



COURS DE FOUILLE DE DONNÉES

M1, P-20- IFI 2016

Enseignante: NGUYỄN Thị Minh Huyền

TP5 – G2

Clustering

Données Utilisées: Census Income Data Set

Étudiant: Ginel Dorleon

Gervais Fotsing Sikadie S.

CLUSTERING

Le **partitionnement de données** (ou ***data clustering*** en anglais) est une des méthodes d'analyse des données. Elle vise à diviser un ensemble de données en différents « paquets » homogènes, en ce sens que les données de chaque sous-ensemble partagent des caractéristiques communes, qui correspondent le plus souvent à des critères de proximité (similarité informatique) que l'on définit en introduisant des mesures et classes de distance entre objets.

RAPPEL/DESCRIPTION SUR NOS DONNÉES: Census Income Data Set [2]

Les données que nous utilisons pour cette 2^e partie ont été extraites de la base de données d'un bureau de recensement aux USA. 32561 instances avec 15 attributs . L'objectif est d'étudier un ensemble de paramètre et d'établir une tâche de prévision afin de déterminer le profil des personnes qui gagnent plus de 50K par année.

NOS VARIABLES

Age: Variable continue représentant l'âge des personnes.

Workclass: Variable discrète prenant les valeurs suivantes : Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked

Fnlwgt: Final weight, une variable continue calculée à partir des données d'origine et du sexe de la personne.

Education: Variable discrète déterminant le niveau d'éducation de la personne, elle prend les valeurs suivantes : Licence, Certains collège, 11e, HS-grad, Prof-école, Assoc-ACDM, Assoc-voc, 9e, septième-huitième, 12e, Masters, 1ere-4ème, 10e, Doctorat, 5ème-6ème, Maternelle.

education-num: continu.

statut matrimonial: Marié-civ-conjoint, Divorcé, jamais marié, Séparé, Veuf, Marié-conjoint absent, Marié-AF-conjoint.

Profession: Tech-support, Craft-réparation, Autre service, ventes, Exec-gestion, Prof spécialité, Handlers-nettoyants, machine-op-inspct, Adm-clérical, agriculture-pêche, Transport-mobile, Priv-ménage serv, protection-serv, armées-Forces.

Relationship : Variable discrète représentant les types de relations des individus. Elle prend les valeurs suivantes : Femme, propre enfant, Mari, Non-in-famille, Autre-parent, Unmarried.

Race: Variable discrète représentant la race des individus. Elle prend les valeurs Noir Blanc, Asie-Pac-Islander, Amer-Indian-Eskimo, Autre.

Sexe: Variable discrète représentant le sexe de l'individu, 2 valeurs : Femme, Homme.

Capital-gain: Variable continue représentant le capital gagné par l'individu.

Capital-loss: Variable continue représentant le capital perdu par l'individu.

Hours-per-week : Variable continue représentant le nombre d'heur que l'individu travail

Pays: Variable discrète représentant le pays d'origine de l'individu.

Salary: <50K, >=50 K

STATISTIQUE SUR LES VARIABLES

Dans la plupart des études sur les données , le nombre de sujets est souvent trop important pour que l'on puisse présenter les données réelles de chaque individu. C'est pourquoi, il est nécessaire de trouver un moyen qui donne le maximum d'informations possible sous le format le plus utile. Une manière courante de présentation de données est la représentation graphique ou le tableau. Les tableaux sont commodes pour présenter l'information relative aux données individuelles et les graphiques pour donner un profil général des observations. Toutefois, il est également utile de donner un résumé chiffré. Pour les variables qualitative ou en catégorie (niveau d'étude, sexe, absence-présence d'une maladie, niveau d'éducation, etc.), la mesure la plus instructive est la proportion d'individus entrant dans chaque catégorie. Les variables quantitatives (poids, taille, âge, salaire, etc.) nécessitent quant à elle deux types de mesure pour avoir une idée complète de la distribution des observations : la mesure de la *position centrale* des observations et la mesure de leur *dispersion*, c'est-à-dire la mesure de la répartition des observations autour de cette position centrale. Ainsi, nous allons faire procéder à l'étude qualitative et quantitative de deux des variables de notre jeu de données.

Étude d'une variable qualitative- Education

Généralement, en statistique, une variable qualitative, on dit aussi catégorielle, est une variable pour laquelle la valeur mesurée sur chaque individu ne représente pas une quantité. Les différentes valeurs que peut prendre cette variable sont appelées les *catégories*, *modalités* ou *niveaux*.

Ainsi, parmi l'ensemble des variables qualitatives de notre jeu de données, nous avons choisi d'étudier la variable *Education*. En effectuant la statistique primaire de cette variable, on peut constater que la plus grande valeur valeur Hs-grad signifiant le degré en High School. On constate que le mode est cette valeur, Hs-grad. Le nombre d'individu ayant ce niveau est de 10501 et le pourcentage d'individus ayant ce niveau est de 32.25 % . Voir capture ci-dessous

Univariate discrete stat 1				
Parameters				
Attributes : 1				
Examples : 32561				
Results				
Attribute	Gini	Distribution		
		Values	Count	Percent
EDUCATION	0,8096	Bachelors	5355	16,45 %
		Hs-grad	10501	32,25 %
		11th	1175	3,61 %
		Masters	1723	5,29 %
		9th	514	1,58 %
		Some-college	7291	22,39 %
		Assoc-acdm	1067	3,28 %
		Assoc-voc	1382	4,24 %
		7th-8th	646	1,98 %
		Doctorate	413	1,27 %
		Prof-school	576	1,77 %
		5th-6th	333	1,02 %
		10th	933	2,87 %
		1st-4th	168	0,52 %

Étude d'une variable quantitative – L'Age

Une variable est *quantitative* si elle reflète une notion de grandeur, c'est-à-dire si les valeurs qu'elle peut prendre sont des nombres. Une grandeur quantitative est souvent exprimée avec une unité de mesure qui sert de référence

Parmi l'ensemble des variables quantitatives de notre jeu de données, nous avons choisi d'étudier la variable quantitative Age. En effectuant l'analyse statistique primaire de cette variable, on constate que l'âge minimum est de 17 ans, le max est de 90 ans et l'âge moyenne est 38,5816. Voir capture ci-dessous.

