Faculté des Sciences
Aix*Marseille Université

Master 1 MAS – DS
Statistique exploratoire
Compte rendu du mini projet

Analyse en composantes principales

Classification ascendante hiérarchique

Réalisé par :

- Saïda GUEZOUI
- Benoit GILLES

Encadré par :

- Thomas Willer

2020/2021

Table des matières

I. Introduction 3
II. Présentation des données 3
III. Etude univariée 3
A. Proc MEANS 3
B. Proc UNIVARIATE 4
C. Procédure SGPLOT 4
IV. Etude bivariée 6
V. Etude multivariée 7
A. Comparaison entre ACP normé et ACP canonique 7
B. ACP 7
B.1. Détermination du nombre d'axes 7
B.2. Représentation des individus sur les axes 8
B.3. Représentation des variables sur les axes 9
B.4. Cercle de corrélation 11
B.5. Carte des individus 11
VI. Classification 12
A. But de la classification 12
B. Procédure CLUSTER 12
C. Choix du nombre de classes 13
D. Interprétation des classes 14
E. Caractérisation des classes 15
VII. Synthèse 15

I. Introduction

Ce rapport traite le sujet de la pauvreté au Kenya. Nous souhaitons, à partir de données récoltées lors d'une étude, tirer des conclusions quant à la répartition de la pauvreté au Kenya, selon les différents comtés et vérifier s'il existe ou non un lien entre la situation géographique d'un foyer et la pauvreté de celui-ci. Notre étude se divisera en 4 parties majeures. Après une présentation très brève des données, nous ferons dans un premier temps une analyse univariée, en présentant des résultats classiques pour les variables que nous avons jugées intéressantes. Par la suite, nous étudierons les corrélations entre les différentes variables dans une étude bivariée. En troisième partie, nous effectuerons une analyse par composantes principales afin de permettre une meilleure visualisation des données, et expliquer le lien entre les variables et les comtés. Enfin nous terminerons en effectuant une classification, et en tirant des conclusions vis-à-vis du lien entre la pauvreté et la localisation géographique du comté.

II. Présentation des données

La table SAS utilisée au long de ce projet est nommée KENYA. En effet, en utilisant la procédure CONTENTS qui affiche le dictionnaire, on en déduit que cette table est composée de 47 lignes (observations) représentant l'ensemble des comtés qui divisent administrativement le Kenya. Elle dispose aussi de 17 colonnes (variables) pour lesquelles les valeurs correspondent au pourcentage d'accès à un bien pour chaque comté. On remarque aussi que toutes les variables sont numériques à part la variable county qui est une chaine de charactère.

III. Etude univariée

A. Proc MEANS

A l'aide de la procédure MEANS, on peut calculer les statistiques univariées classiques telles que : la moyenne, la médiane, l'écart type, le min, le max etc. La sortie de cette procédure est donnée par le tableau ci-dessous.

Variable	Minimum	Mean	Maximum	Median	Lower 95% CL for Mean	Upper 95% CL for Mean	Variance	Lower Quartile	Upper Quartile
electricity	0.04	0.26	0.88	0.21	0.20	0.31	0.03	0.14	0.33
radio	0.09	0.62	0.84	0.66	0.57	0.67	0.03	0.57	0.72
television	0.05	0.26	0.69	0.23	0.22	0.30	0.02	0.17	0.33
refrigerator	0.00	0.04	0.17	0.03	0.03	0.06	0.00	0.02	0.06
landlinephone	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.01
mobilephone	0.33	0.81	0.98	0.84	0.77	0.84	0.02	0.77	0.89
solarpanel	0.01	0.10	0.24	0.11	0.09	0.12	0.00	0.06	0.14
table	0.17	0.78	0.97	0.87	0.73	0.84	0.04	0.71	0.92
chair	0.24	0.80	0.98	0.86	0.75	0.86	0.04	0.70	0.94
sofa	0.08	0.47	0.84	0.54	0.42	0.53	0.04	0.35	0.62
bed	0.16	0.91	0.99	0.93	0.87	0.95	0.02	0.91	0.97
cupboard	0.05	0.40	0.77	0.40	0.34	0.45	0.03	0.29	0.50
clock	0.01	0.16	0.35	0.15	0.13	0.18	0.01	0.09	0.22
microwave	0.00	0.02	0.12	0.01	0.01	0.03	0.00	0.01	0.03
dvdplayer	0.02	0.16	0.51	0.12	0.13	0.19	0.01	0.09	0.21
cdplayer	0.01	0.09	0.26	0.08	0.07	0.10	0.00	0.05	0.11

