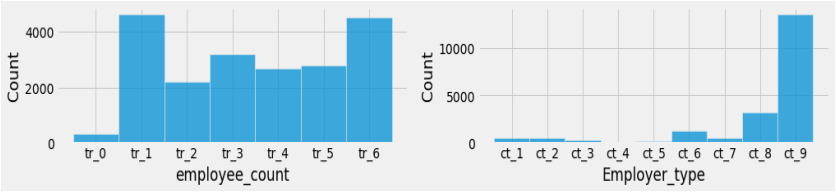
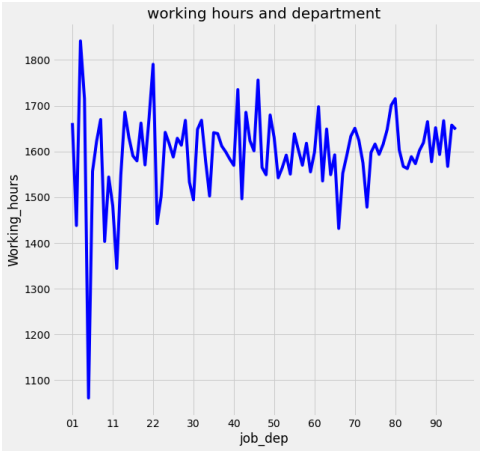
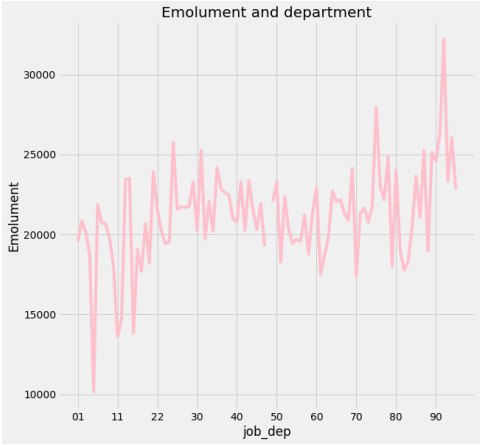


Figure 7: Employee count and employer type

Working hours and emolument dispersions are noteworthy (standard deviations of 95.05 and 3006.11 respectively). Their volatility is confirmed when looking at the plot between working hours or emolument and department. For instance, working hours median is at 1600.15 but some departments are way above 1700 and others way below 1500. This could constitute clusters. Emolument's mean is 21187.64, the first quarter is 19560.97 and the third is 22898.54. The variations observed on the graph, even though not that extreme, could signal clusters too.



(a) Working data



(b) Emolument

1.3 Clustering algorithm

In order to give an overview of potential clusters of the data set, one can use hierarchical clustering n.d. The dendrogram, which is a tree representing the order and distances of merges in the hierarchical clustering, is useful to get the number of clusters. The dendrogram can be completed by a Hierarchically-clustered Heatmap. Changes in color expresses data clustering for rows and columns. It is possible to add dendrograms to present the clustering on the map.
On the first data set dendrogram, two clusters can be seen. This is less clear on the heatmap.