Univariate continuous stat 1					
Parameters					
Attributes : 1					
Examples : 32561					
Results					
Attribute	Min	Max	Average	Std-dev	Std-dev/avg
AGE	17	90	38,5816	13,6404	0,3535

More Univariate cont stat 1					
Parameters					
Attributes : 1					
Examples : 32561					
Results					
Attribute	Stats		Histogram		
	Statistics		Values	Count	Percent
AGE	Average	38,5816	x_<_24,3000	5570	17,11%
	Median	37,0000	24,3000_=<_x_<_31,6000	5890	18,09%
	Std dev. [Coef of variation]	13,6404 [0,3535]	31,6000_=<_x_<_38,9000	6048	18,57%
	MAD [MAD/STDDEV]	11,1892 [0,8203]	38,9000_=<_x_<_46,2000	6163	18,93%
	Min * Max [Full range]	17,00 * 90,00 [73,00]	46,2000_=<_x_<_53,5000	3967	12,18%
	1st * 3rd quartile [Range]	28,00 * 48,00 [20,00]	53,5000_=<_x_<_60,8000	2591	7,96%
	Skewness (std-dev)	0,5587 (0,0136)	60,8000_=<_x_<_68,1000	1595	4,90%
	Kurtosis (std-dev)	-0,1661 (0,0271)	68,1000_=<_x_<_75,4000	496	1,52%
			75,4000_=<_x_<_82,7000	174	0,53%
			x>=_82,7000	67	0,21%

On peut ensuite observer les informations suivantes :

- La médiane est de 37
- La classe modale est [38,9000 ; 46,2000[, le mode est 42.5500
- L'écart-type est de 13,6404

On peut ensuite constater la dispersion de la variable Age.

L'étendue (différence entre la valeur max et la valeur min) est de 73,00

la dispersion interquartile (différence entre le troisième et le premier quartile) est de 20,00;

La représentation graphique ici est sous forme d'histogramme.

La forme de la distribution de l'Age est, suite aux statistiques primaires, étalée à droite car la moyenne est > médiane > mode (on peut aussi le constater en observant que le coefficient d'asymétrie de Fisher (Skewness = 0,5587) est positif.

La forme de la distribution de l'Age est moins aplatie qu'une distribution normale car coefficient d'aplatissement observé ici (Kurtosis = -01661) est négatif.

Corrélation entre chaque paire d variable quantitative

En probabilité et en statistique, étudier la corrélation entre deux ou plusieurs variables aléatoires ou statistiques *numériques*, c'est étudier l'intensité de la liaison qui peut exister entre ces variables. La mesure de la corrélation linéaire entre les deux se fait alors par le calcul du coefficient de corrélation linéaire, noté r . Dans le tableau suivant, nous présentons la corrélation entre chaque paire de nos variables quantitatives.

Linear correlation 2

Parameters

Cross-tab parameters

Sort results

non

Input list

Target (Y) and input (X)

Results

Y	X	r	r ²	t	Pr(> t)
AGE	FNLWGT	-0,0766	0,0059	-13,8709	0,0000
AGE	EDUCATION-NUM	0,0365	0,0013	6,5954	0,0000
AGE	CAPITAL-GAIN	0,0777	0,0060	14,0581	0,0000
AGE	CAPITAL-LOSS	0,0578	0,0033	10,4423	0,0000
AGE	HOURS-PER-WEEK	0,0688	0,0047	12,4358	0,0000
FNLWGT	AGE	-0,0766	0,0059	-13,8709	0,0000
FNLWGT	EDUCATION-NUM	-0,0432	0,0019	-7,8014	0,0000
FNLWGT	CAPITAL-GAIN	0,0004	0,0000	0,0779	0,9379
FNLWGT	CAPITAL-LOSS	-0,0103	0,0001	-1,8499	0,0643
FNLWGT	HOURS-PER-WEEK	-0,0188	0,0004	-3,3872	0,0007
EDUCATION-NUM	AGE	0,0365	0,0013	6,5954	0,0000
EDUCATION-NUM	FNLWGT	-0,0432	0,0019	-7,8014	0,0000
EDUCATION-NUM	CAPITAL-GAIN	0,1226	0,0150	22,2958	0,0000
EDUCATION-NUM	CAPITAL-LOSS	0,0799	0,0064	14,4677	0,0000
EDUCATION-NUM	HOURS-PER-WEEK	0,1481	0,0219	27,0256	0,0000

Dans le tableau ci-dessus, r représente le coefficient de corrélation pour chaque paire de variable, r^2 est l'écart-type

Contingence des variables qualitatives

Pour évaluer le lien existant entre deux de nos variables qualitatives, on va présenter un tableau de contingence. Le tableau de contingence est une méthode de représentation de données issues d'un comptage permettant d'estimer la dépendance entre deux caractères. Elle consiste à croiser deux caractères d'une population, par exemple dans notre cas, *workclass* et *education*, en dénombrant l'effectif correspondant à la conjonction «caractère 1» et «caractère 2». Les effectifs partiels sont rassemblés dans un tableau à double entrée, par ligne pour le premier caractère, et par colonne en fonction du second caractère: c'est le «tableau de contingence».

