02_analyse_donnees

January 16, 2020

1 Caractériser les adhérents de la Médiathèque de Roubaix selon leur lieu d'habitation - partie 2 : Analyse des données

1.1 Transformation du jeu de donnée : on bascule les IRIS en individus

On va modifier le jeu de données pour avoir en ligne les IRIS et en colonnes les différentes variables.

Pour simplifier, on ne conserve que les variables quantitatives donnant des informations sur les pratiques des adhérents et on laisse de côté (même s'il serait intéressant de les utiliser) : - les variables permettant de les qualifier (âge, sexe, ...). - les variables qualitatives relatives aux pratiques.

[1]:
```
import pandas as pd

adh = pd.read_csv("data/adh.csv")
iris1 = pd.pivot_table(adh,
                       values=['nb_venues', 'nb_venues_postes_informatiques',␣
 ↪'nb_venues_prets', 'nb_venues_prets_bus', 'nb_venues_prets_mediatheque',␣
 ↪'nb_venues_salle_etude', 'nb_venues_wifi'],
                       index='Code IRIS de Roubaix',
                       aggfunc='mean',
                       fill_value=0)
iris2 = adh['date_extraction'].groupby(adh['Code IRIS de Roubaix']).count()
iris = pd.merge(iris1, iris2, on='Code IRIS de Roubaix')
iris = iris.rename(columns={"date_extraction" : "nb_inscrits"})
```

On enrichit les données IRIS avec le nb d'habitants, le libellé, le revenu médian (sous dormes absolues et d'indice) :

[2]:
```
iris_lib = pd.read_csv("data/iris_data.csv")
iris = pd.merge(iris, iris_lib, on='Code IRIS de Roubaix')
iris
```

[2]:

	Code IRIS de Roubaix	nb_venues	nb_venues_postes_informatiques	\
0	595120101.0	6.755208	3.739583	
1	595120102.0	6.339888	2.176966	
2	595120201.0	8.912587	6.059441	
3	595120202.0	7.448276	4.563218	
4	595120203.0	6.719745	3.961783	
5	595120301.0	8.170306	4.864629	

6	595120302.0	5.839286	1.380952
7	595120303.0	8.966667	5.309524
8	595120401.0	7.900000	4.857895
9	595120402.0	12.063197	8.762082
10	595120403.0	6.143498	2.780269
11	595120501.0	8.512987	4.474026
12	595120502.0	8.607143	5.200893
13	595120503.0	8.258929	2.888393
14	595120504.0	7.212821	3.538462
15	595120601.0	9.599515	4.939320
16	595120602.0	10.399038	6.264423
17	595120603.0	7.371345	3.391813
18	595120701.0	10.105769	1.564904
19	595120702.0	7.862832	1.938053
20	595120703.0	6.057143	1.214286
21	595120801.0	8.967033	4.835165
22	595120802.0	6.004016	2.361446
23	595120803.0	6.726471	1.291176
24	595120901.0	6.125523	2.895397
25	595120902.0	7.423645	2.748768
26	595120903.0	7.690104	3.484375
27	595121001.0	5.466667	2.528205
28	595121002.0	4.628866	2.005155
29	595121003.0	8.213178	4.527132
30	595121101.0	15.244444	12.133333
31	595121102.0	7.977011	4.390805
32	595121103.0	6.307087	2.755906
33	595121201.0	6.920973	2.346505
34	595121202.0	6.732919	2.726708
35	595121203.0	6.290598	1.692308
36	595121301.0	7.011111	2.788889
37	595121302.0	5.614865	1.527027

	nb_venues_prets	nb_venues_prets_bus	nb_venues_prets_mediatheque	\
0	3.130208	0.161458	2.968750	
1	4.129213	0.365169	3.764045	
2	2.895105	0.048951	2.846154	
3	2.465517	0.390805	2.074713	
4	2.331210	0.331210	2.000000	
5	3.179039	1.165939	2.013100	
6	4.297619	0.482143	3.815476	
7	3.747619	0.295238	3.452381	
8	3.068421	0.447368	2.621053	
9	3.249071	0.215613	3.033457	
10	3.246637	0.022422	3.224215	
11	4.487013	0.084416	4.402597	
12	3.883929	0.241071	3.642857	