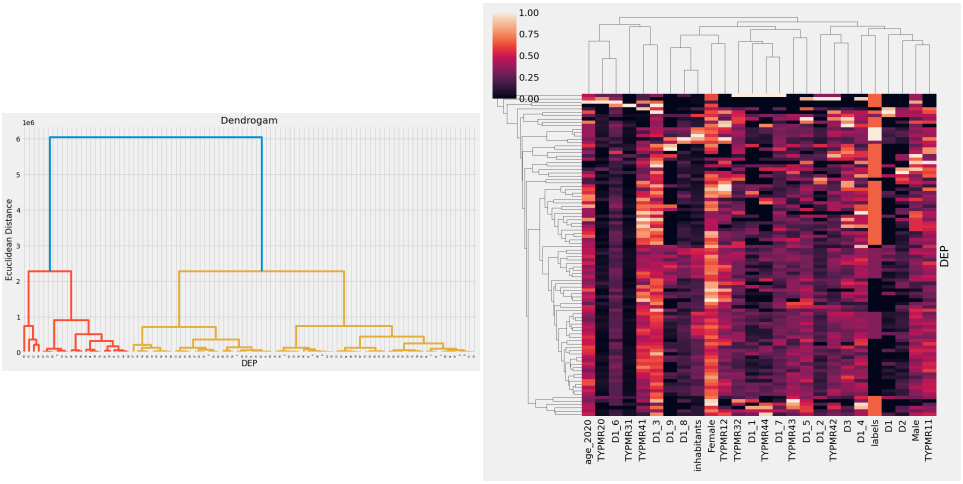


Figure 9: data set 1

The second data set dendrogram identifies three clusters. Albeit, the heatmap does not give a clear number of clusters in the data set, it allows us to compare with the first data set. Demographic information on departments seems to explain more heterogeneity than economic characteristics.

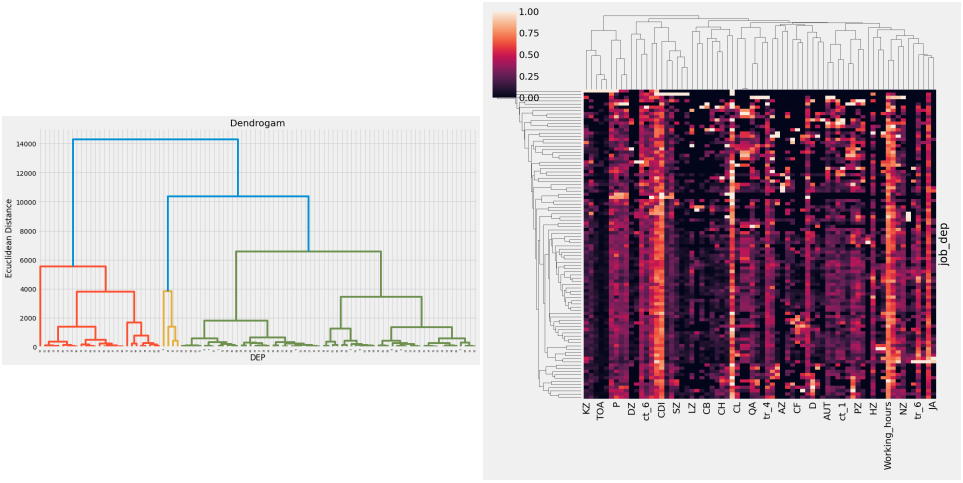


Figure 10: data set 2

To deepen the analysis, particular characteristics of the data set will be analysed as sources of potential clusters. The main method to conduct this study will be K-means clustering. K-means clustering is a method that aims at partitioning the observations of the data set into clusters. Each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid). A cluster refers to a collection of data points grouped together because of some similarities.

K-means algorithm starts by finding k number of centroids, and then allocates every data point to the closest cluster, while keeping the centroids as small as possible. To find the optimal number of clusters, one can use the Elbow method.

1.3.1 Data set 1

First, it would be interesting to verify whether it is possible to cluster departments according to gender and age. The Elbow method gives the optimal number of cluster of 4, which can also be seen on the dendrogram.
Gender is not a good way to divide departments into categories. Is age a better way to cluster? Probably not, since there is a huge concentration around 43 years old. Hence, this clustering is not very satisfying and departments seems to be homogeneous in terms of age and gender.