Cet outil simple répond à un problème crucial en statistique: la détection d'éventuelles dépendances entre les qualités relevées sur les individus d'une population.

Ainsi, nous présentons le tableau suivant avec la variable *Workclass* d'une part et le niveau d'éducation (variable *education*) de l'individu.

Contingency Chi-Square 2												
Parameters												
Cross-tab parameters												
Sort results	non											
Input list	Target (Row) and input (Column)											
Additional information	0											
Contribution threshold	2,0											
Results												
Row (Y)	Column (X)	Statistical indicator										
WORKCLASS	EDUCATION	Stat	Value		Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm	Ass
		d.f.	120	State-gov	270	268	14	169	6	325	41	
		Tschuprow's t	0,084918	Self-emp-not-inc	399	866	60	124	34	486	71	
		Cramer's v	0,099369	Private	3551	7780	923	894	387	5094	729	
		Phi²	0,078993	Federal-gov	212	263	9	67	3	254	55	
		Chi² (p-value)	2572,10 (0,0000)	Local-gov	477	503	36	342	23	387	88	
		Lambda	0,000000	?	173	532	118	48	51	514	47	
		Tau (p-value)	0,0177 (0,0000)	Self-emp-inc	273	279	14	79	10	226	35	
		U(R/C)	0,0286 (0,0000)	Without-pay	0	9	0	0	0	3	1	
		(p-value)		Never-worked	0	1	1	0	0	2	0	
		Sum	5355	10501	1175	1723	514	7291	1067			

On peut constater que, le secteur prive est le secteur qui offre le plus grand nombre d'emploi.

Les individus avec le niveau de Hs-grad représentent le plus grande valeur des employés. Il n'y a pas de Bacheliers, ni de niveau Masters, ni de niveau de 9th sans emplois ou qui ne sont jamais été employés .

Test du Khi 2

Le *Test du Khi 2* est un test statistique permettant de tester l'adéquation d'une série de données à une famille de lois de probabilités ou de tester l'indépendance entre deux variables aléatoires.

Il caractérise le degré de liaison entre les variables *Workclass* et *Education*. Il quantifie l'écart entre le tableau construit sur les données et le tableau que l'on aurait obtenu si l'hypothèse d'indépendance entre les variables était vraie. Pour nos deux variables, le *Test du Khi 2* est égal déjà 2572,10.

Results											
Row (Y)	Column (X)	Statistical indicator									
WORKCLASS	EDUCATION	Stat	Value		Bachelors	HS-grad	11th	Masters	9th	Some-college	Assoc-acdm
		d.f.	120	State-gov	270 (+ 1 %)	268 (- 2 %)	14 (- 1 %)	169 (+ 6 %)	6 (- 0 %)	325 (+ 0 %)	41 (- 0 %)
		Tschuprow's t	0,084918	Self-emp-not-inc	399 (- 0 %)	866 (+ 0 %)	60 (- 0 %)	124 (- 0 %)	34 (- 0 %)	486 (- 0 %)	71 (- 0 %)
		Cramer's v	0,099369	Private	3551 (- 0 %)	7780 (+ 1 %)	923 (+ 1 %)	894 (- 3 %)	387 (+ 0 %)	5094 (+ 0 %)	729 (- 0 %)
		Phi²	0,078993	Federal-gov	212 (+ 1 %)	263 (- 0 %)	9 (- 1 %)	67 (+ 0 %)	3 (- 0 %)	254 (+ 0 %)	55 (+ 1 %)
		Chi² (p-value)	2572,10 (0,0000)	Local-gov	477 (+ 2 %)	503 (- 2 %)	36 (- 1 %)	342 (+ 19 %)	23 (- 0 %)	387 (- 1 %)	88 (+ 0 %)
		Lambda	0,000000	?	173 (- 2 %)	532 (- 0 %)	118 (+ 2 %)	48 (- 1 %)	51 (+ 1 %)	514 (+ 1 %)	47 (- 0 %)
		Tau (p-value)	0,0177 (0,0000)	Self-emp-inc	273 (+ 2 %)	279 (- 1 %)	14 (- 1 %)	79 (+ 0 %)	10 (- 0 %)	226 (- 0 %)	35 (- 0 %)
		U(R/C) (p-value)	0,0286 (0,0000)	Without-pay	0 (- 0 %)	9 (+ 0 %)	0 (- 0 %)	0 (- 0 %)	0 (- 0 %)	3 (- 0 %)	1 (+ 0 %)
				Never-worked	0 (- 0 %)	1 (- 0 %)	1 (+ 0 %)	0 (- 0 %)	0 (- 0 %)	2 (+ 0 %)	0 (- 0 %)
				Sum	5355	10501	1175	1723	514	7291	1067

Observons la p-value de la statistique du KHI-2 de la figure ci-dessus, elle permet de déterminer s'il y a lieu de rejeter ou non l'hypothèse d'indépendance (H0), si elle est inférieure au niveau de signification (par défaut 5% sur Tanagra). Dans notre cas elle est de 0 % donc il ne faut pas rejeter H0. Le CHI-2 étant un critère additif, il serait intéressant pour nous de savoir quelles sont les cases qui y ont le plus contribué.