Si on s'intéresse à la moyenne des variables, on voit que le nombre moyen des ménages interrogés ayant accès à l'électricité ou à la télévision est de 26%, ils ont plus facilement accès à la radio (moyenne de 0.91) et à des biens comme des chaises, des tables et des téléphones mobiles. Les autres variables ont une moyenne faible par rapport à celles d'un pays développé.

Les variables table chair, sofa, cupboard, electricity et radio ont une forte variance par rapport à landlinephone, cdplayer, solarpanel et microwave. Ces variables vont contribuer le plus à la formation des valeurs propres des axes.

Concernant la statistique « minimum », on remarque que les variables microwave et refrigerator ont un minimum de 0, ce qui nous montre qu'il existe des ménages qui n'ont pas accès à ces biens.

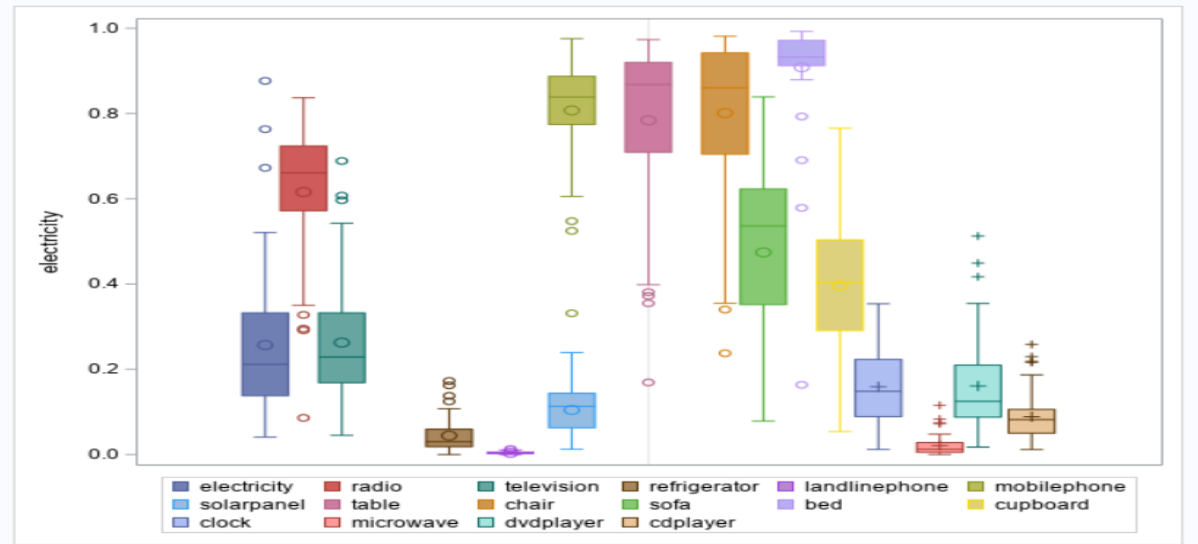
On remarque également qu'il y a au maximum, par comté, 88% des ménages interrogés ayant accès à l'électricité et au maximum 69% possédant la radio tandis que les médianes sont égales à 21%, 23 % respectivement, on en déduit donc qu'il y a une inégalité de possession de biens au sein des ménages kenyans ayant répondu à l'enquête.

B. Proc UNIVARIATE

La procédure UNIVARIATE permet de calculer plus de statistiques. En effet, si aucune variable n'est précisée en tant que paramètre, la procédure UNIVARIATE affichera, pour chaque variable de la table, un ensemble plus complet de statistiques, tels que divers quantiles, les tests de tendance centrale, ou les valeurs extrêmes prise par les variables. En résumé, la procédure UNIVARIATE reprend toutes les options et toutes les instructions de la procédure MEANS.

