

Master 2 Data Analyst

Projet Data Mining

Réalisé par

Jivaldo Julien et Ilyes Hadi

Novembre 2021

Table des matières

<i>Table des matières.....</i>	<i>2</i>
<i>Introduction.....</i>	<i>3</i>
<i>1) Construction de la base des données et présentation des variables.....</i>	<i>4</i>
<i>2) Statistiques descriptives</i>	<i>5</i>
a) Moyennes, écart-types et coefficients de variations.....	5
b) Matrice de corrélation	6
<i>3) L'analyse en composante principale.....</i>	<i>7</i>
<i>4) Classification : K means et CAH.....</i>	<i>10</i>
a) K-means.....	10
b) Classification Ascendante Hiérarchique	11
<i>5) Arbre de décision</i>	<i>14</i>
<i>6) Régression logistique</i>	<i>17</i>
<i>Conclusion.....</i>	<i>20</i>

Introduction

Le projet que nous avons réalisé est une brève étude sur 100 quartiers politiques en France. Elle porte plus précisément sur des données concernant le tissu économique des QPV. A l'aide de différentes techniques d'analyse de données et de classification, nous allons explorer et faire ressortir les informations principales dans notre échantillon.

Le but serait d'appréhender la structure des QPV en France et comprendre ce qui pourrait expliquer de potentielles différences entre les différents types d'établissement dans ces quartiers. Pour finir, l'objectif est également de voir si la répartition des différents types d'établissements influence le taux d'emploi précaire dans les QPV. Ici ce qui définit les emplois précaires, ce sont tous les types de contrats qui ne sont pas des contrats à durées indéterminées tels que les contrats à durées déterminés, l'intérim, les contrats aidés.

Dans une première partie, nous créerons et présenterons notre base de données. Dans une seconde partie, nous présenterons quelques statistiques descriptives telles que les moyennes, les écarts-types, les coefficients de variations ainsi que la matrice de corrélation. Ensuite nous réaliserons une analyse en composante principale sur l'échantillon. Nous continuerons avec une classification, plus précisément avec la méthode des K-means et de classification ascendante hiérarchique. Enfin, après avoir fait une partie sur l'arbre de décision, nous terminerons avec une partie sur une régression logistique que nous aurons réalisé.

1) Construction de la base des données et présentation des variables

Les données que nous avons utilisées pour ce projet sont issues des études de l'INSEE sur l'insertion professionnelle dans différents quartiers politiques en France. Nous avons concentré notre étude sur 100 QPV choisie aléatoirement en France. A partir de plusieurs études et données sur l'année 2017 de l'INSEE, nous avons pu construire une base de données regroupant des variables économiques, démographiques et sur l'insertion professionnelle. Nous avons cependant transformé plusieurs variables pour notre base de données. Nous avons créé la variable **TX_INDUS** qui est le ratio de la variable INDUS (nombre d'établissement industrielle du QPV) sur NBETAB (nombre d'établissements total du QPV), nous obtenons au final la part des établissement industrielle dans le QPV. De ce même procédé nous avons créé les variable **TX_CONSTR**, **TX_COM_TRANSP**, **TX_COM_GROS** et **TX_COM_DETAIL**, **TX_SERV_ENT** et **TX_SERV_PAR**. Le tableau ci-dessus présente l'ensemble des variables de notre base de données finale ainsi que leurs significations.

Nom de la variable	Explication
CODGEO	Code Géographique
LIBGEO	Libellé géographique
TX_INDUS	Part des établissements industriels (ratio de INDUS et NBETAB)
TX_CONSTR	Part des établissements de constructions (ratio de CONSTR et NBETAB)
TX_COM_TRANSP	Part des établissements de commerce, transport, hébergement et transport (ratio de COM_TRANSP et NBETAB)
TX_COM_GROS	Part des établissements de commerce de gros (ratio de COM_GROS et NBETAB)
TX_COM_DETAIL	Part des établissements de commerce de détail (ratio de COM_DETAIL et NBETAB)
TX_SERV_ENT	Part des d'établissements proposant des services aux entreprises (ratio de SERV_ENT et NBETAB)
TX_SERV_PAR	Part des établissements proposant des services aux particuliers (ratio de SERV_PAR et NBETAB)
TX_TOT_EPREC	Part des personnes en emploi précaire (contrat d'apprentissage, Placés par une agence d'intérim, Emplois-jeunes, CES, contrats de qualification, stagiaires rémunérés en entreprise, autres emplois à durée limitée) parmi les personnes ayant un emploi.

2) Statistiques descriptives

a) Moyennes, écart-types et coefficients de variations

Statistiques / Variables	moyenne	ecart-type	coefficient de variation
TX_INDUS	8,43	2,81	0,33
TX_CONSTR	13,51	2,43	0,18
TX_COM_TRANSP	31,97	3,14	<u>0,10</u>
TX_COM_GROS	4,73	1,44	0,30
TX_COM_DETAIL	14,48	2,25	<u>0,16</u>
TX_SERV_ENT	24,66	4,30	<u>0,17</u>
TX_SERV_PAR	21,43	2,36	<u>0,11</u>
TX_TOT_EPREC	13,33	1,94	<u>0,15</u>

Le tableau ci-dessus représente l'ensemble des moyennes, des écart-types ainsi que des coefficients de variation de l'ensemble de nos variables. Dans un premier temps, nous allons nous focaliser sur les moyennes de nos variables. Nous pouvons voir dans un premier temps que le taux d'emploi précaire (**TX_TOT_EPREC**) a une moyenne de 13,33% sur l'ensemble de nos QPV. Que le taux d'établissement proposant des services aux entreprises (**TX_SERV_ENT**) est supérieur à ceux des établissements aux particuliers (**TX_TOT_PAR**) avec des valeurs respectives de 24,66% contre 21,43%. La moyenne de la part des établissements de gros (**TX_COM_GROS**) est uniquement de 4,73%, on retrouve ensuite la part des établissements industriels (**TX_INDUS**) avec une moyenne de 8,43%, puis les établissements de construction (**TX_CONSTR**) avec une moyenne de 13,51%. Enfin, on retrouve la part des établissements de commerce en détail (**TX_COM_DETAIL**) avec une moyenne de 14,48% et pour finir le "secteur" avec la plus grande part et le secteur des commerces et transports (**TX_COM_TRANSP**) avec une part moyenne de 31,97%.