Les contributions sont indiquées en pourcentage du total. Le signe apparu dans les cases permet de déterminer s'il s'agit d'une attraction ou une répulsion entre les caractéristiques des variables étudiées. Lorsque la contribution d'une case est 2 fois plus élevée que la contribution moyenne, elle est surlignée en rouge, d'où la présence de la couleur rouge dans notre tableau. On peut déduire que la liaison entre la variable

secteur de travail(workclass) et Education repose avant tout sur une forte attraction évaluée à 19 % entre la valeur Federal-gov et le niveau d'éducation Masters.

Application des méthodes factorielles aux jeux de données « Census Income Data Set »

ANALYSE EN COMPOSANTES PRINCIPALE (ACP)


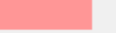
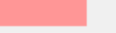
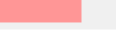

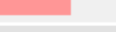
L'analyse en composante principale est une méthode de la famille de l'analyse des données et plus généralement de la statistique multivariée, qui consiste à transformer des variables liées entre elles (dites « corrélées » en statistique) en nouvelles variables décorréliées les unes des autres. Ces nouvelles variables sont nommées « composantes principales », ou axes principaux. Elle permet au praticien de réduire le nombre de variables et de rendre l'information moins redondante.

Il s'agit d'une approche à la fois géométrique (les variables étant représentées dans un nouvel espace, selon des directions d'inertie maximale) et statistique (la recherche portant sur des axes indépendants expliquant au mieux la variabilité — la variance — des données).

Ici, notre objectif de cette analyse en composante principale est de partir d'un ensemble de données contenant 32561 observations et 15 variables continues et discrètes, pour chercher à résumer l'information disponible à l'aide de variables synthétiques appelées composantes principales. Grâce à la fonction «Principal Component Analysis» située sous l'onglet «Factorial Analysis», nous avons réalisé l'ACP. Les résultats obtenus sont présentés dans la suite.

VALEURS PROPRES

Le tableau contenant les valeurs propres de la matrice des corrélations nous est présenté dans le tableau ci-dessous :

Principal Component Analysis 1					
Parameters					
Number of asked factors : 5 Compute COS2 and CTR : 0 Standardizing attributes : 1 Bartlett's test and MSA (KMO indices) : 0 Correlations and partial correlations : 0 Reproduced correlations : 0 Sort variables according to loadings : 1					
Results					
Eigen values					
Matrix trace		6,000000			
Average		1,000000			
Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	1,310633	0,269666	21,84 %		21,84 %
2	1,040966	0,022367	17,35 %		39,19 %
3	1,018599	0,076807	16,98 %		56,17 %
4	0,941792	0,055348	15,70 %		71,87 %
5	0,886443	0,084877	14,77 %		86,64 %
6	0,801566	-	13,36 %		100,00 %
Tot.	6,000000	-	-	-	-

En étudiant le tableau précédent, on peut constater que l'inertie expliquée par le premier axe principal est évalué à $\Delta=1.32$. Un autre constat nous permet de constater que la première composante principale occupe à elle seule 21,84 %.

Se basant sur le critère de Kaiser-Guttman utilisé par tanagra pour la définition des composantes principales les plus significatives, on observe que seules les quatre premières valeurs propres sont importantes. Nous ne retiendrons que ces derniers dans la suite de notre étude.

Voir tableau suivant.

Significance of Principal Components

Global critical values	
Kaiser-Guttman	1
Karlis-Saporta-Spinaki	1,02478

Eigenvalue table - Test for significance

Eigenvalues - Significance		
Axis	Eigenvalue	Broken-stick critical values
1	1,310633	2,450000
2	1,040966	1,450000
3	1,018599	0,950000
4	0,941792	0,616667
5	0,886443	0,366667
6	0,801566	0,166667

CORRÉLATION ENTRE LES VARIABLES ET LES AXES PRINCIPAUX

Dans la seconde partie des résultats, on peut constater suivant une indication claire, la corrélation qui existe entre les variables avec les axes factoriels, voir schéma suivant.

Factor Loadings [Communality Estimates]

Attribute	Axis_1		Axis_2		Axis_3		Axis_4		Axis_5	
	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)
EDUCATION-NUM	-0,63063	40 % (40 %)	-0,18199	3 % (43 %)	-0,21939	5 % (48 %)	-0,35561	13 % (61 %)	0,27537	8 % (68 %)
HOURS-PER-WEEK	-0,58574	34 % (34 %)	-0,13766	2 % (36 %)	-0,21256	5 % (41 %)	-0,14421	2 % (43 %)	-0,70870	50 % (93 %)
FNLWGT	0,24081	6 % (6 %)	-0,55212	30 % (36 %)	-0,52660	28 % (64 %)	0,55470	31 % (95 %)	-0,09724	1 % (96 %)
CAPITAL-GAIN	-0,47510	23 % (23 %)	-0,53437	29 % (51 %)	0,32861	11 % (62 %)	0,21325	5 % (66 %)	0,39990	16 % (82 %)
CAPITAL-LOSS	-0,30579	9 % (9 %)	0,51934	27 % (36 %)	-0,62359	39 % (75 %)	0,21094	4 % (80 %)	0,33282	11 % (91 %)
AGE	-0,43889	19 % (19 %)	0,35890	13 % (32 %)	0,38874	15 % (47 %)	0,62997	40 % (87 %)	-0,16798	3 % (90 %)
Var. Expl.	1,31063	22 % (22 %)	1,04097	17 % (39 %)	1,01860	17 % (56 %)	0,94179	16 % (72 %)	0,88644	15 % (87 %)