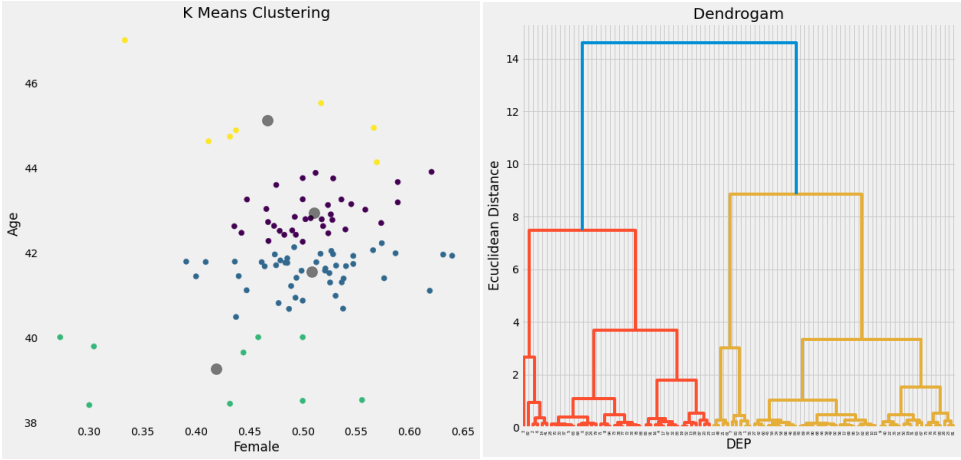


Figure 11: Cluster by age and gender

Second, one should verify if departments can be clustered by inhabitants. Gender is still not convincing but clustering by inhabitants seems feasible. The Elbow method gives an optimal number of cluster of 3 whereas it seems to be 4 on the dendrogram. This is not surprising since on the elbow graph there is a kink at 4. The three clusters are departments between 0 and 500 000 inhabitants, departments between 500 000 and 1 200 000 inhabitants and departments above 1 200 000 inhabitants. Hence, departments are not homogeneous in population.

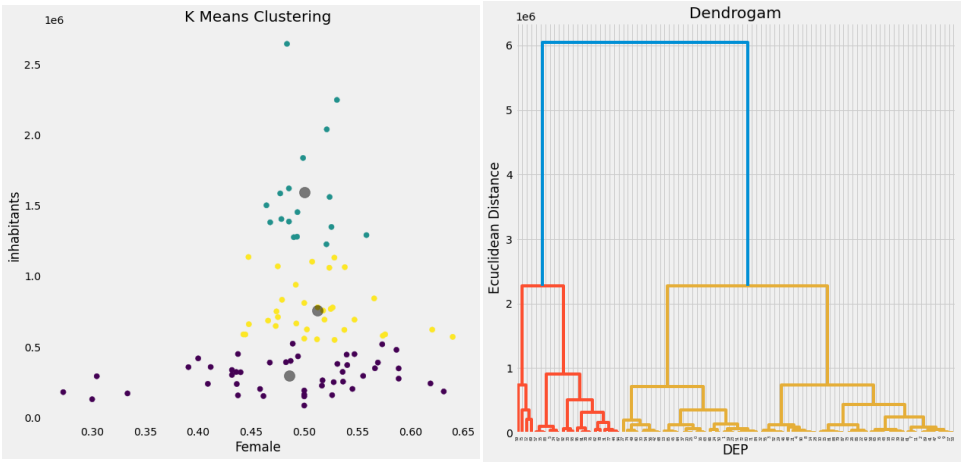


Figure 12: Cluster by inhabitants and gender

The last cluster analysis for this data set would be on household type and degree. Considering only degree, there are not clear clusters since there is a huge concentration of share of high degree diploma ($D1_8$) between 0 and 0.1. In addition, the graph on the right indicates that CAP and BEP do not tend to provide a better cluster. One can see that across departments with different levels of population, the share of $D1_3$ remains around 0.3. Thus, there is not a clear heterogeneity in term of degree.

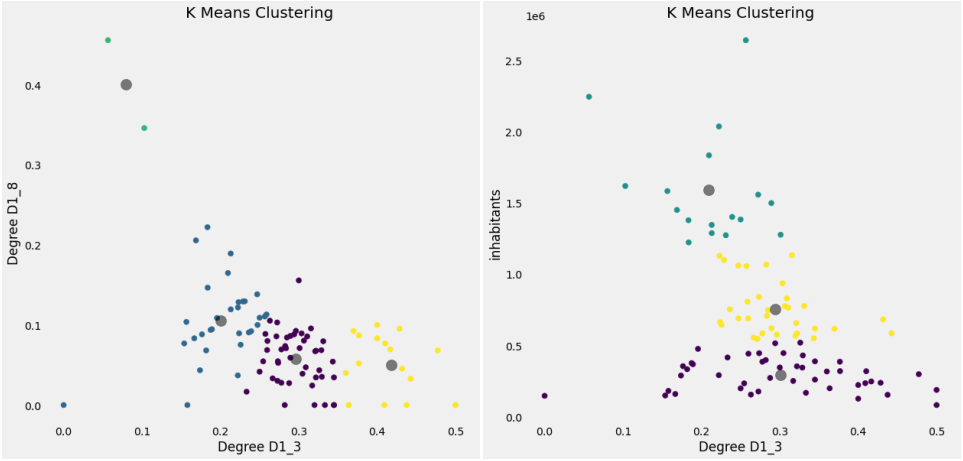


Figure 13: Cluster by degree

The conclusion is different for households. Considering the type of household that is the most present in the data set, one can perceive several clusters. The Elbow method give 3 optimal clusters, whereas the dendrogram gives 4.

