DATA MINING

TD2 : Analyse en Composantes Principales

Exercice 1 :

Le tableau suivant fournit la structure du bilan d'un groupe pétrolier de 1969 à 1984 :

Année	NET	INT	SUB	LMT	DCT	IMM	EXP	VRD
1969	17.93	3.96	0.88	7.38	19.86	25.45	5.34	19.21
1970	16.21	3.93	0.94	9.82	19.11	26.58	5.01	18.40
1971	19.01	3.56	1.91	9.43	17.87	25.94	5.40	16.88
1972	18.05	3.33	1.73	9.72	18.83	26.05	5.08	17.21
1973	16.56	3.10	2.14	9.39	20.36	23.95	6.19	18.31
1974	13.09	2.64	2.44	8.10	25.05	19.48	11.61	17.59
1975	13.43	2.42	2.45	10.83	22.07	22.13	11.17	15.49
1976	9.83	2.46	1.79	11.81	24.10	22.39	11.31	16.30
1977	9.46	2.33	2.30	11.46	24.45	23.07	11.16	15.77
1978	10.93	2.95	2.25	10.72	23.16	24.17	9.64	16.20
1979	13.02	3.74	2.21	7.99	23.04	19.53	12.60	17.87
1980	13.43	3.60	2.29	7.09	23.59	17.61	16.67	15.72
1981	13.37	3.35	2.58	6.76	23.94	18.04	15.42	16.54
1982	11.75	2.74	3.11	7.37	25.04	18.11	14.71	17.18
1983	12.59	3.05	3.85	7.12	23.40	19.17	11.86	18.97
1984	13.00	3.00	4.00	7.00	24.00	20.00	12.00	17.00

Les postes de bilan sont les suivants :

NET : Situation nette ; représente l'ensemble des capitaux propres de l'entreprise.

INT : Intérêts ; représente l'ensemble des frais financiers supportés par l'entreprise.

SUB : Subventions ; représente le montant total des subventions accordées par l'Etat.

LMT : Dettes à long et moyen terme.

DCT : Dettes à court terme.

IMM : Immobilisations ; représente l'ensemble des terrains et du matériel de l'entreprise.

EXP : Valeurs d'exploitation.

VRD : Valeurs réalisables et disponibles ; ensemble des créances à court terme de l'entreprise.

Les données ont été ventilées en pourcentage par année, la somme des éléments d'une même ligne vaut 100, de manière à éviter les effets dus à l'inflation. On propose d'appliquer une Analyse en Composantes Principales (ACP) afin d'analyser l'évolution de la structure de bilan sur 15 ans. Les résultats de l'ACP sont présentés dans les tableaux et les figures ci-dessous :

Tableau1 :

Eigenvalues of the Correlation Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	4.47037150	2.35552573	0.5588	0.5588
2	2.11484576	1.43418677	0.2644	0.8232
3	0.68065899	0.17991239	0.0851	0.9082
4	0.50074660	0.34116829	0.0626	0.9708
5	0.15957831	0.09542998	0.0199	0.9908
6	0.06414833	0.05449844	0.0080	0.9988
7	0.00964990	0.00964928	0.0012	1.0000
8	0.00000062	0.0000	1.0000	

Tableau 3 :