Concernant les écarts-types, leurs interprétations sont limitées. En effet, plus une moyenne est grande, plus l'écart-type associé à cette moyenne sera grand. Son interprétation n'est donc pas intéressante, l'interprétation des coefficients de variations est plus intéressante. L'interprétation pour les coefficients de variation est la suivante : plus la valeur de coefficient sera faible (proche de 0) plus les observations de la variable en question seront proches de la moyenne. Le coefficient de variation le moins élevé est celui de la variable **TX_COM_TRANSP** avec une valeur de 0,10. Donc cette variable a les observations les moins disposées autour de la moyenne. On retrouve aussi la variable **TX_SERV_PAR** avec un coefficient de variation de 0,11, puis la variable **TX_TOT_EPREC** avec une valeur de 0,15. Ensuite, nous avons la variable **TX_COM_DETAIL** (CV = 0,16), puis **TX_SERV_ENT** avec un coefficient de 0,18. Le 3ème plus fort coefficient de variation est celui de la variable **TX_CONSTR** avec une valeur de 0,18, puis en 2ème il y a la variable **TX_COM_GROS** avec un coefficient de variation de 0,30. Enfin, le plus fort coefficient de variation est de 0,33 pour la variable **TX_INDUS**, cette variable a donc les observations les plus dispersées autour de la moyenne de notre base de données.

b) Matrice de corrélation

Matrice de corrélation :



La figure ci-dessus est la matrice de corrélation de notre base de données, elle permet de savoir si des variables de notre base de données sont corrélées entre elles, et à quelles hauteurs elles sont corrélées. Lorsque la valeur de corrélation entre deux variables tend vers le bleu foncé cela veut dire que ces deux variables sont fortement corrélées positivement (la valeur se rapproche de 1), et lorsque cette valeur tend vers le rouge foncé (la valeur se rapproche de -1) alors les deux variables en question seront fortement corrélées négativement. Les corrélations qui ressortent le plus de cette matrice de corrélation sont les corrélations entre :

Les variables **TX_COM_DETAIL** et **TX_COM_TRANS**, leur corrélation s'élève à 0,76. C'est plutôt une forte corrélation ce qui veut dire que la part des établissements de commerce de détail est corrélée positivement à la part des établissements de commerce et transport. Ce qui peut s'expliquer par le fait que les établissements de commerce et transport peuvent être également des établissements de détail.

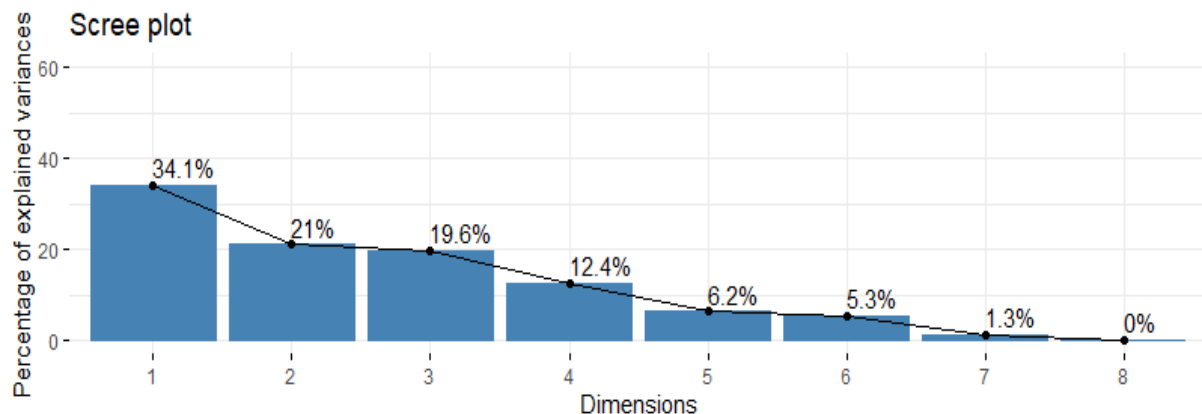
Les variables **TX_SERV_ENT** et **TX_INDUS**, leur corrélation est négative et s'élève à -0,62. Donc le taux des établissements proposant des services aux entreprises est négativement corrélé au taux des établissements industriels. Cela peut s'expliquer par le fait que les établissements industriels ne proposent pas majoritairement des services aux entreprises.

Les variables **TX_SERV_ENT** et **TX_COM_DETAIL**, la valeur de leur corrélation est de -0,53, elle est donc négative. Le taux des établissements proposant des services aux entreprises est donc corrélé négativement aux taux des établissements de commerce de gros. On peut donc supposer que les établissements proposant des services aux entreprises ne sont pas des établissements de détail, mais plutôt des établissements de commerce de gros.

3) L'analyse en composante principale

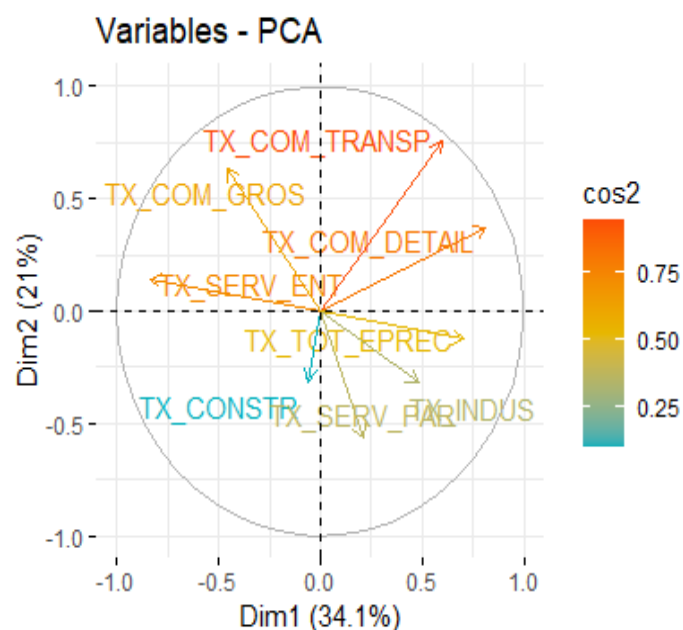
Dans cette partie nous mettons en place une analyse en composante principale sur nos données dont nous analyserons les principaux résultats. Elle va nous permettre de mieux appréhender nos données, repérer les éventuelles corrélations entre variables, visualiser les variables qui vont expliquer la variabilité des données et enfin avoir un premier aperçu des différents groupes d'individus.