13	5.535714	0.062500	5.473214
14	4.184615	0.010256	4.174359
15	4.716019	0.053398	4.662621
16	4.170673	0.019231	4.151442
17	4.122807	0.005848	4.116959
18	8.201923	0.091346	8.110577
19	5.946903	0.628319	5.318584
20	4.945714	0.677143	4.268571
21	3.347985	0.054945	3.293040
22	3.381526	0.337349	3.044177
23	5.388235	0.905882	4.482353
24	3.368201	0.125523	3.242678
25	4.458128	0.561576	3.896552
26	4.328125	0.388021	3.940104
27	2.974359	0.794872	2.179487
28	2.680412	0.500000	2.180412
29	3.356589	0.255814	3.100775
30	3.377778	0.044444	3.333333
31	4.109195	0.965517	3.143678
32	3.228346	0.519685	2.708661
33	5.130699	0.948328	4.182371
34	4.180124	0.552795	3.627329
35	4.602564	0.337607	4.264957
36	4.316667	1.005556	3.311111
37	4.297297	0.216216	4.081081

	nb_venues_salle_etude	nb_venues_wifi	nb_inscrits	\
0	0.057292	0.458333	192	
1	0.171348	0.325843	356	
2	0.195804	0.517483	286	
3	0.293103	0.683908	174	
4	0.515924	0.936306	157	
5	0.174672	0.834061	229	
6	0.148810	0.279762	168	
7	0.509524	0.747619	210	
8	0.947368	0.278947	190	
9	0.479554	1.505576	269	
10	0.452915	0.327354	223	
11	0.662338	0.253247	154	
12	1.151786	0.151786	224	
13	0.191964	0.294643	224	
14	0.197436	0.397436	390	
15	0.332524	0.519417	412	
16	0.163462	0.992788	416	
17	0.230994	0.426901	342	
18	0.663462	0.718750	416	
19	0.022124	0.066372	226	

20	0.200000	0.071429	350
21	0.794872	1.014652	273
22	0.333333	0.634538	249
23	0.144118	0.208824	340
24	0.422594	0.401674	239
25	0.369458	0.522167	203
26	0.450521	0.442708	384
27	0.220513	0.312821	195
28	0.175258	0.278351	194
29	0.058140	1.089147	258
30	1.022222	0.666667	45
31	0.574713	0.609195	174
32	0.259843	0.488189	127
33	0.060790	0.103343	329
34	0.217391	0.111801	161
35	0.153846	0.273504	234
36	0.616667	0.055556	180
37	0.067568	0.121622	148

	revenu_fiscal_median_par_uc	pos	indice	\
0	8472	G3	0,888050314465409	
1	10913	G5	1,14392033542977	
2	7222	G2	0,757023060796646	
3	4105	G1	0,430293501048218	
4	6063	G1	0,635534591194969	
5	9073	G4	0,95104821802935	
6	10237	G5	1,0730607966457	
7	7078	G2	0,741928721174004	
8	6545	G1	0,686058700209644	
9	8003	G3	0,838888888888889	
10	8859	G3	0,928616352201258	
11	6568	G2	0,688469601677149	
12	7338	G2	0,769182389937107	
13	11818	G5	1,23878406708595	
14	9041	G4	0,94769392033543	
15	10640	G5	1,11530398322851	
16	9469	G4	0,992557651991614	
17	9584	G4	1,00461215932914	
18	24340	G6	2,55136268343815	
19	27475	G6	2,87997903563941	
20	16318	G6	1,7104821802935	
21	8381	G3	0,878511530398323	
22	9423	G4	0,987735849056604	
23	14598	G6	1,53018867924528	
24	7920	G3	0,830188679245283	
25	7941	G3	0,832389937106918	
26	8677	G3	0,909538784067086	