Coordonnees des variables sur les axes

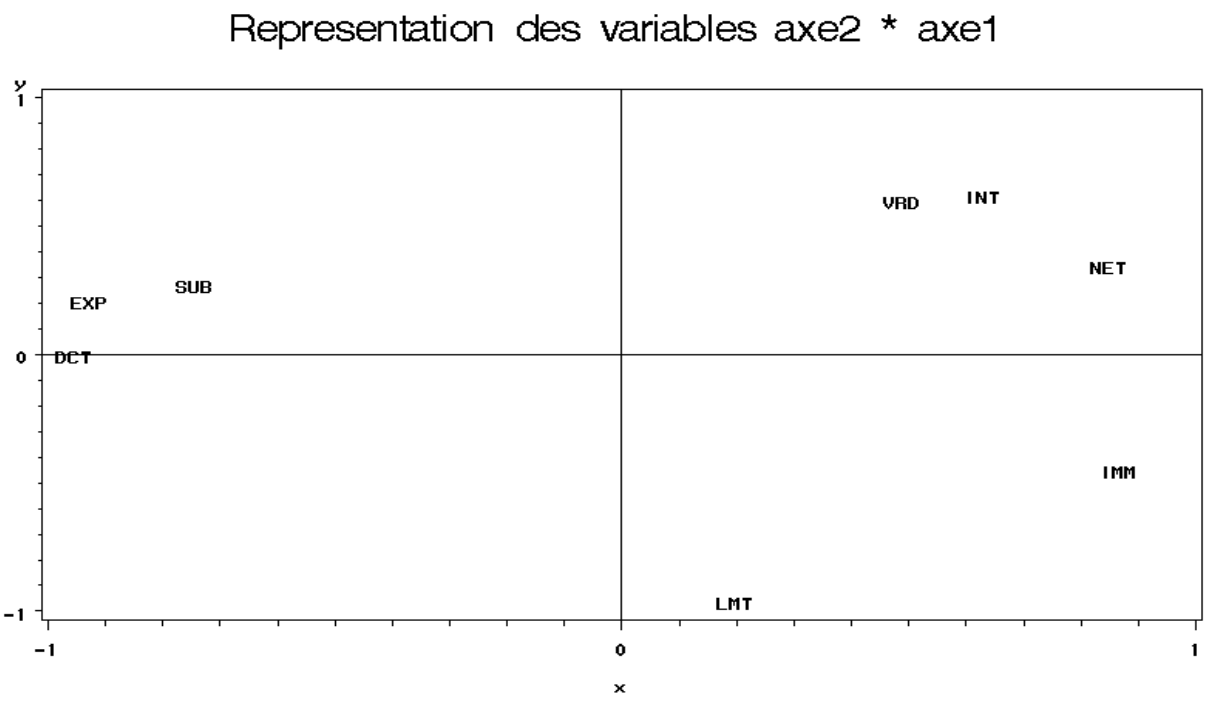
Pearson Correlation Coefficients, N = 16

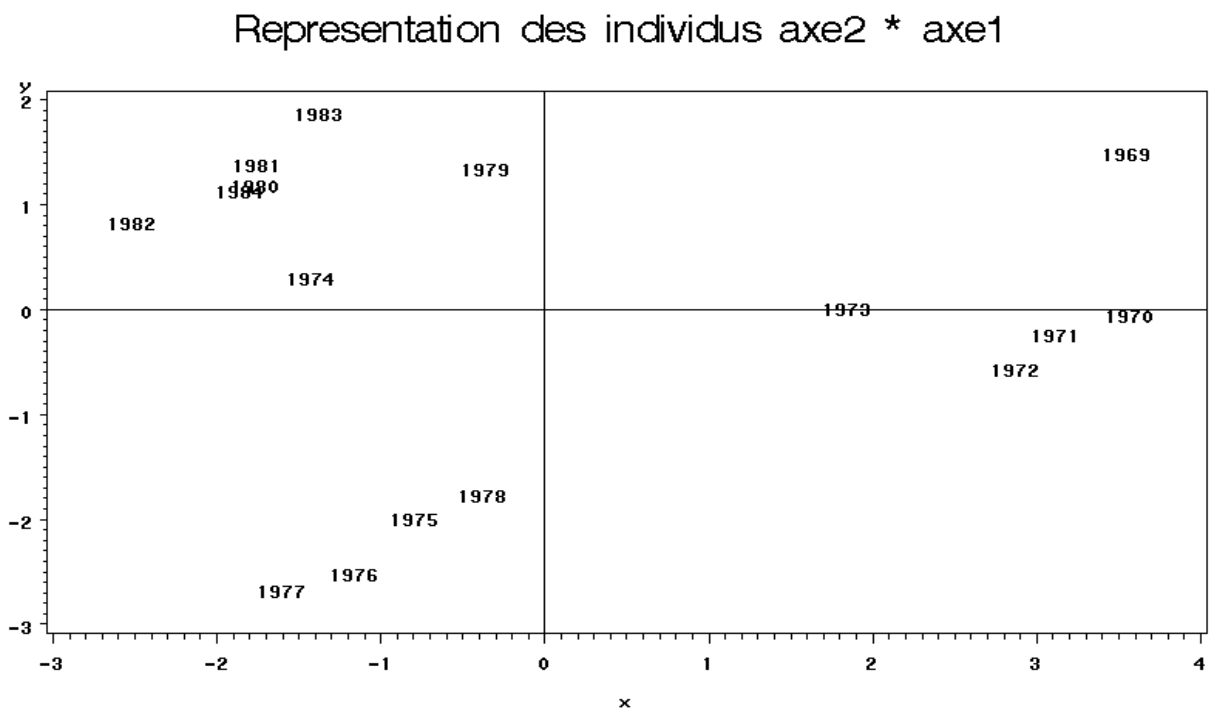
	Prin1	Prin2
NET	0.85014	0.34678
INT	0.62963	0.62173
SUB	-0.74214	0.27580
LMT	0.20017	-0.96163
DCT	-0.95386	0.00168
IMM	0.86787	-0.44767
EXP	-0.92571	0.20985
VRD	0.49025	0.60233