Histogramme des valeurs propres :



Nous pouvons voir que les trois premières composantes principales expliquent environ 75% de l'inertie totale. En effet, la première dimension explique 34% de l'inertie totale, la deuxième 21% et enfin la dernière 20%. Nous allons donc interpréter ces trois composantes.

Cercle des corrélations avec contributions des variables :



Projection des individus sur le plan factoriel :



Tableau de résultat de l'analyse en composante principale :

Call:
PCA(X = Etablissements[, 3:10])

Eigenvalues

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5	Dim.6	Dim.7	Dim.8
Variance	2.727	1.677	1.571	0.993	0.499	0.428	0.105	0.000
% of var.	34.083	20.960	19.640	12.419	6.240	5.346	1.312	0.000
Cumulative % of var.	34.083	55.043	74.683	87.102	93.342	98.688	100.000	100.000

Individuals (the 10 first)

	Dist	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
1	1.956	-0.002	0.000	0.000	-0.214	0.027	0.012	1.877	2.241	0.920
2	3.105	-0.690	0.175	0.049	-1.856	2.054	0.357	-0.779	0.386	0.063
3	3.965	3.235	3.838	0.666	1.705	1.734	0.185	0.703	0.315	0.031
4	2.610	0.492	0.089	0.035	-0.446	0.119	0.029	-2.346	3.503	0.808
5	7.920	-0.806	0.238	0.010	7.144	30.440	0.814	-0.450	0.129	0.003
6	1.771	1.013	0.376	0.327	0.282	0.047	0.025	0.654	0.272	0.136
7	2.052	1.723	1.089	0.705	-0.678	0.274	0.109	0.421	0.113	0.042
8	2.355	-0.071	0.002	0.001	0.789	0.371	0.112	1.348	1.157	0.328
9	3.379	1.044	0.399	0.095	-2.721	4.417	0.649	-0.863	0.474	0.065
10	2.086	-1.834	1.234	0.773	-0.277	0.046	0.018	-0.774	0.382	0.138

Variables

	Dim.1	ctr	cos2	Dim.2	ctr	cos2	Dim.3	ctr	cos2
TX_INDUS	0.484	8.590	0.234	-0.320	6.116	0.103	-0.430	11.774	0.185
TX_CONSTR	-0.060	0.130	0.004	-0.316	5.942	0.100	-0.796	40.300	0.633
TX_COM_TRANSP	0.603	13.337	0.364	0.758	34.232	0.574	-0.069	0.302	0.005
TX_COM_GROS	-0.460	7.766	0.212	0.634	23.949	0.402	-0.094	0.560	0.009
TX_COM_DETAIL	0.813	24.242	0.661	0.374	8.331	0.140	0.189	2.264	0.036
TX_SERV_ENT	-0.840	25.882	0.706	0.146	1.263	0.021	0.425	11.471	0.180
TX_SERV_PAR	0.213	1.662	0.045	-0.567	19.205	0.322	0.651	26.971	0.424
TX_TOT_EPREC	0.708	18.390	0.501	-0.127	0.963	0.016	0.316	6.358	0.100

Tout d'abord sur le premier axe, les variables les mieux représentées sont la part de commerce de détail dans les QPV, ensuite la part d'établissement de service aux entreprises et enfin le taux d'emploi précaire avec respectivement des $\cos^2 = 0,661 ; 0,706 ; 0,501$. Ensuite, concernant la contribution aux axes des variables, la part de commerce de détail contribue à 24% à la formation de l'axe, la part d'établissement aux services des entreprises à hauteur de 26% et le taux de personnes en emploi précaire 18%. Enfin, le cercle de corrélation nous montre que le taux de commerce de détail et le taux de personnes en emploi précaire dans les QPV sont positivement liés. Ces dernières sont cependant négativement liées à la part d'établissement de service aux entreprises. Finalement, ce que l'on peut dire du premier axe est qu'il oppose les QPV en fonction des trois variables les mieux représentées. Plus un QPV se situe à droite du plan factorielle, plus sa part de commerce de détail et son taux de personnes en emploi précaire seront élevés et inversement pour les QPV à gauche du plan. Plus on se situe à gauche du plan, plus la part d'établissement au service des entreprises sera élevée et inversement.

Concernant le second axe, la variable la mieux représentée est la part de commerce lié aux transports avec un $\cos^2 = 0.574$. Elle contribue à hauteur de 34% à la formation de l'axe. Le cercle des corrélations nous permet de conclure que plus les QPV situés vers le haut du plan factoriel, plus ces QPV ont une part de commerce liée au transport élevé.

Enfin, lorsque l'on s'intéresse au troisième axe, l'unique variable bien représentée est la part d'entreprise dans la construction avec un $\cos^2 = 0,633$. Elle contribue à hauteur de 40% au troisième axe et semble y être corrélée négativement. Le troisième axe opposerait donc les QPV en fonction de la part d'établissement de construction.

La projection des individus sur le plan factoriel nous permet de supposer qu'il existe 4 groupes plutôt homogènes dans notre échantillon. Nous allons vérifier cette hypothèse en mettant en place une classification avec la méthode des K-Means et la classification ascendante hiérarchique.

4) Classification : K means et CAH

Dans cette partie, l'objectif est de pouvoir connaître la répartition adéquate de nos observations dans notre base de données. En d'autres termes, regrouper les individus en classes homogènes en fonction des variables. Pour cela nous utiliserons deux méthodes de classification non supervisée : les K-Means et la classification ascendante hiérarchique.

a) K-means

Afin de connaître quelle est la meilleure répartition des individus en classe dans notre échantillon et de les visualiser, nous mettons en place la méthode des K-means.