27	7277	G2	0,762788259958071
28	6419	G1	0,672851153039832
29	7885	G2	0,826519916142558
30	4648	G1	0,487211740041929
31	6362	G1	0,666876310272537
32	9988	G5	1,04696016771488
33	15408	G6	1,61509433962264
34	15934	G6	1,67023060796646
35	12881	G5	1,35020964360587
36	9747	G4	1,02169811320755
37	9127	G4	0,956708595387841

	Nom de l'IRIS à Roubaix	nb_hab	indice_nb_hab
0	MACKELLERIE	2267	0,911241101367719
1	FRESNOY	3656	1,46956218200281
2	FOSSE AUX CHENES	2832	1,13834794842231
3	ALMA SUD	2284	0,918074404730423
4	ALMA NORD	1806	0,725937992532024
5	ENTREPONT-CARTIGNY	2782	1,11824999735553
6	HUTIN-ORAN	2528	1,0161524059363
7	CUL DE FOUR	3189	1,28184731903911
8	HOMMELET NORD	2182	0,877074584554196
9	HOMMELET CENTRE	2944	1,18336735881189
10	HOMMELET SUD	2208	0,88752551910892
11	EPEULE NORD	1934	0,777388747262976
12	EPEULE CENTRE	2695	1,08327956249934
13	EPEULE SUD	2011	0,808339591905815
14	TRICHON	2684	1,07885801326465
15	ESPERANCE CENTRE	2028	0,815172895268519
16	NATIONS UNIES	2645	1,06318161143256
17	ANSEELE	3324	1,33611178691941
18	BARBIEUX-VAUBAN	3722	1,49609147741096
19	BARBIEUX SUD	2088	0,839290436548653
20	EDOUART VAILLANT	2979	1,19743592455864
21	MOULIN NORD	2549	1,02459354538435
22	MOULIN SUD	2385	0,958672265885315
23	POTENNERIE	2957	1,18859282608926
24	SAINTE-ELISABETH CENTRE	2477	0,995652495848186
25	SAINTE-ELISABETH SUD	1903	0,764928017601574
26	SAINTE-ELISABETH NORD	2818	1,13272052212361
27	PILE EST	2207	0,887123560087585
28	PILE CENTRE	2616	1,05152479981383
29	PILE NORD	2142	0,860996223700773
30	TROIS PONTS NORD	1040	0,418037382188984
31	TROIS PONTS SUD	2264	0,910035224303712
32	SARTEL	1796	0,721918402318669
33	LINNE CHEMIN NEUF	3041	1,22235738388144

34	JUSTICE	2109	0,8477315759967
35	FRATERNITE	2679	1,07684821815797
36	NOUVEAU ROUBAIX	2302	0,925309667114463
37	HAUTS CHAMPS	2464	0,990427028570824