Tableau 2 :

Coordonnees et qualite de representation des individus sur les axes

annee	Prin1	Prin2	cos2_prin1	cos2_prin2
1969	3.55662	1.50535	0.78441	0.14052
1970	3.57546	-0.04273	0.93110	0.00013
1971	3.12027	-0.21808	0.83031	0.00406
1972	2.87553	-0.54758	0.89332	0.03239
1973	1.84936	0.02352	0.75517	0.00012
1974	-1.42432	0.32194	0.57269	0.02926
1975	-0.79476	-1.97215	0.11144	0.68621
1976	-1.16070	-2.50400	0.15851	0.73770
1977	-1.59726	-2.65758	0.25931	0.71786
1978	-0.37918	-1.74803	0.03739	0.79463
1979	-0.36150	1.35612	0.04004	0.56350
1980	-1.75965	1.20307	0.34868	0.16299
1981	-1.75001	1.40025	0.49152	0.31468
1982	-2.51840	0.84115	0.87166	0.09724
1983	-1.37918	1.88579	0.21797	0.40752
1984	-1.85228	1.15298	0.50000	0.19373





1. Expliquer les objectifs de l'Analyse en Composantes Principales (ACP) en Informatique décisionnelle.

2. Etude du tableau des valeurs propres.

2.1. A quoi correspond la somme des valeurs propres ?

2.2. On choisit de n'étudier que les deux premières composantes principales. Justifier ce choix en analysant le tableau 1 des valeurs propres (Eigenvalues).

2.3. Calculer le pourcentage d'information quantifié par les deux premières composantes principales sélectionnées.

3. Analyse des résultats de l'ACP

3.1. Sélectionner les individus (les années) qui sont bien représentés sur le plan factoriel en analysant les qualités de leurs représentations (cos2) dans le tableau 2.

3.2. Sélectionner les variables corrélées avec les premières composantes principales à partir du tableau 3.

3.3. Commenter les positions des années bien représentées sur le plan factoriel par rapport aux variables corrélées avec les deux premières composantes principales.

Exercice 2 :

On a rassemblé les résultats de 15 enfants de 10 ans à 6 subtests du WISC (scores 0 à 5). Les variables observées sont : CUB (Cubes de Kohs), PUZ (Assemblage d'objets), CAL (Calcul mental), MEM (Mémoire immédiate des chiffres), COM (Compréhension de phrases), VOC (Vocabulaire).

On traite ces données par une analyse en composantes principales normée. Les principaux résultats de cette ACP sont indiqués ci-dessous.

I. Etude du tableau des valeurs propres :

	Val. propr	% Total variance	Cumul Val. propr	Cumul %
1	3,2581	54,3020	3,2581	**54,3020**
2	1,8372	30,6194	5,0953	**84,9214**
3	0,4430	7,3831	5,5383	**92,3044**
4	0,2538	4,2292	5,7920	**96,5337**
5	0,1679	2,7990	5,9600	**99,3327**
6	0,0400	0,6673	6,0000	**100,0000**

Valeurs propres & statistiques associées

1. A quoi correspond la **somme** des valeurs propres ?
2. **On choisit de n'étudier que les deux premières composantes principales**. Justifier ce choix en analysant le tableau des valeurs propres.
3. **II.** Etude des qualités de représentation dans le premier plan principal