Résultats K means avec l'hypothèse des 4 groupes :

```
k-means clustering with 4 clusters of sizes 23, 45, 10, 22

Cluster means:
  TX_INDUS TX_CONSTR TX_COM_TRANSP TX_COM_GROS TX_COM_DETAIL TX_SERV_ENT TX_SERV_PAR TX_TOT_EPPEC
1  1.15471195  0.7577163   -0.6109294   -0.3617119   -0.6179928   -0.60354285  -0.243424842   0.1904714
2  -0.00131558 -0.4419999    0.5100814   -0.2965255   -0.7810960   -0.38291847   0.475958386   0.4837855
3  -0.41102113  0.5300546    1.0180523    1.5218939   -0.2179239   0.08603598  -1.569820300  -0.7764243
4  -1.01768011 -0.1290012   -0.8674005    0.2929128   -0.8525567    1.37511168  -0.005506956  -0.8357704

Clustering vector:
[1] 2 1 2 1 3 2 2 2 1 4 2 4 4 2 2 2 1 2 4 2 2 3 2 1 4 2 2 1 1 2 2 2 3 2 4 2 1 2 2 2 1 2 2 2 1 2 4 4 4 1 3 1 2 3 1 1 1 2 1 1 4 3 1 3 4 2 2 2 4 2 2 2
[88] 2 4 3 4 4 4 3 4 4 4 4 3 4

within cluster sum of squares by cluster:
[1] 103.08015 175.28327 85.79806 89.98157
(between_SS / total_SS = 42.7 %)
```

La partition que nous avons choisie explique seulement 42,7% de l'inertie totale.

Description des classes avec les moyennes :

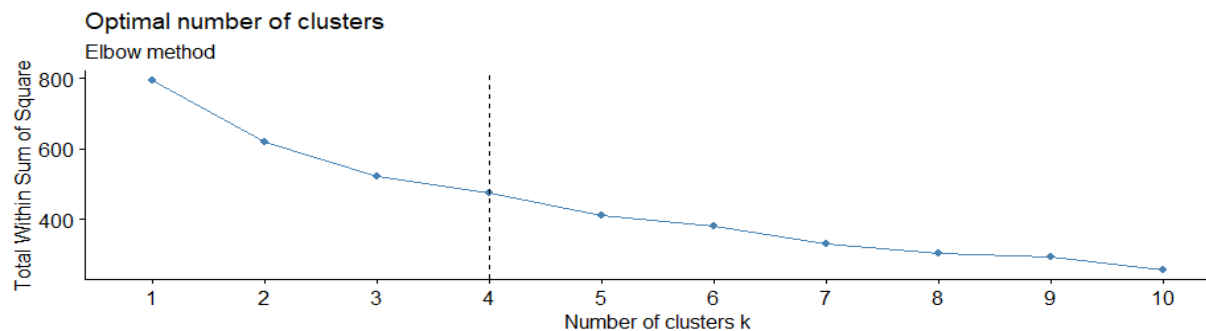
	TX_INDUS	TX_CONSTR	TX_COM_TRANSP	TX_COM_GROS	TX_COM_DETAIL	TX_SERV_ENT	TX_SERV_PAR	TX_TOT_EPPEC
clus_1	11.676385	15.35409	30.05042	4.206874	13.08770	22.06784	20.85126	13.69915
clus_2	5.565367	13.19626	29.24407	5.148297	12.55981	30.58158	21.41273	11.71122
clus_3	7.271921	14.80008	35.17197	6.915709	13.98807	25.03495	17.72108	11.82618
clus_4	8.424437	12.43457	33.57490	4.300619	16.23640	23.01714	22.54895	14.26733

Nous avons calculé les moyennes de nos classes afin de pouvoir les décrire brièvement. Nous pouvons dire que la première classe se caractérise par un taux d'établissement industriel et de construction plus élevé que les autres classes. La classe 1 va également avoir des QPV avec une part de commerce en gros faible comparé aux autres groupes. Quant à la classe 2, elle se caractérise par des QPV ayant des taux d'établissement industriel les plus faibles en moyenne et une part des personnes en emploi précaires le plus faible également. Ensuite, la classe trois va plutôt concentrer les QPV avec beaucoup de commerce en gros et de transport. Enfin, la dernière classe a le taux de personnes en emploi précaire le plus élevé.

On peut ensuite opposer les classes 1 et 4 aux classes 2 et 3 en termes de taux de personnes en emploi précaire et de taux d'établissement d'industrie. Le groupe 3 et 4 semblent concentrer davantage de QPV ayant des taux de commerce de transport plus élevés que le groupe 1 et 2. Enfin, il est intéressant de noter que le cluster 2 a nettement davantage de QPV ayant des taux d'établissement de service aux entreprises et le groupe 3 un taux d'établissement de service au particulier plus bas que les autres groupes.

Nous cherchons donc à connaître la répartition adéquate pour nos données, pour cela nous utiliserons la méthode du “coude”. Cette méthode consiste à repérer sur une courbe d’inertie le point à partir duquel la variance ne diminue plus significativement et crée une cassure, d’où l’appellation “coude”.

Coude d’inertie, méthode du “coude” :

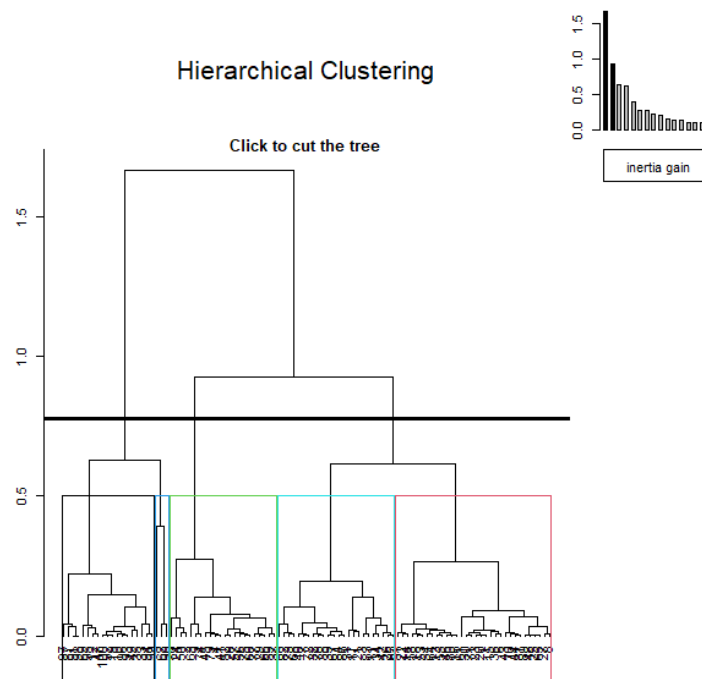


Nous pouvons voir ici que la cassure se fait lorsque k vaut 4. Donc, la répartition adéquate de nos données serait bien en 4 classes.

b) Classification Ascendante Hiérarchique

Après avoir pu déterminer le nombre de classes optimal avec la méthode des K-means, nous allons tenter de connaître le découpage de notre échantillon avec la méthode de la classification ascendante hiérarchique.