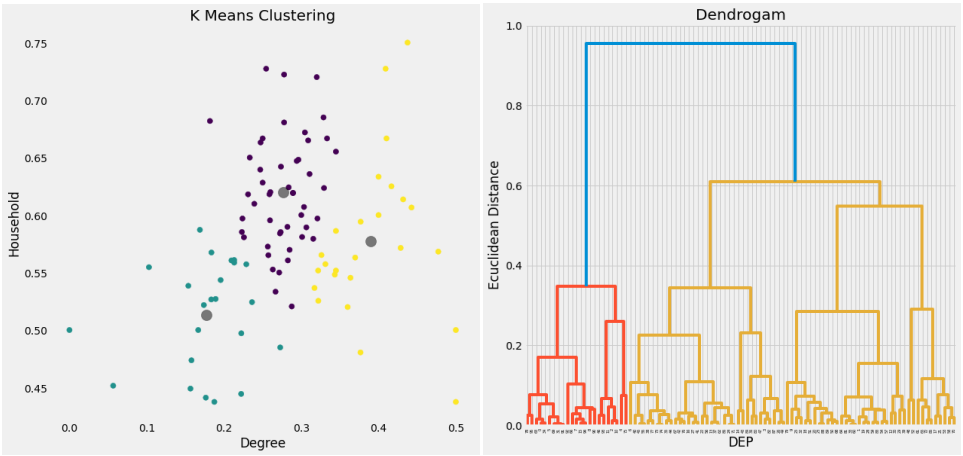


Figure 14: Cluster by households and degree

Partial conclusion : Gender, age and diploma do not appear to explain heterogeneity between french departments. However, the type of household and inhabitants are relatively good clustering characteristics.

1.3.2 Data set 2

In this data set, departments could be clustered by hours worked, emolument, the type of job, the economic activity, the type of contract... We focus on characteristics that are the most represented in the data set because the other ones might be not at all represented in too many departments.
First, focusing on the two economic sectors that are highly represented in the data set, the Elbow method identifies three clusters. The dendrogram clearly defines the clusters. For instance, two groups that can be easily seen as clusters on the graph are departments with relative high share of QA (0.10-0.15) and low share of GZ (0.05-0.15), and departments for which it is the contrary.

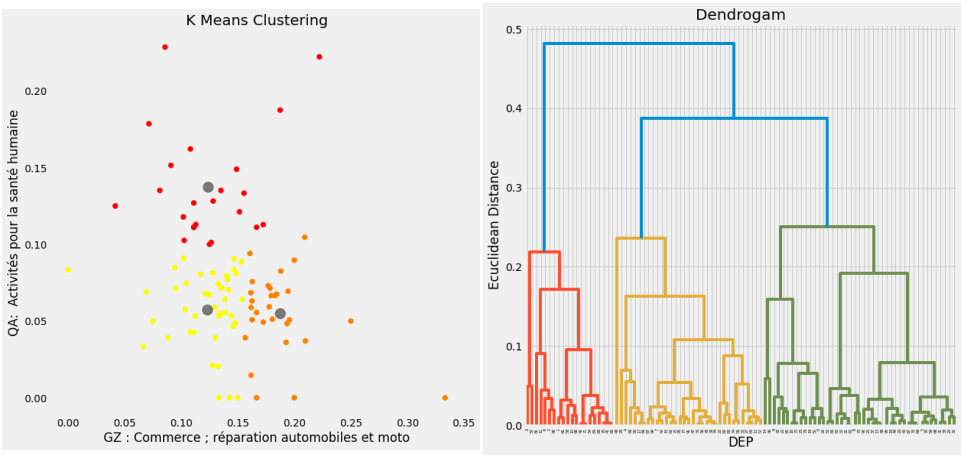


Figure 15: Cluster by economic sector

Second, clustering can be done for Employer type and for employee count. Recall that the subcategories chosen here for Employer type and employee count are the most represented ones in the data set. The dendrogram presents two clusters whereas the K-means algorithm displays four. Clusters are identified but not sharply.
Working hours and Emolument produce similar results.