Factor Score Coefficients

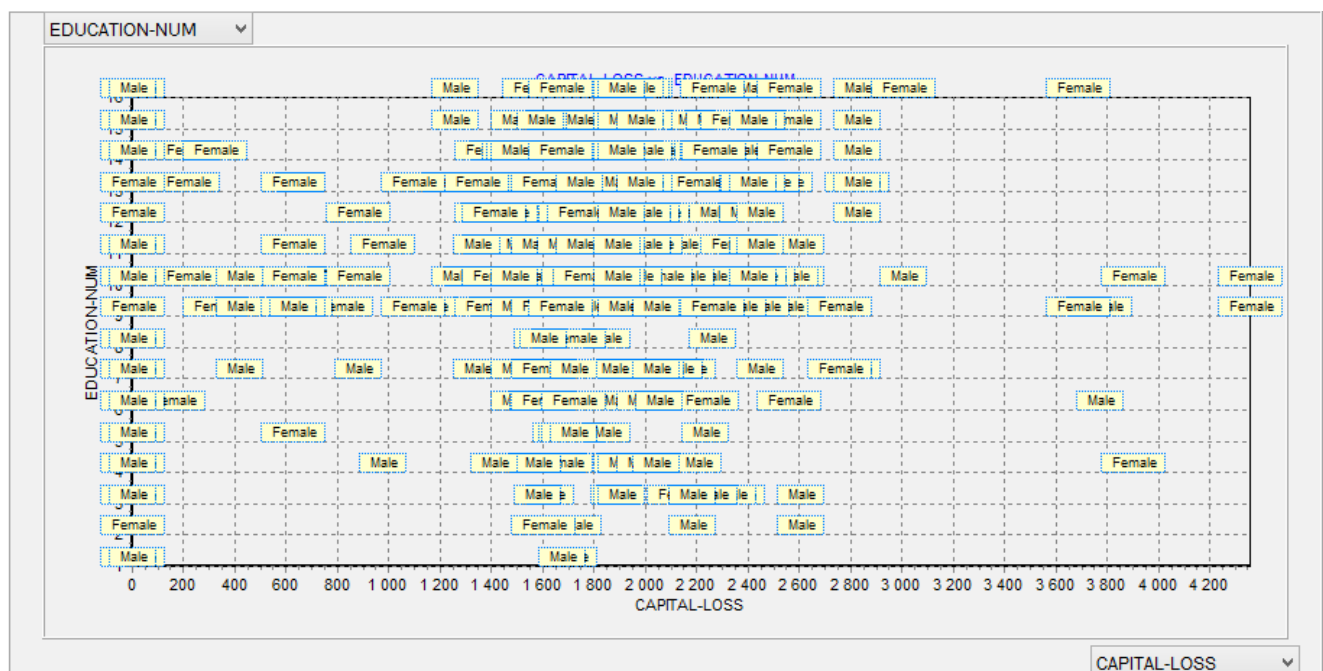
Attribute	Mean	Std-dev	Axis_1	Axis_2	Axis_3	Axis_4	Axis_5
AGE	38,5816468	13,6402231	-0,3833714	0,3517641	0,3851754	0,6491510	-0,1784146
FNLWGT	189778,3665121	105548,3568809	0,2103434	-0,5411439	-0,5217732	0,5715859	-0,1032819
EDUCATION-NUM	10,0806793	2,5726808	-0,5508545	-0,1783703	-0,2173817	-0,3664390	0,2924763
CAPITAL-GAIN	1077,6488437	7385,1786769	-0,4149957	-0,5237458	0,3255972	0,2197423	0,4247427
CAPITAL-LOSS	87,3038297	402,9540308	-0,2671034	0,5090136	-0,6178736	0,2173597	0,3534959
HOURS-PER-WEEK	40,4374559	12,3472391	-0,5116402	-0,1349254	-0,2106097	-0,1486037	-0,7577238

Que peut-on en déduire du tableau précédent ?

Nous notons que le premier axe est fortement corrélé négativement avec les variables : EDUCATION-NUM et HOURS-PER-WEEK, c'est-à-dire qu'il prend en compte le niveau d'éducation des gens et aussi le nombre d'heures de travail par semaine . On peut cependant constater que de la variable opposée associée à cet axe, est corrélée positivement pour ces mêmes attributs. Le premier axe ainsi défini est donc crucial dans la détermination des indices pouvant influencer le revenu annuel des gens. Le deuxième axe par contre est corrélé négativement pour les variables FNLWGT, CAPITAL-LOSS et CAPITAL-GAIN, ce qui signifie qu'il fait référence aux données liées aux origines des gens (FNLWGT), de pertes ou dépenses qui peuvent affecter les indices.

La prochaine étape de notre analyse en composante principale concerne le plan factoriel.