Dendrogramme :



Ici, le dendrogramme nous suggère une fragmentation de notre échantillon en 5 groupes. On voit par ailleurs que le second cluster est composé de très peu individus.

Projection des groupes sur le plan factoriel :

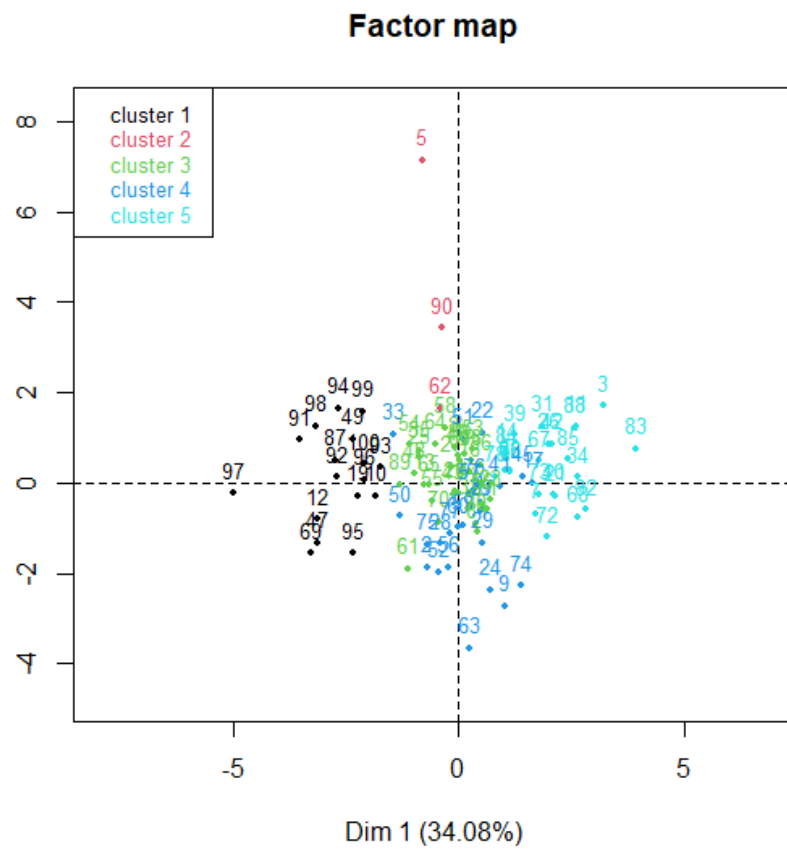


Tableau description des groupes :

Description of each cluster by quantitative variables						
=====						
\$`1`						
	v.test	Mean in category	overall mean	sd in category	overall sd	p.value
TX_SERV_ENT	6.747291	31.079722	24.664758	3.3878095	4.281223	1.506311e-11
TX_COM_GROS	2.648446	5.568642	4.727056	0.8794510	1.430903	8.086273e-03
TX_COM_TRANSP	-4.020400	29.178208	31.971197	2.3536722	3.128261	5.809944e-05
TX_COM_DETAIL	-5.149748	11.917657	14.478518	1.1456591	2.239253	2.608363e-07
TX_INDUS	-5.170983	5.214010	8.428138	0.9270905	2.798936	2.328655e-07
TX_TOT_EPPEC	-5.485295	10.982364	13.330189	1.1672192	1.927387	4.127811e-08
\$`2`						
	v.test	Mean in category	overall mean	sd in category	overall sd	p.value
TX_COM_TRANSP	4.583468	40.165364	31.971197	2.2286532	3.128261	4.573280e-06
TX_COM_GROS	3.560192	7.638385	4.727056	4.5208331	1.430903	3.705832e-04
TX_SERV_PAR	-5.182316	14.471540	21.425723	0.5697622	2.348089	2.191471e-07
\$`3`						
	v.test	Mean in category	overall mean	sd in category	overall sd	p.value
TX_SERV_PAR	3.105148	22.518981	21.425723	1.436505	2.348089	0.001901837
TX_INDUS	-2.282290	7.470305	8.428138	1.388893	2.798936	0.022472229
TX_CONSTR	-3.014102	12.415889	13.510185	1.543278	2.421310	0.002577410
\$`4`						
	v.test	Mean in category	overall mean	sd in category	overall sd	p.value
TX_INDUS	6.427665	11.55990	8.428138	2.593226	2.798936	1.295791e-10
TX_CONSTR	4.430020	15.37742	13.510185	2.037092	2.421310	9.422454e-06
TX_SERV_PAR	-2.041390	20.59130	21.425723	2.176347	2.348089	4.121206e-02
TX_COM_TRANSP	-2.831200	30.42944	31.971197	2.281935	3.128261	4.637371e-03
TX_COM_DETAIL	-3.272602	13.20285	14.478518	1.434992	2.239253	1.065625e-03
TX_SERV_ENT	-3.519316	22.04193	24.664758	2.211933	4.281223	4.326605e-04
\$`5`						
	v.test	Mean in category	overall mean	sd in category	overall sd	p.value
TX_COM_DETAIL	6.954387	17.263649	14.478518	1.4911625	2.239253	3.540972e-12
TX_COM_TRANSP	5.216005	34.889461	31.971197	1.7033372	3.128261	1.828228e-07
TX_TOT_EPPEC	4.696163	14.948998	13.330189	1.6949632	1.927387	2.650944e-06
TX_CONSTR	-2.142743	12.582278	13.510185	1.8745721	2.421310	3.213375e-02
TX_COM_GROS	-3.240797	3.897691	4.727056	0.7742545	1.430903	1.191961e-03
TX_SERV_ENT	-4.400687	21.295208	24.664758	2.3609634	4.281223	1.079088e-05

Nous voyons que ces groupes sont plutôt différents de la méthode des K-means. Nous remarquons que la description des groupes fait écho avec l'analyse en composante principale. Ici, le premier groupe se caractérise par des QPV avec une forte part d'établissement de service aux entreprises (moyenne de classe 31% contre 24 % moyenne globale) et un taux de personnes en emploi précaire nettement inférieur à la moyenne globale.