Figure 16: Cluster by employee count and by Employer type

Last, the analysis is conducted on the type of contract (here CDD) along with another Employer type. Four clusters are computed using the Elbow method. The dendrogram gives three. Further analysis gives that the clustering is actually driven by CDD (similar cluster for different tests). Hence, among the different types of contract, CDD explains heterogeneity across departments.

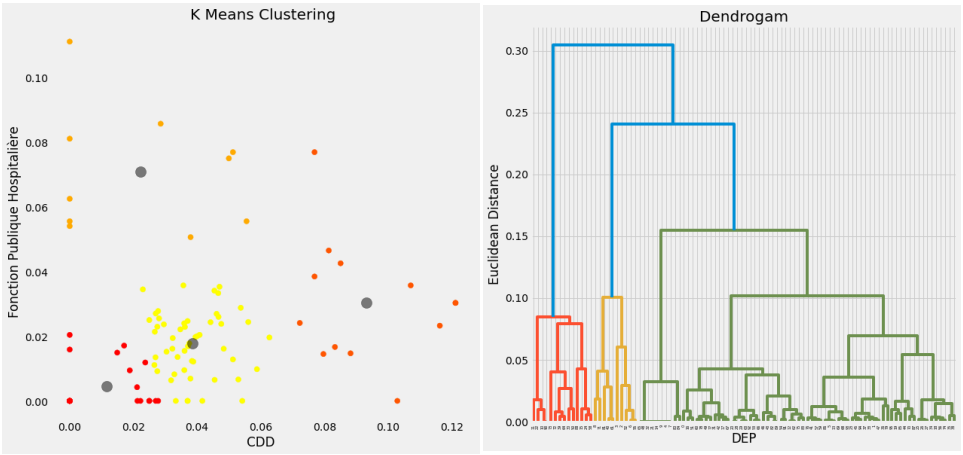


Figure 17: Cluster by CDD and by Employer type

Partial conclusion : Employee count and employer type and do not explain heterogeneity across French departments. On the other hand, the type of contract (especially CDD) and the economic sector show heterogeneity.

2 Conclusion

The exploratory analysis allowed us to see the distribution of the variables in the data sets. We observed that some characteristics were equally distributed across departments (gender) whereas some where heterogeneously distributed (inhabitants). We then studied the possibility of heterogeneity across departments through clustering analysis. The first data set seems more helpful to determine clusters. Maybe this could driven by the fact that the first data set contains fewer characteristics, hence it is easier to cluster. Within this data set, inhabitants and the type of household induced heterogeneity. As for the second data set, the type of contract (especially CDD) and the economic sector were the characteristics that induced the more heterogeneity.

3 Annex

	Code	Libellé
0	D1	Pas de scolarité ou arrêt avant la fin du primaire
1	D2	Aucun diplôme et scolarité interrompue à la fin du primaire ou avant la fin du collège
2	D3	Aucun diplôme et scolarité jusqu'à la fin du collège ou au-delà
3	D1_1	CEP (certificat d'études primaires)
4	D1_2	BEPC, brevet élémentaire, brevet des collèges, DNB
5	D1_3	CAP, BEP ou diplôme de niveau équivalent
6	D1_4	Baccalauréat général ou technologique, brevet supérieur, capacité en droit, DAEU, ESEU
7	D1_5	Baccalauréat professionnel, brevet professionnel, de technicien ou d'enseignement, diplôme équivalent
8	D1_6	BTS, DUT, Deug, Deust, diplôme de la santé ou du social de niveau bac+2, diplôme équivalent
9	D1_7	Licence, licence pro, maîtrise, diplôme équivalent de niveau bac+3 ou bac+4
10	D1_8	Master, DEA, DESS, diplôme grande école niveau bac+5, doctorat de santé
11	D1_9	Doctorat de recherche (hors santé)