1.2 Première analyse : répartition des adhérents par IRIS

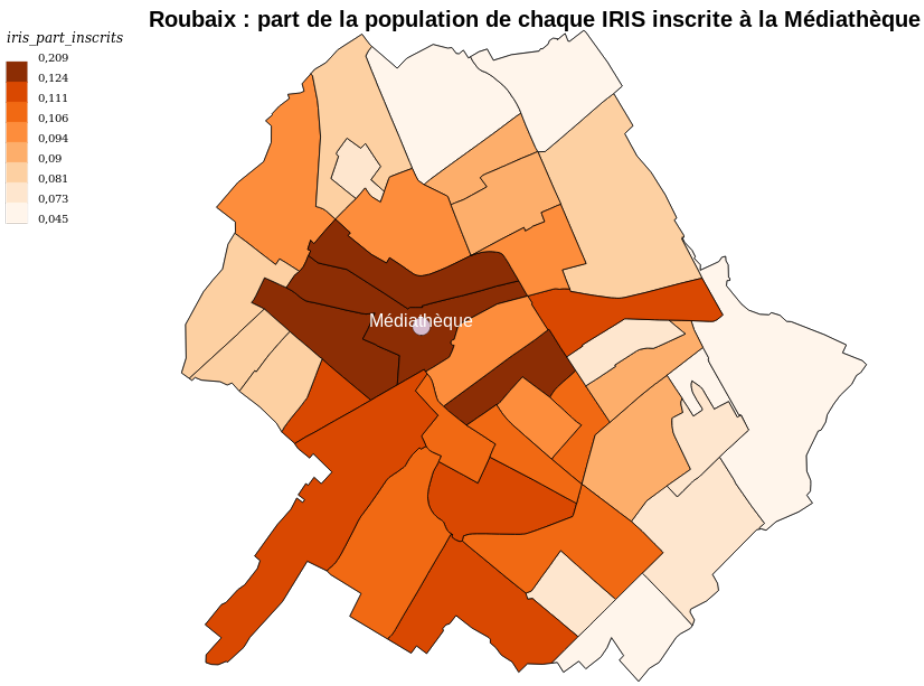
On peut maintenant regarder comment se répartissent les inscrits à la Médiathèque au sein des différents IRIS. On calcule la part d'inscrits par rapport au nombre d'habitants de chaque IRIS.

[4]:
```
iris['iris_part_inscrits'] = iris['nb_inscrits'] / iris['nb_hab']
iris[['Code IRIS de Roubaix', 'Nom de l'IRIS à Roubaix', 'iris_part_inscrits']].
↪to_csv("data/iris_carte_inscrits.csv", header=True, index=False)
```

À partir du fichier csv obtenu et à un fond de carte, on réalise une carte avec l'outil Magrit.

[5]:
```
from IPython.display import Image
Image(filename='data/iris_carte_inscrits.png')
```

[5]:



On constate une double tendance, à savoir le taux d'inscription a tendance à diminuer plus : - on s'éloigne de la Médiathèque, - les quartiers sont pauvres

On va maintenant se demander comment regrouper les quartiers en fonction des pratiques de leurs habitants. On va donc recourir à une ACP.

1.3 Réalisation d'une ACP pour regrouper les quartiers

On exporte le jeu de données en csv pour pouvoir réaliser l'ACP avec R.

[ ]:
```
iris.to_csv("data/iris.csv", header=True, index=False)
```

Même si l'ACP est effectuée avec R, par commodité, on dépose le code et les résultats dans ce notebook Python.

Pour cette ACP, on ne conserve que les variables quantitatives suivantes : - `nb_venues_postes_informatiques', - `nb_venues_prets', - `nb_venues_prets_bus', - `nb_venues_prets_mediatheque', - `nb_venues_salle_etude', - `nb_venues_wifi'.

On exclut notamment les variables `nb_venues' et `nb_inscrits', en présupposant qu'elles synthétisent les autres.

On ajoute par ailleurs une variable qualitative `pos', qui regroupe les IRIS en 6 groupes selon le niveau de revenus de leurs habitants. On essaiera de voir si cette variable permet de regrouper les résultats.