	Score Fact. 1	Score Fact. 2	Contribution Fact.1	Contribution Fact.2	Cos² Fact.1	Cos² Fact. 2
I1	-2,5616	3,0568	13,43	33,91	0,4078	0,5807
I2	-0,9661	0,9370	1,91	3,19	0,3907	0,3676
I3	0,6765	-0,6624	0,94	1,59	0,4446	0,4263
I4	-2,7969	-1,4636	16,01	7,77	0,7160	0,1961
I5	-1,8423	0,1211	6,95	0,05	0,8142	0,0035
I6	1,8891	0,1350	7,30	0,07	0,8426	0,0043
I7	-2,3396	-1,5487	11,20	8,70	0,6028	0,2641
I8	0,7275	-2,2054	1,08	17,65	0,0816	0,7499
I9	2,8400	0,5423	16,50	1,07	0,8745	0,0319
I10	2,1733	0,6117	9,66	1,36	0,7433	0,0589
I11	1,2940	2,0373	3,43	15,06	0,2256	0,5592

I12	-0,9947	0,8181	2,02	2,43	0,3120	0,2110
I13	-0,6099	-0,8730	0,76	2,77	0,1949	0,3994
I14	2,0150	-0,9470	8,31	3,25	0,7548	0,1667
I15	0,4957	-0,5591	0,50	1,13	0,1151	0,1464

Scores, contributions et qualités de représentation des individus

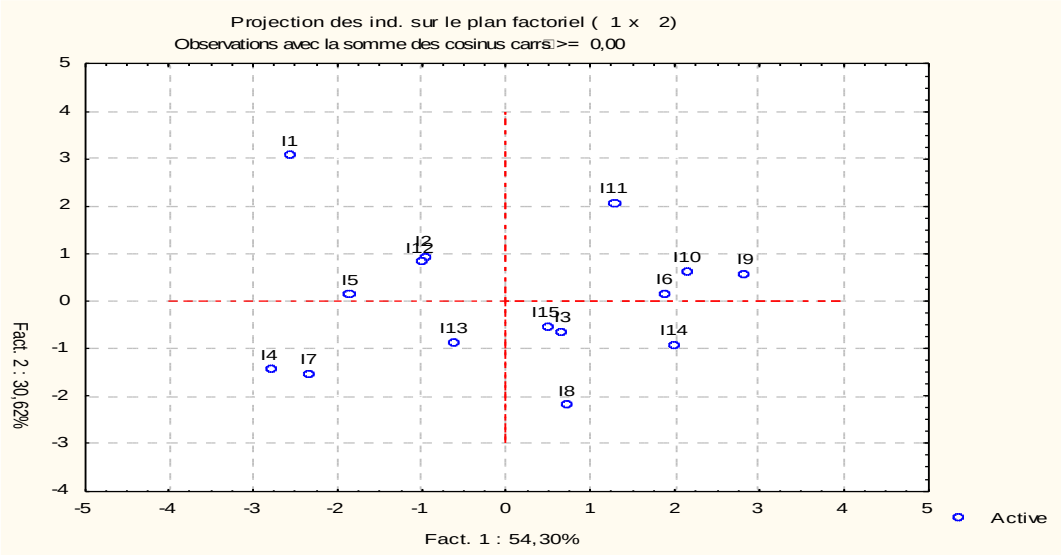
	Saturation Fact. 1	Saturation Fact. 2	Contribution Fact.1	Contribution Fact.2	Cos² Fact.1	Cos² Fact.1&2
CUB	-0,8970	0,2018	0,25	0,02	0,8046	0,8453
PUZ	-0,8652	0,2883	0,23	0,05	0,7485	0,8316
CAL	-0,9458	0,0390	0,27	0,00	0,8945	0,8960
MEM	0,4449	-0,7861	0,06	0,34	0,1980	0,8160
COM	-0,5382	-0,7627	0,09	0,32	0,2897	0,8714
VOC	-0,5683	-0,7156	0,10	0,28	0,3229	0,8350