On peut opposer à ce groupe le cluster 5 qui se caractérise par des QPV qui ont un taux d'établissement de service aux entreprises plutôt faible comparé à la moyenne générale (21% contre 24%). De plus, elle se caractérise également par plus d'établissements de transport et de commerce de détail.

Dans la seconde classe et troisième classe, on voit que peu de variables y sont significativement associées. Lorsque l'on s'intéresse au troisième groupe, nous constatons que pour le peu de variables significatives liées au groupe, les moyennes sont plutôt homogènes si on les compare avec la moyenne globale.

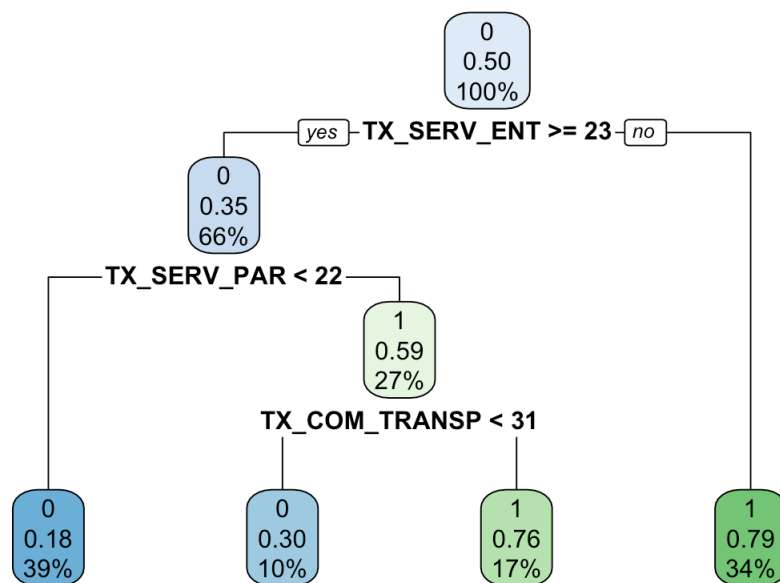
Enfin, la quatrième classe va regrouper des QPV qui ont des classes plutôt similaires à la moyenne générale dans les différentes variables significativement liées au groupe. Il est pertinent de noter que ce groupe se caractérise particulièrement par moins d'établissements de transport que la moyenne générale.

Selon la méthode, nous n'obtenons pas la même classification. Toutefois, dans la classification ascendante hiérarchique, la classe en plus est composée de seulement 3 individus qui semblent avoir des valeurs aberrantes. Finalement, nous avons une classification plutôt similaire. D'autant plus que globalement les classes s'opposent toujours par rapport aux mêmes variables.

5) Arbre de décision

Nous appliquons dans cette partie, une méthode de classification hiérarchique descendante supervisée. L'arbre de décision permet d'expliquer une variable en fonction d'autres variables, on aura donc une variable à expliquer à partir d'autres variables dites explicatives. Ici nous allons créer une variable à expliquer binaire. Nous allons créer la variable **discretise** qui prendra une valeur de 1 lorsque la variable **TXT_TOT_EPREC** de l'observation est au-dessus de la médiane (La médiane de **TXT_TOT_EPREC** est égale à 13,209) sinon la variable **discretise** prend la valeur 0. L'objectif est d'expliquer la variable **discretise** à partir des variables **TX_INDUS**, **TX_CONSTR**, **TX_COM_TRANSP**, **TX_COM_GROS**, **TX_COM_DETAIL**, **TX_SERV_ENT**, **TX_SERV_PAR**. Pour cela, nous réaliserons dans un premier temps un arbre de décision dit globale avec l'ensemble des observations des variables citées précédemment. Cependant, certains individus de cette base sont mal classés, et le taux d'erreur globale ne sera pas réaliste. Le taux d'erreur global est le rapport entre le nombre de fois où l'arbre s'est trompé dans la classification des individus et le nombre total d'individus dans la base de données. Il existe cependant une méthode afin d'améliorer ce taux d'erreur global et de rendre l'arbre encore plus performant.

Arbre de décision globale :



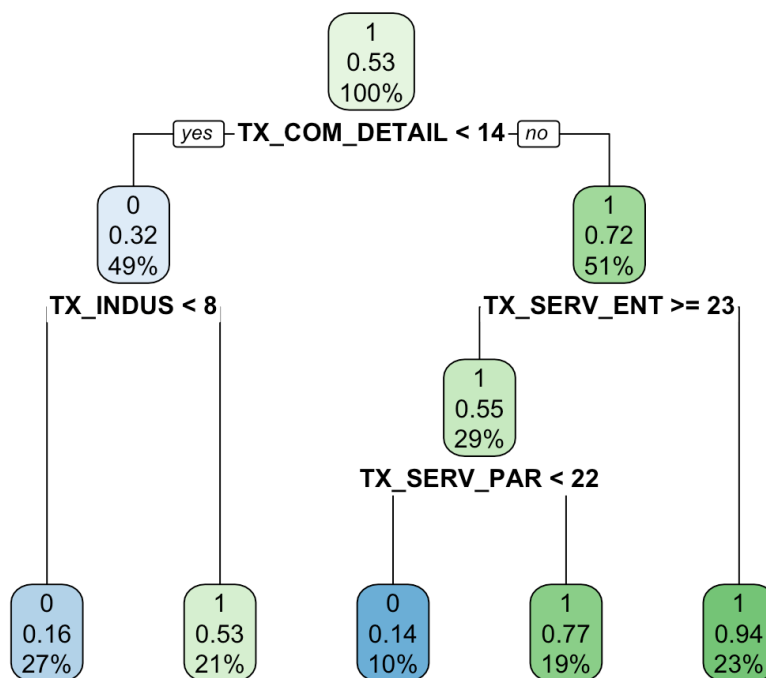
L'arbre de décision ci-dessus est l'arbre de décision réalisé avec l'ensemble des observations de la base de données. Concernant l'interprétation de cet arbre, nous pouvons dire que si le taux d'établissement de service aux entreprises est inférieur à 23% les QPV seront dans le groupe supérieur à la médiane du taux de personnes en emploi précaire.