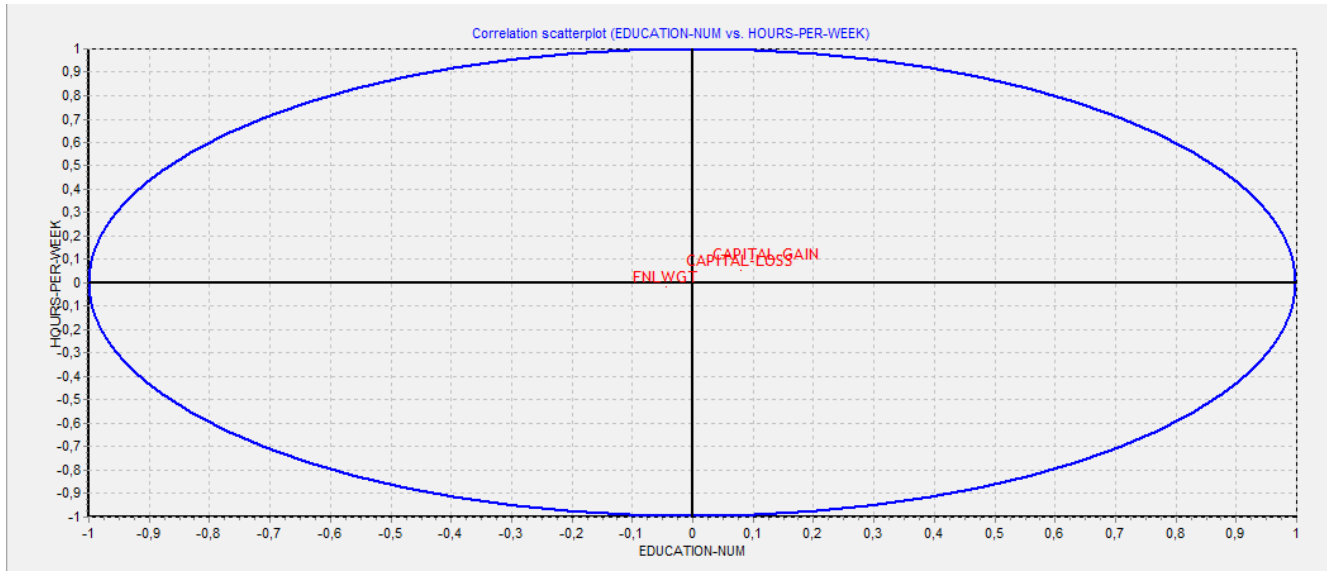
PLANS FACTORIELS



Considérant les résultats du tableau précédent, on peut constater que la majorité des gens qui travaillent sont de sexe M et ont aussi un niveau d'éducation plus élevé, ces variables ont toutes une valeur positive pour l'axe principal choisi ; cependant, on peut aussi constater que les femmes avec un niveau d'éducation élevé et qui travaillent beaucoup par semaine perdent beaucoup par année en capital. Ce qui équivaut à des valeurs élevées des indices du de CAPITAL-LOSS de l'enquête. Autrement dit, dans la

majorité des cas, les gens de sexe M (Male) travaillent plus en semaine et ont un niveau d'éducation plus élevé.

CERCLE DES CORRÉLATIONS



Nous constatons que les variables liées aux indices EDUCATION, HOURS-PER-WEEK et CAPITAL-GAIN sont très proches du cercle et du deuxième axe principal ce qui confirme leur forte corrélation positive avec la première composante principale.

En somme, l'analyse du cercle des corrélations confirme toutes les observations que nous avons précédemment faites.

ANALYSE FACTORIELLE DES CORRESPONDANCES

Nous avons effectué une analyse factorielle des correspondances entre les variables qualitatives de notre jeu de données. La relation entre ces variables est présentée dans le tableau de contingence à la figure du paragraphe étude d'une variable qualitative.

Résultats numériques de l'AFC

Statistical indicator	
Stat	Value
d.f.	120
Tschuprow's t	0,084918
Cramer's v	0,099369
Phi ²	0,078993
Chi ² (p-value)	2572,10 (0,0000)

Le test de CHI 2

Ce test indique s'il existe un lien entre les lignes et les colonnes du tableau de contingence.







Les valeurs propres:

Le tableau suivant permet d'observer la valeur propre calculée, le pourcentage d'inertie associée à chaque axe et le pourcentage cumulé qui permet d'avoir une idée du nombre d'axes à retenir. Dans notre cas, les quatre premiers axes résument 95,17% de l'information disponible.

Results

Eigen values

Matrix trace	6,000000
Average	1,000000

Axis	Eigen value	Difference	Proportion (%)	Histogram	Cumulative (%)
1	1,310633	0,269666	21,84 %		21,84 %
2	1,040966	0,022367	17,35 %		39,19 %
3	1,018599	0,076807	16,98 %		56,17 %
4	0,941792	0,055348	15,70 %		71,87 %
5	0,886443	0,084877	14,77 %		86,64 %
6	0,801566	-	13,36 %		100,00 %
Tot.	6,000000	-	-	-	-

REPRÉSENTATION GRAPHIQUE DE L'AFC

Plan factoriel Dans l'onglet CHART de la fenêtre de visualisation, nous pouvons observer les graphiques suivant les différents plans factoriels ; par exemple, le premier plan factoriel est présenté sur la Figure 16.