Saturations, contributions et qualités de représentation des variables

3. Comment quantifie-t-on la qualité de représentation des individus par le plan factoriel ?

4. Quel est l'individu le moins représenté par le premier plan principal ? Quel est l'individu le mieux représenté ?

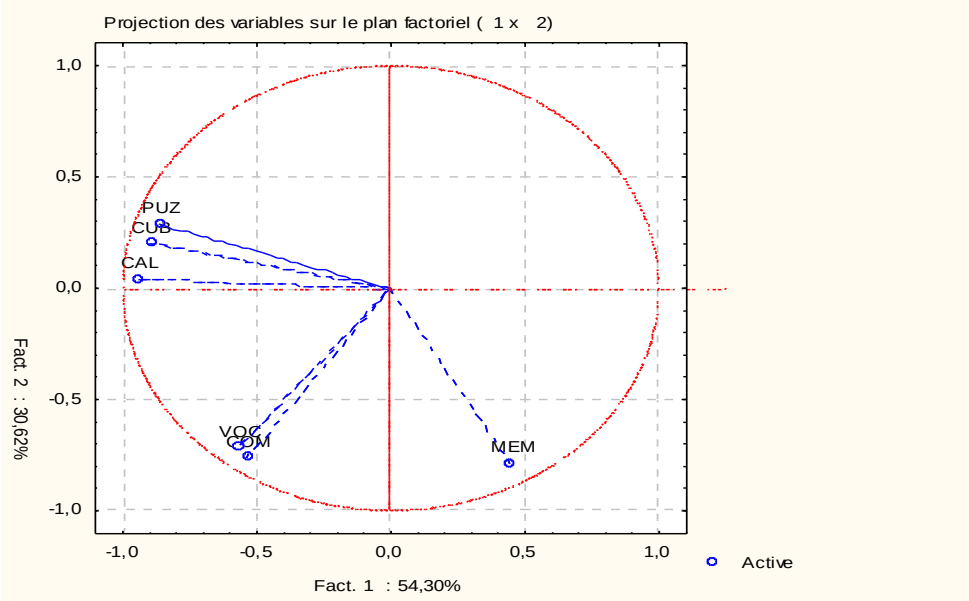
III. Etude du nuage des individus.



5. Quels sont les individus dont la contribution à la formation de la première composante principale est supérieure à la moyenne ? Pour chacun d'eux, préciser le signe de la coordonnée correspondante. Caractériser cet axe en termes d'opposition entre individus.

6. Même question pour la deuxième composante principale.

IV. Etude du nuage des variables



7. La représentation graphique des variables montre qu'elles sont toutes très bien représentées dans le plan (CP1, CP2). Justifier cette affirmation.

8. Quelles sont les variables qui sont corrélées positivement avec le premier facteur principal ? Quelles sont celles qui sont corrélées négativement ? Comment peut-on caractériser cet axe par rapport aux variables de départ ? (1.5)

9. Quelles sont les variables qui ont joué un rôle dominant dans la formation du deuxième axe

Exercice 3 :

On donne sur l'annexe 1 un échantillon de données pour quelques villes. Ces dernières sont décrites par :

Annexe 1 : Données PCA

avganncount	avgdeathsperyear	target_deathrate	incidencerate	medincome	popest2015	povertypercent	studypercap	medianage	medianagemale	medianagefemale	percentmarried	pctnohs18_24	pcths18_24
1397	469	164,9	489,8	61898	260131	11,2	499,7482038	39,3	36,9	41,7	52,5	11,5	39,5
173	70	161,3	411,6	48127	43269	18,6	23,11123437	33	32,2	33,7	44,5	6,1	22,4
102	50	174,7	349,7	49348	21026	14,6	47,56016361	45	44	45,8	54,2	24	36,6
427	202	194,8	430,4	44243	75882	17,1	342,6372526	42,8	42,2	43,4	52,7	20,2	41,2
57	26	144,4	350,1	49955	10321	12,5	0	48,3	47,8	48,9	57,8	14,9	43
428	152	176	505,4	52313	61023	15,6	180,259902	45,4	43,5	48	50,4	29,9	35,1
250	97	175,9	461,8	37782	41516	23,2	0	42,6	42,2	43,5	54,1	26,1	41,4
146	71	183,6	404	40189	20848	17,8	0	51,7	50,8	52,5	52,7	27,3	33,9
88	36	190,5	459,4	42579	13088	22,3	0	49,3	48,4	49,8	55,9	34,7	39,4
4025	1380	177,8	510,9	60397	843954	13,1	427,7484318	35,8	34,7	37	50	15,6	36,3
113	36	121,4	413,3	54721	16252	12,7	0	54,4	54	54,6	56,8	17,7	32,4
740	269	172,7	499,3	51395	121846	15,7	837,1222691	41	40	42,2	53,6	25,5	33,8
55	26	188,3	398,9	52673	11339	12,6	0	45,2	44,9	45,5	54,4	20	43,8
3438	1118	165,3	493,4	71890	772501	9,9	138,5111476	37,6	36,6	38,7	52,1	15,4	33,3
2265	901	171	440,7	50083	490945	16,3	462,3735856	37,2	35,7	38,7	49,4	10,9	29,3
251	106	174,2	423,8	43823	43791	19,3	0	46,2	45,6	46,8	57,2	25,1	35,3
1390	483	169,9	495,9	61653	269536	11,9	207,7644545	38,5	37,1	39,9	52,4	14,8	31,7
32	12	153,8	463,2	51022	4042	13,9	0	52,1	51,5	53,1	53,5	37,2	20,8
305	120	162,8	442,5	49819	60338	15,7	464,0525042	36,9	35,5	39,1	46,5	13,7	29,2
1081	367	163,3	490,9	53733	212284	15,7	249,6655424	36,5	35,6	37,5	47,5	9,4	22,1
134	50	140,8	390,5	41837	48177	28,4	0	24,2	24,4	23,9	36,1	1,2	16,4
958	403	169,4	411	44342	248830	20,5	301,4106016	32,5	31,6	33,6	48,6	28,7	39,3
94	41	189,7	445,2	35615	16704	21,5	0	41,5	40,9	42,1	52	9,8	36,1
499	215	206,1	463,1	56737	111901	13,2	89,3647063	38,4	38	38,6	51,8	17	40,8