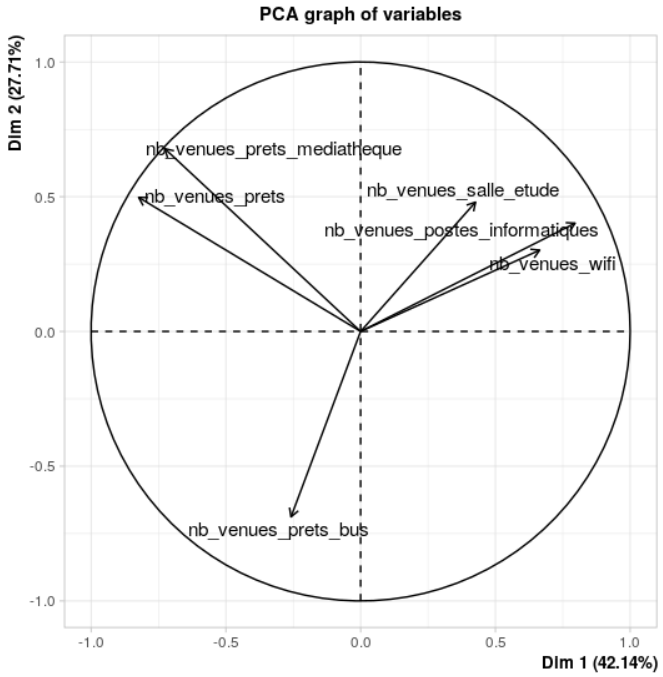
```
library(FactoMineR)
empr <-read.table("data/iris.csv",
                  header=TRUE, sep=',', row.names=13)
summary(empr)
res <-PCA(empr[,c(3,4,5,6,7,8,11)], quali.sup = 7)
barplot(res$eig[,1], main="Eigenvalues", names.arg=1:nrow(res$eig))

library(factoextra)
fviz_pca_ind(res,
             habillage=empr$pos, repel=TRUE)
```

---

[7]:
```
Image(filename='data/pca_variables.png')
```

[7]:

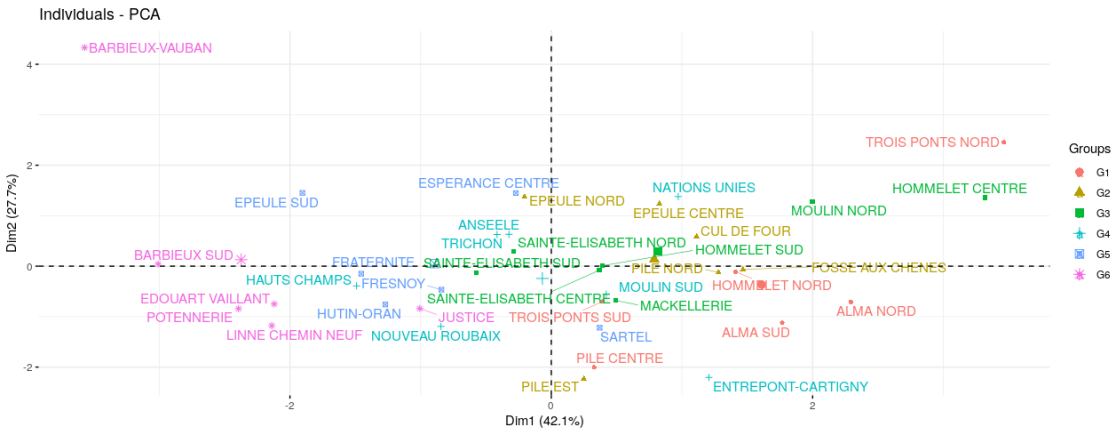


On constate que les deux premiers axes permettent d'expliquer une grande partie des phénomènes observés (presque 70 % de variance cumulée).

On constate d'autre part que 3 types d'usages ont tendance à s'opposer : - prêts - usages de l'informatique ou salle étude - bibliobus.

[8]:
```
Image(filename='data/pca_individus.png')
```

[8]:



L'ACP tend à montrer que : - il y a une forme de corrélation entre le niveau de revenus et l'intensité des usages (à gauche, utilisation forte, à droite, utilisation faible) : information nouvelle par rapport à carte précédente, qui

ne s'attache qu'au fait de s'inscrire ou non. - on voit bien deux quartiers aux niveaux d'usage un peu hors norme (Barbieux-Vauban et Trois Ponts Nord) - pour le reste, il semble que de manière verticale, les IRIS placé en bas sont plutôt utilisateurs du biblbiobus tandis que les IRIS placéss en haut sont plutôt consommateurs desservics informatiques.