Coordonnées factorielles

nous obtenons aussi les coordonnées factorielles de chaque point individus dans le tableau ci-dessous

Rows analysis

Characterization				Coord.		Contributions (%)		COS²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos² 1	cos² 2
Handlers-cleaners	0,00003	1,02460	0,00003	-1,00357	-0,13202	0,07	0,00	0,98 (0,98)	0,02 (1,00)
?	0,00003	0,89551	0,00003	-0,94625	0,01020	0,07	0,00	1,00 (1,00)	0,00 (1,00)
?	0,00004	0,83403	0,00003	-0,91322	-0,00664	0,07	0,00	1,00 (1,00)	0,00 (1,00)
?	0,00003	0,98278	0,00003	-0,99018	-0,04812	0,06	0,00	1,00 (1,00)	0,00 (1,00)
?	0,00003	1,00613	0,00003	-0,99559	-0,12214	0,06	0,00	0,99 (0,99)	0,01 (1,00)
?	0,00003	0,87577	0,00003	-0,93122	0,09272	0,06	0,00	0,99 (0,99)	0,01 (1,00)
?	0,00003	0,97958	0,00003	-0,98869	-0,04545	0,06	0,00	1,00 (1,00)	0,00 (1,00)
?	0,00003	0,89890	0,00003	-0,93737	-0,14220	0,06	0,01	0,98 (0,98)	0,02 (1,00)
?	0,00003	0,90708	0,00003	-0,95140	0,04367	0,06	0,00	1,00 (1,00)	0,00 (1,00)
?	0,00003	0,84066	0,00003	-0,91665	0,02014	0,06	0,00	1,00 (1,00)	0,00 (1,00)
Other-service	0,00003	0,97631	0,00003	-0,98715	-0,04271	0,06	0,00	1,00 (1,00)	0,00 (1,00)
?	0,00003	0,99944	0,00003	-0,99266	-0,11852	0,06	0,00	0,99 (0,99)	0,01 (1,00)
?	0,00003	0,82072	0,00003	-0,90459	0,04930	0,06	0,00	1,00 (1,00)	0,00 (1,00)
Exec-managerial	0,00003	0,78403	0,00003	-0,87710	0,12133	0,06	0,00	0,98 (0,98)	0,02 (1,00)
Priv-house-serv	0,00003	0,86305	0,00003	-0,92263	-0,10860	0,06	0,00	0,99 (0,99)	0,01 (1,00)
Adm-clerical	0,00003	0,88254	0,00003	-0,93615	0,07842	0,06	0,00	0,99 (0,99)	0,01 (1,00)

Interprétations des axes factoriels par les points individus

- ▲ L'axe « *characterization* » définit les différents types d'emploi pour les différents clusters.
- ▲ L'inertie est fixé à 0.00003 pour ce cluster.
- ▲ La colonne Sq. Dist fait surtout ressortir l'importance des valeurs de chaque catégorie d'emploi dans la détermination du salaire annuel.

Columns analysis

Characterization				Coord.		Contributions (%)		COS ²	
Values	Weight	Sq. Dist.	Inertia	coord 1	coord 2	ctr 1	ctr 2	cos ² 1	cos ² 2
AGE	0,43302	0,05483	0,02374	-0,23237	-0,02848	53,58	3,08	0,98 (0,98)	0,01 (1,00)
HOURS-PER-WEEK	0,45384	0,04623	0,02098	0,20982	-0,04696	45,79	8,78	0,95 (0,95)	0,05 (1,00)
EDUCATION-NUM	0,11314	0,09101	0,01030	0,04791	0,29787	0,60	88,08	0,03 (0,03)	0,97 (1,00)

Dans le tableau ci-dessus, nous présentons les résultats descriptive des valeurs et coordonnées pour le 1^{er} cluster

CLASSIFICATIONS

Classifications des individus

K-means

Afin de classier les individus de notre ensemble de données, nous avons appliqué, la méthode K-means avec les paramètres suivants:

5 clusters;

05 essais;

10 itérations au maximum.

Pour ce faire, nous avons considéré toutes les variables continues et nous avons utilisé le composant K-means du groupe Clustering de Tanagra. La figure ci-dessous montre la répartition des individus dans les différents clusters, dont 3.

Cluster size and WSS

Clusters	3		
Cluster	Description	Size	WSS
cluster n°1	c_kmeans_1	15810	53983,5783
cluster n°2	c_kmeans_2	159	554,4779
cluster n°3	c_kmeans_3	16592	87101,3781

On remarque que le cluster No3 a le plus grand nombre d'individus (16592) tandis que le cluster N°2 a le plus petit nombre d'individus (159).

Explication de la partition dans l'inertie totale

Global evaluation

D'après la figure la figure ci-contre, la partition explique 27,5% de l'inertie totale.

Within Sum of Squares	141639,4343
Total Sum of Squares	195366,0000
R-Square	0,2750

Coordonnées des centres des clusters

Les coordonnées des centres de chacun des 3 clusters sont présentées à la figure suivante

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
AGE	29,292853	46,358491	47,358124
FNLWGT	222523,017900	192968,886792	158546,433582
EDUCATION-NUM	9,187666	12,918239	10,904412
CAPITAL-GAIN	235,199241	99999,000000	932,436295
CAPITAL-LOSS	19,461290	0,000000	152,785499
HOURS-PER-WEEK	36,553890	49,798742	44,048276

Caractérisation des clusters

Afin de comprendre la répartition des individus dans les différents groupes, nous avons effectué une caractérisation de nos groupes grâce à l'outil Group Characterization de Tanagra. Les résultats obtenus sont présentés à la figure suivante.