	avganncount	avgdeathsperyear	target_deathrate	incidencerate	medincome	popest2015	povertypercent	studypercap	medianage
1	1397	469	164.9	489.8	61898	260131	11.2	499.74820	39.3
2	173	70	161.3	411.6	48127	43269	18.6	23.11123	33.0
3	102	50	174.7	349.7	49348	21026	14.6	47.56016	45.0
4	427	202	194.8	430.4	44243	75882	17.1	342.63725	42.8

	medianagemale	medianagefemale	percentmarried	pctnohs18_24	pcths18_24
1	36.9	41.7	52.5	11.5	39.5
2	32.2	33.7	44.5	6.1	22.4
3	44.0	45.8	54.2	24.0	36.6
4	42.2	43.4	52.7	20.2	41.2

Figure 1: statistiques descriptives

Annexe 2 : Résultats PCA

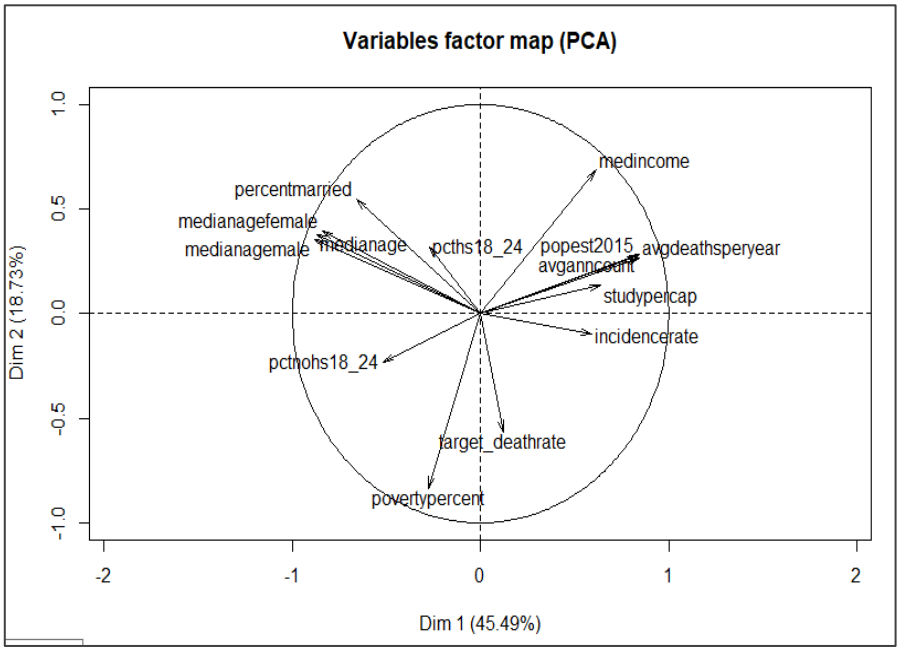


Figure 2: Cercle de corrélation

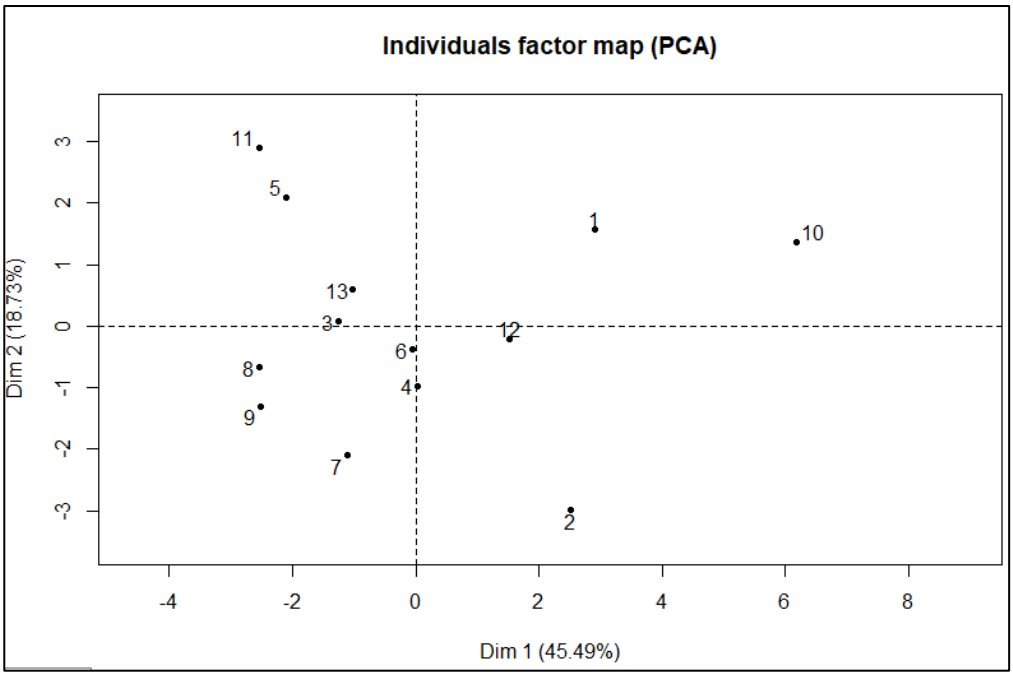


Figure 3: carte des individus

	eigenvalue	percentage of variance	cumulative percentage of variance
comp 1	6.368503e+00	4.548930e+01	45.48930
comp 2	2.622454e+00	1.873181e+01	64.22112
comp 3	2.253734e+00	1.609810e+01	80.31922
comp 4	1.019427e+00	7.281619e+00	87.60084
comp 5	8.307547e-01	5.933962e+00	93.53480
comp 6	4.448084e-01	3.177203e+00	96.71200
comp 7	2.859710e-01	2.042650e+00	98.75465
comp 8	1.223788e-01	8.741339e-01	99.62879
comp 9	4.098827e-02	2.927733e-01	99.92156