Ensuite, si le taux d'établissement de service aux entreprises est supérieur à 23% et si le taux d'établissement de services est inférieur à 22% alors les QPV auront un taux de personnes en emploi précaire inférieur à 13%.

Enfin, si le taux d'établissement de service aux entreprises est supérieur à 23%, le taux d'établissement de services est supérieur à 22% et le taux de commerce de transport est inférieur à 31% alors les QPV auront un taux de personnes en emploi précaire inférieur à 13% et dans le cas contraire (taux de commerce de transport est supérieur à 31%) les QPV seront dans le groupe supérieur à la médiane du taux de personnes en emploi précaire dans l'échantillon.

Avec cet arbre de décision nous obtenons un taux d'erreur globale de 0,21. Donc l'arbre s'est trompé 2 fois sur 10 dans la classification, ce qui est un résultat assez satisfaisant, on peut donc dire que notre arbre est performant. Cependant il est possible de rendre cet arbre encore plus performant en améliorant le taux d'erreur avec une méthode que nous avons énoncé au début. Cette méthode est la méthode dite d'apprentissage-test. Dans cette méthode, il y aura la création d'un sous-échantillon de notre base de données. Dans ce sous échantillon, 70% de notre base de données sera prise, ce sous échantillon sera le sous échantillon dit d'apprentissage. Les 30% restants seront le sous échantillon dit de test. On réalisera un nouvel arbre avec l'échantillon apprentissage. Finalement, nous obtenons un arbre de décision d'apprentissage et un nous calculerons le nouveau terme d'erreur sur le sous échantillon test.

Arbre de décision apprentissage :



L'arbre de décision ci-dessus provient du sous-échantillon d'apprentissage. Concernant l'interprétation de l'arbre, nous pouvons dire que dans un premier temps si le taux de commerce de détail est supérieur à 14% et que le taux d'établissement de service aux entreprises est inférieur à 23% alors les QPV seront dans le groupe supérieur à la médiane du taux de personnes en emploi précaire dans l'échantillon.

Dans un autre cas, si le taux de commerce de détail est supérieur à 14% et que le taux d'établissement de service aux entreprises est supérieur à 23% et le taux d'établissement de service proposant des services aux particuliers est inférieur à 22% alors les QPV auront un taux de personnes en emploi précaire inférieur à 13%. Dans le cas contraire, les QPV seront dans le groupe supérieur à la médiane du taux de personnes en emploi précaire dans l'échantillon.

Enfin, si le taux de commerce de détail est inférieur à 14% et le taux d'établissement industriel est supérieur à 8% alors les QPV seront dans le groupe supérieur à la médiane du taux de personnes en emploi précaire dans l'échantillon. Dans le cas contraire, si le taux d'établissement industriel est inférieur à 8%, les QPV auront un taux de personnes en emploi précaire inférieur à 13%

Pour synthétiser, cet arbre nous montre que pour être un QPV avec un taux de personne en emploi précaire en dessous de la médiane, il faut un taux d'établissement commerce de détail et d'établissement industriel moindre

Le terme d'erreur associé à ce nouvel arbre du sous échantillon est le terme d'erreur du sous échantillon test. Celui s'élève également à 0.1666667, cela signifie que l'arbre de décision issu de la méthode apprentissage-test est plus performant que l'arbre de décision globale. En effet, cela signifie que cet arbre se trompe 1,6 fois sur 10.

6) Régression logistique

Afin de réaliser la régression logistique, nous avons repris en tant que variable expliquée la variable **discretise** de l'arbre de décision (Partie 5). Cette variable prend une valeur 1 si le taux des personnes en emploi précaire (**TX_TOT_EPREC**) se situe au-dessus de la médiane qui est de 13,209. Cela signifie que si une observation a pour valeur 1 pour la variable discretise alors elle se situe dans les 50% des QPV ayant le plus fort de taux de personnes en emploi précaire. Si cette variable prend 0 cela signifie que le QPV en question se situe dans les 50% des QPV ayant le plus faible taux des personnes en emploi précaire. Notre régression logistique aura le modèle suivant :

$$\Pr (\text{Discretise} \mid \mathbf{x}_i) = \beta_1 (\text{TX_INDUS}) + \beta_2 (\text{TX_CONSTR}) + \beta_3 (\text{TX_COM_TRANSP}) + \beta_4 (\text{TX_COM_GROS}) + \beta_5 (\text{TX_COM_DETAIL}) + \beta_6 (\text{TX_SERV_ENT}) + \beta_7 (\text{TX_SERV_PAR})$$

Après avoir effectué la régression logistique nous obtenons les résultats suivants :

Coefficients: (1 not defined because of singularities)

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	14.5514	9.9157	1.468	0.14224
TX_INDUS	0.0478	0.1600	0.299	0.76518
TX_CONSTR	-0.5200	0.1689	-3.078	0.00208 **
TX_COM_TRANSP	0.1846	0.1692	1.091	0.27518
TX_COM_GROS	-0.6570	0.2609	-2.518	0.01181 *
TX_COM_DETAIL	-0.2674	0.2498	-1.070	0.28449
TX_SERV_ENT	-0.2786	0.1555	-1.792	0.07314 .
TX_SERV_PAR	NA	NA	NA	NA