OCCUPATION= Adm-clerical				OCCUPATION= Exec-managerial			
Examples		[11,6 %] 3770		Examples		[12,5 %] 4066	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
EDUCATION-NUM	0,83	10,11 (1,70)	10,08 (2,57)	EDUCATION-NUM	36,25	11,45 (2,14)	10,08 (2,57)
AGE	-7,74	36,96 (13,36)	38,58 (13,64)	HOURS-PER-WEEK	25,12	44,99 (11,11)	40,44 (12,35)
HOURS-PER-WEEK	-15,23	37,56 (9,59)	40,44 (12,35)	AGE	17,93	42,17 (11,97)	38,58 (13,64)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

RELATIONSHIP= Not-in-family				RELATIONSHIP= Husband			
Examples		[25,5 %] 8305		Examples		[40,5 %] 13193	
Att - Desc	Test value	Group	Overall	Att - Desc	Test value	Group	Overall
Continuous attributes : Mean (StdDev)				Continuous attributes : Mean (StdDev)			
EDUCATION-NUM	9,64	10,32 (2,57)	10,08 (2,57)	AGE	57,18	43,82 (12,02)	38,58 (13,64)
HOURS-PER-WEEK	1,35	40,60 (11,76)	40,44 (12,35)	HOURS-PER-WEEK	44,42	44,12 (11,67)	40,44 (12,35)
AGE	-1,82	38,35 (13,87)	38,58 (13,64)	EDUCATION-NUM	14,23	10,33 (2,73)	10,08 (2,57)
Discrete attributes : [Recall] Accuracy				Discrete attributes : [Recall] Accuracy			

On note ainsi, entre autres, après analyse de ces résultats que les individus de chaque groupe, on peut constater par exemple pour une occupation de Adm-clerical du groupe 1, les valeurs pour les attributs AGES, HOUR-PER-WEEK sont supérieures à la moyenne générale; tandis que pour les individus du groupe 3 toutes les valeurs des variables sont supérieures à la moyenne à l'exception des valeurs de EDUXCATION-NUM. Le terme «moyenne générale» est ici relatif à l'espérance globale; c'est-à-dire celle qui prend en compte tous les individus de chaque groupe ou cluster du jeu de données.

HAC

Pour illustrer la classification ascendante hiérarchique (HAC), nous considérerons toutes les variables quantitatives sauf la variable salary (elle indique la surface brûlée par l'incendie). Par la suite nous essayerons de déterminer quel groupe de variables expliquerait le mieux les valeurs de salary

Après application du composant HAC de Tanagra aux données issues du clustering effectué avec la méthode K-Means, on observe que seul 04 clusters sont nécessaires à la classification des individus , voir figure suivante

Best cluster selection

Clusters	BSS ratio	Gap
1	0,0000	0,0000
2	0,0387	6,2532
3	0,0450	1,0293
4	0,0460	0,0344
5	0,0468	0,0904
6	0,0471	0,0116
7	0,0473	0,0319
8	0,0474	0,0112
9	0,0474	0,0031

Ici, dans le tableau ci-contre, la sélection du nombre optimal de clusters.

Résultat du Clustering

Clustering results

Clusters	From the dendrogram	After one-pass relocation
cluster n°1	22703	22703
cluster n°2	3671	3671
cluster n°3	6187	6187

La figure ci-contre, on présente les trois clusters obtenus après réorganisation des données. Le cluster No1 avec 22703 données.

Coordonnées des centres des clusters

Cluster centroids

Attribute	Cluster n°1	Cluster n°2	Cluster n°3
AGE	36,792582	45,298829	41,160983
EDUCATION-NUM	9,878959	10,498774	10,572814

Étant donné que notre variable de sortie est Salary, nous l'avons exclue de la classification afin de comparer les classes qu'elle contient avec celles issues de nos opérations de clustering et de HAC. Dans l'ensemble de données qui est l'objet de notre analyse, nous pouvons facilement regrouper les individus en 3 classes selon la valeur de la variable.

Dans la figure suivante, nous tenons à montrer qu'après la classification de nos individus avec respectivement K-means et HAC, nous obtenons le même nombre de classes . Dans

l'objectif de comparer le résultat ci-dessus du tableau de clustering avec celui produit par Tanagra, nous utilisons le composant «View dataset » de l'onglet «Data visualisation» dans Tanagra au composant HAC déjà exécuté. Ce dernier présente pour chaque individu la classe à laquelle l'individu appartient à l'issue de la HAC. Clairement, nous y distinguons 03 classes qui sont: c_hac_1, c_hac_2, et c_hac_3. Nous choisissons d'exporter le résultat obtenu dans Microsoft Office Excel 2010 afin de mieux le visualiser et l'analyser. Les 03 classes issues de la classification HAC, sont réparties comme présenté dans la figure ci-dessous.

No	Classe	Nombre d'Individus
1	c_hac_1	22703
2	c_hac_2	3671
3	c_hac_3	6187
	Total	32561

Donc, en comparant ce résultat obtenu manuellement, nous constatons que nos deux classifications, K-Means et HAC, ont exactement le même nombre d'individu. Effectivement, la répartition des individus par classe est exactement la même d'une formule déjà une autre.

Conclusion

En effet, dans ce travail, nous avons présenté l'Analyse Factorielle des Correspondances de notre jeu de données, l'analyse factorielle, le clustering, la répartition des classes, le méthode K-means, le Hac, *ect.*, sont quelques paramètres statistique présents dans ce travail. Dans un prochain travail, nous présenterons une autre partie d'analyse de notre jeu de données.

Références

- [1] <http://www.analyse-donnees.fr/services-analyse-enquetes-traitement/>
- [2] <http://archive.ics.uci.edu/ml/machine-learning-databases/adult/adult.names>
- [3] Cours de Fouilles de Données, Master 1 , IFI 2016, Enseignante NGUYỄN Thị Minh Huyền