comp 10	9.236089e-03	6.597206e-02	99.98753
comp 11	1.709144e-03	1.220817e-02	99.99974
comp 12	3.643788e-05	2.602706e-04	100.00000

Figure 4: valeurs propres, variance

	Dim.1	Dim.2	Dim.3	Dim.4	Dim.5
avganncount	0.8437465	0.28134063	0.3255090	0.04747150	0.31022852
avgdeathsperyear	0.8444693	0.26536159	0.3348084	0.03649259	0.30839640
target_deathrate	0.1204549	-0.56681584	0.6112364	-0.35548804	-0.04154321
incidencerate	0.5882692	-0.09866841	0.5339943	0.44534413	-0.21297600
medincome	0.6131181	0.68643531	-0.1519976	0.04819867	-0.17388042
popest2015	0.8373729	0.27509000	0.3036888	0.02828112	0.35316429
povertypercent	-0.2781171	-0.83387133	0.2583468	0.15632484	0.18968014
studypercap	0.6354490	0.13614854	0.3151271	0.14983004	-0.59611933
medianage	-0.8685415	0.37765320	0.1770190	0.21259126	0.10080660
medianagemale	-0.8834139	0.35544799	0.1577312	0.18281515	0.12948946
medianagefemale	-0.8423059	0.39527423	0.2086886	0.24857987	0.05925756
percentmarried	-0.6599164	0.54818918	0.3411178	-0.17008025	-0.08390842
pctnohs18_24	-0.5197228	-0.22990574	0.7316509	0.24476804	-0.04938326
pcths18_24	-0.2734582	0.32023150	0.5866027	-0.64178020	-0.07812471