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

A la suite de notre régression logistique, un coefficient n'est pas défini, c'est le coefficient de la variable **TX_SERV_PAR** aucune valeur n'apparaît. Cette absence de valeur est dû au fait que la variable en question est linéairement liée aux autres variables, d'où le NA à la place du coefficient estimé. La seule solution pour pallier ce problème est de supprimer la variable **TX_SERV_PAR** de notre modèle et d'effectuer une nouvelle fois la régression logistique. Nous obtenons donc ce nouveau modèle :

$$\Pr(\text{Discretise} \mid \mathbf{x}_i) = \beta_1(\text{TX_INDUS}) + \beta_2(\text{TX_CONSTR}) + \beta_3(\text{TX_COM_TRANSP}) + \beta_4(\text{TX_COM_GROS}) + \beta_5(\text{TX_COM_DETAIL}) + \beta_6(\text{TX_SERV_ENT})$$







Nous obtenons les résultats suivants pour cette nouvelle régression logistique sur R :

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	14.5514	9.9157	1.468	0.14224
TX_INDUS	0.0478	0.1600	0.299	0.76518
TX_CONSTR	-0.5200	0.1689	-3.078	0.00208 **
TX_COM_TRANSP	0.1846	0.1692	1.091	0.27518
TX_COM_GROS	-0.6570	0.2609	-2.518	0.01181 *
TX_COM_DETAIL	-0.2674	0.2498	-1.070	0.28449
TX_SERV_ENT	-0.2786	0.1555	-1.792	0.07314 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Il n'y a pas de coefficient manquant dans cette nouvelle régression logistique, cependant les coefficients des différentes valeurs des coefficient estimés des variables ne sont pas directement interprétables. En effet, il est nécessaire de convertir les coefficients suivants en termes d'odds ratio afin de pouvoir les interpréter convenablement, actuellement ces coefficients sont en log d'odds ratio. Pour transformer ces coefficients en odds ratio il suffit d'appliquer la fonction exponentielle à chacun des coefficients. Une fois la fonction exponentielle appliquée, nos coefficients seront en odds ratio. Nous utilisons la fonction **odds.ratio()** de la library **questionr** pour effectuer cette transformation.

Variable	N	Odds ratio	p
TX_INDUS	100		0.765
TX_CONSTR	100		0.002
TX_COM_TRANSP	100		0.275
TX_COM_GROS	100		0.012
TX_COM_DETAIL	100		0.284
TX_SERV_ENT	100		0.073

Le tableau ci-dessus nous présente l'ensemble des coefficients estimé en termes d'odds ratio pour l'ensemble de nos variables de notre régression logistique ainsi que les p-values associées. Les coefficients des variables suivants sont significatifs :

- La variable **TX_CONSTR** a un coefficient significatif au seuil de 1% (p-value = 0,002) avec une valeur de 0,59. Cela signifie entre autre que lorsque le taux d'établissement de construction augmente de 1% alors les chances pour un QPV que son taux de personnes ayant un emploi précaire soit au-dessus de la médiane décroît d'un facteur de 0,59. On peut donc supposer que les établissements de construction favorisent la baisse des personnes en ayant un emploi précaire dans les QPV.
- La variable **TX_COM_GROS** a également un coefficient significatif cependant celui-ci est significatif au seuil de 5% (p-value = 0,012), la valeur de son coefficient est de 0,52. Donc, lorsque le taux d'établissement de commerce de gros augmente de 1%, alors les chances pour un QPV que son taux de personne ayant un emploi précaire décroît d'un facteur de 0,52. A la suite de cette analyse, nous pouvons supposer à l'aide de nos résultats que les établissement de commerce de gros favorisent la baisse des personnes ayant un emploi précaire dans les QPV
- La variable **TX_SERV_ENT** a un coefficient d'une valeur de 0,76 et est significatif au seuil de 10% (p-value = 0,073). Donc lorsque le taux d'établissement de commerce proposant des services aux entreprises augmente de 1% alors les chances pour un QPV que son taux d'emploi précaire soit au-dessus de la médiane décroissent d'un facteur de 0,73. On peut donc supposer que les établissements proposant des services aux entreprises favorisent eux aussi la baisse des personnes ayant un emploi précaire.

A la suite de cette régression nous avons pu voir que seulement trois coefficients de notre régression logistique étaient significatifs et donc interprétables (**TX_CONSTR**, **TX_COM_GROS** et **TX_SERV_ENT**). Les trois coefficients estimés associés à ces variables entraînent tout un décroissement d'un facteur plus ou moins élevé des chances que le taux de personnes ayant un emploi précaire d'un QPV soit au-dessus de la médiane de ce même taux. La variable entraînant un décroissement le plus élevé est la variable **TX_COM_GROS** car son coefficient est de 0,52 , puis la variable **TX_CONSTR** avec un coefficient d'une valeur de 0,59. Enfin, la variable qui entraîne un décroissement le moins fort est la variable **TX_SERV_ENT** avec un coefficient d'une valeur de 0,73.

Conclusion

Pour conclure, cette étude nous a permis d'explorer de façon globale nos données. D'abord, la mise en place de l'ACP nous a permis de comprendre quelles variables expliquent le plus la variabilité de nos données. Nous avons pu voir que ce sont la part de personne en emploi précaire, la proportion de commerce de détail, de commerce lié aux transports qui vont expliquer la variabilité de nos données.

Ensuite, la mise en place de la classification non supervisée avec les méthodes des K-means et la classification ascendante hiérarchique nous a permis de choisir une répartition de notre échantillon en 4 classes homogènes.

Nous avons enrichi l'analyse avec un arbre à décision. Après avoir discrétisé notre variable d'intérêt, l'arbre nous a permis de comprendre que pour être un QPV avec un taux de personne en emploi précaire en dessous de la médiane, il fallait un taux de commerce de détail et des taux d'établissement industriel moindre.

Enfin, une régression logistique a été faite et nous a permis de dégager des interprétations. Plusieurs taux ont été impliqués dans l'explication de notre variable d'intérêt. Le but ici était de savoir si nos différents taux d'établissement avaient un impact sur le taux de personnes ayant un emploi précaire. Le taux d'établissement de construction, le taux d'établissement de commerce de gros ainsi que le taux d'établissement proposant des services aux entreprises entraînent un décroissement des chances qu'un QPV soit au-dessus de la médiane de la variable **TXT_TOT_EPREC** lorsqu'ils augmentent de 1%. A l'aide de ces résultats, on peut donc supposer que les établissements de construction, de commerce de gros et d'établissement proposant des services aux entreprises entraînent peut être une baisse des personnes ayant un emploi précaire.