Figure 18: Degree

	Code	Libellé
0	TYPMR11	Homme vivant seul
1	TYPMR12	Femme vivant seule
2	TYPMR20	Plusieurs personnes sans famille
3	TYPMR31	Famille principale monoparentale composée d'un homme avec enfant(s)
4	TYPMR32	Famille principale monoparentale composée d'une femme avec enfant(s)
5	TYPMR41	Famille principale composée d'un couple de deux 'actifs ayant un emploi'
6	TYPMR42	Famille principale composée d'un couple où seul un homme a le statut 'd'actif ayant un emploi'
7	TYPMR43	Famille principale composée d'un couple où seule une femme a le statut 'd'actif ayant un emploi'
8	TYPMR44	Famille principale composée d'un couple d'aucun 'actif ayant un emploi'

Figure 19: Household

	Code	Libellé
0	AZ	Agriculture, sylviculture et pêche
1	BZ	Industries extractives
2	CA	Fabrication de denrées alimentaires, de boisso...
3	CB	Fabrication de textiles, industries de l'habil...
4	CC	Travail du bois, industries du papier et impri...
5	CD	Cokéfaction et raffinage
6	CE	Industrie chimique
7	CF	Industrie pharmaceutique
8	CG	Fabrication de produits en caoutchouc et en pl...
9	CH	Métallurgie et fabrication de produits métalli...
10	CI	Fabrication de produits informatiques, électro...
11	CJ	Fabrication d'équipements électriques
12	CK	Fabrication de machines et équipements n.c.a.
13	CL	Fabrication de matériels de transport
14	CM	Autres industries manufacturières ; réparation...
15	DZ	Production et distribution d'électricité, de g...
16	EZ	Production et distribution d'eau ; assainissem...
17	FZ	Construction
18	GZ	Commerce ; réparation d'automobiles et de moto...
19	HZ	Transports et entreposage
20	IZ	Hébergement et restauration
21	JA	Edition, audiovisuel et diffusion
22	JB	Télécommunications
23	JC	Activités informatiques et services d'information
24	KZ	Activités financières et d'assurance
25	LZ	Activités immobilières
26	MA	Activités juridiques, comptables, de gestion, ...
27	MB	Recherche-développement scientifique
28	MC	Autres activités spécialisées, scientifiques e...
29	NZ	Activités de services administratifs et de sou...
30	OZ	Administration publique
31	PZ	Enseignement
32	QA	Activités pour la santé humaine
33	QB	Hébergement médico-social et social et action ...
34	RZ	Arts, spectacles et activités récréatives
35	SZ	Autres activités de services
36	TZ	Activités des ménages en tant qu'employeurs ; ...
37	UZ	Activités extra-territoriales

Figure 20: Economic Sector

	Code	Libellé
0	tr_0	0 poste (salariés présents en cours d'année, m...
1	tr_1	[1;10[postes
2	tr_2	[10;20[postes
3	tr_3	[20;50[postes
4	tr_4	[50;100[postes
5	tr_5	[100;250[postes
6	tr_6	[250;+[postes

Figure 21: Employee count

	Code	Libellé
0	ct_1	Fonction Publique d'État
1	ct_2	Fonction Publique Territoriale
2	ct_3	Fonction Publique Hospitalière
3	ct_4	Autres organismes publics administratifs
4	ct_5	Personnes morales de droit public soumises au ...
5	ct_6	Entreprises individuelles
6	ct_7	Particuliers Employeurs
7	ct_8	Organismes privés spécialisés et groupements d...
8	ct_9	Autres sociétés privées

Figure 22: Employer type

	Code	Libellé
0	O	Emploi ordinaire
1	A	Apprenti
2	X	Autres (emploi aidé, stagiaire, indemnité de c...

Figure 23: Job category

	Code	Libellé
0	C	Temps complet
1	D	Travail à domicile
2	F	Faible temps partiel
3	K	Condition d'emploi mixte à dominante temps com...
4	P	Temps partiel
5	Y	Condition d'emploi mixte à dominante temps non...

Figure 24: Job condition

	Code	Libellé
0	CDI	CDI
1	CDD	CDD
2	APP	Contrat d'apprentissage
3	TOA	Travail occasionnel ou à l'acte
4	TTP	Contrat de travail temporaire
5	AUT	Autre

Figure 25: Type of contract

References

(N.d.). URL: https://www.kaggle.com/roshansharma/mall-customers-clustering-analysis/notebook.