Figure 5: coordonnées des variables sur les axes

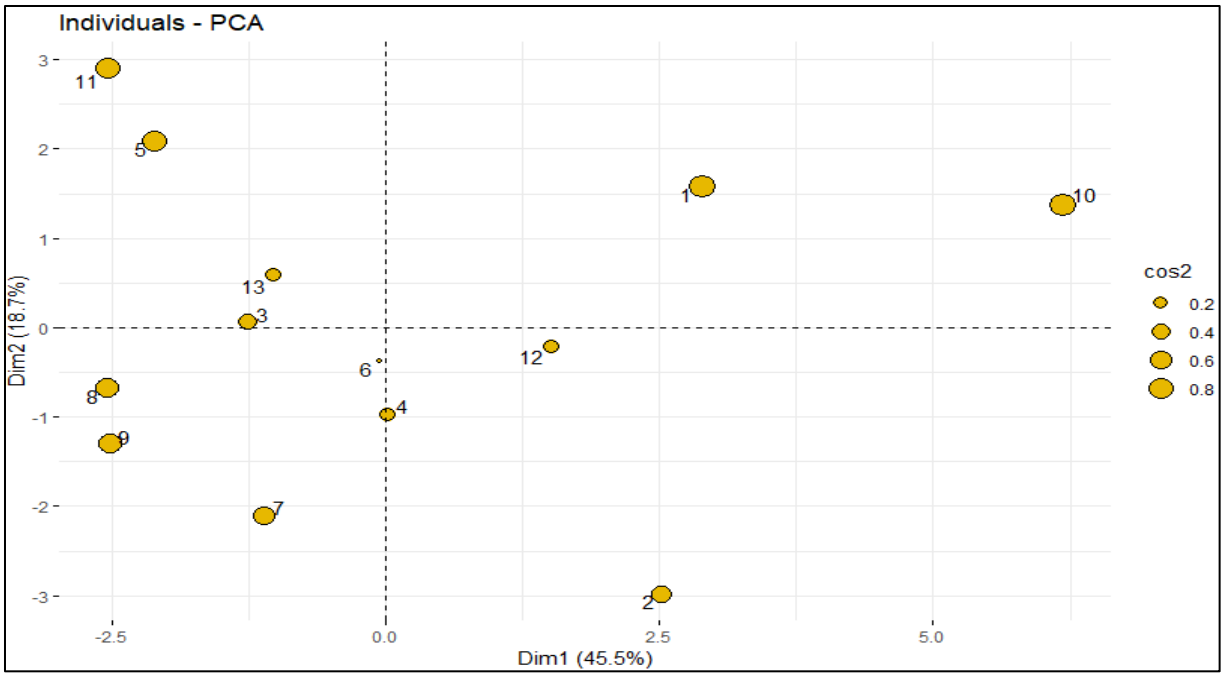


Figure 6: carte des individus selon le cos2

Le nombre de personnes atteintes de cancer le nombre de mortalités du au cancer le revenu moyen de la ville le nombre de population le pourcentage de pauvreté l'âge moyen de la population l'âge moyen des femmes le pourcentage des mariés le pourcentage de la population atteinte du cancer dont l'âge est entre 18 et 24, etc.

Les statistiques descriptives sont données par la figure 1 de l'annexe1. On se propose de réaliser une analyse en composantes principales afin de comprendre les données.

Les résultats sont illustrés dans les graphes de l'annexe

1- Définir l'analyse en composantes principales et préciser son utilité. [1 pt]

2- En se référant au tableau des valeurs propres donné par la figure 4, comment choisir les <axes factoriels les plus adéquats. Quel est le critère utilisé ? [1 pt]

3- Pour des raisons de visualisation, on a choisi de représenter nos variables sur les axes Dim1 et Dim2. Comment jugez-vous ce repère de projection. Interpréter les corrélations variables/ variables et variables /dimensions. [1 pt]

4- En se référant à la carte des individus représentée par la figure 3 préciser la liste des individus mal représentés sur les axes Dim1 et Dim 2. [0.5 pt]

5- On donne la figure 6, carte des individus selon le cos2. En se référant à cette dernière, vérifier les résultats obtenus dans la question précédente. Expliquer et réordonner les individus par ordre décroissant selon leur contribution dans l'axe Dim1 et Dim2. [1 pt]