C. Procédure SGPLOT

A l'aide de la procédure SGPLOT, on obtient les boxplots de chaque variable comme suit :



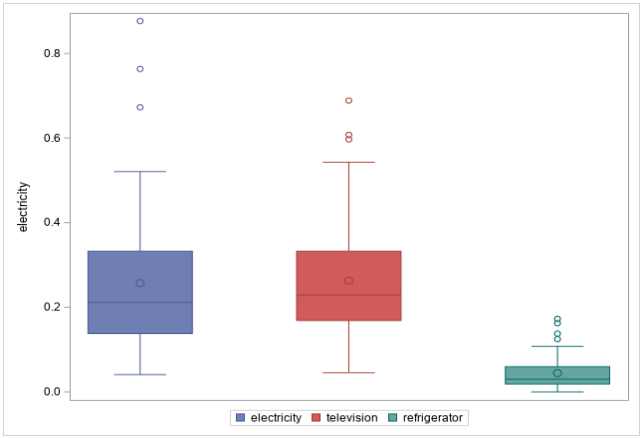
A l'aide de la procédure MEANS, nous avons pu constater une inégalité dans la possession de biens au sein des ménages interrogés au Kenya. En effet, en affichant ces boxplots ci-dessus, nous remarquons qu'on peut les séparer en trois ensembles selon la statistique « médiane ».

Le premier ensemble correspond aux boxplots des variables refrigerator, landlinephone, solarpanel, clock, microwave, dvdplayer et cdplayer. Ce sont les variables ayant une médiane inférieure à 0.2.

Le deuxième ensemble contient les boxplots des variables electricity, radio, television, sofa, cupboard qui ont une médiane entre 0.2 et 0.8.

Concernant le dernier ensemble, il est constitué des variables mobilephone, table, chair et bed avec une médiane supérieure à 0.8.

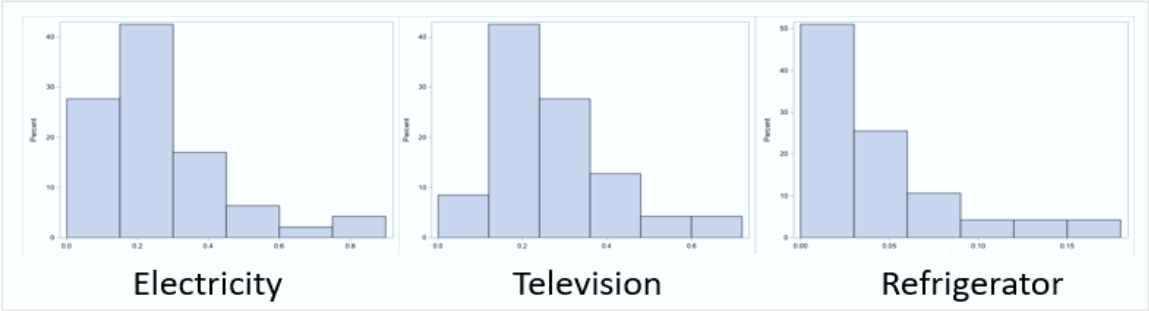
Prenons l'exemple des trois variables : electricity, television et refrigerator et faisons une comparaison :



On remarque une similarité entre le boxplot de la variable electricity et celui de la variable television tandis que celui de refrigirateur est plus écrasé.

La variable electricity a une moyenne de 0.26 et une médiane de 0.21. On sépare le boxplot en 4 segments. Par exemple, le segment [0.15, 0.21] correspond aux comtés qui sont pauvres en électricité. Le segment [0.35,0.38] représente les comtés les mieux équipées d'électricité. Il y a également des points atypiques correspondant aux comtés Kiambu, Mombasa et Nairobi. Concernant la possession du réfrigérateur, elle est plus faible et rare.

On peut également afficher les histogrammes de ces trois variables :



Concernant l'histogramme de l'électricité, le pic des données est fixé au environ de 20% des ménages interrogés et la dispersion des données s'étend de 0.04 à 0.88. Il y a une similarité avec l'histogramme de la variable television comme on l'a déjà mentionné avec les boxplots.

En revanche, la dispersion des données de la variable refrigerator s'étend de 0 à 0.17 et le pic des données est autour des 4% des ménages interrogés. On remarque une asymétrie à droite, ce qui nous indique que les données peuvent ne pas être normalement distribuées.

IV. Etude bivariée

Pearson Correlation Coefficients, N = 47
Prob > |r| under H0: Rho=0

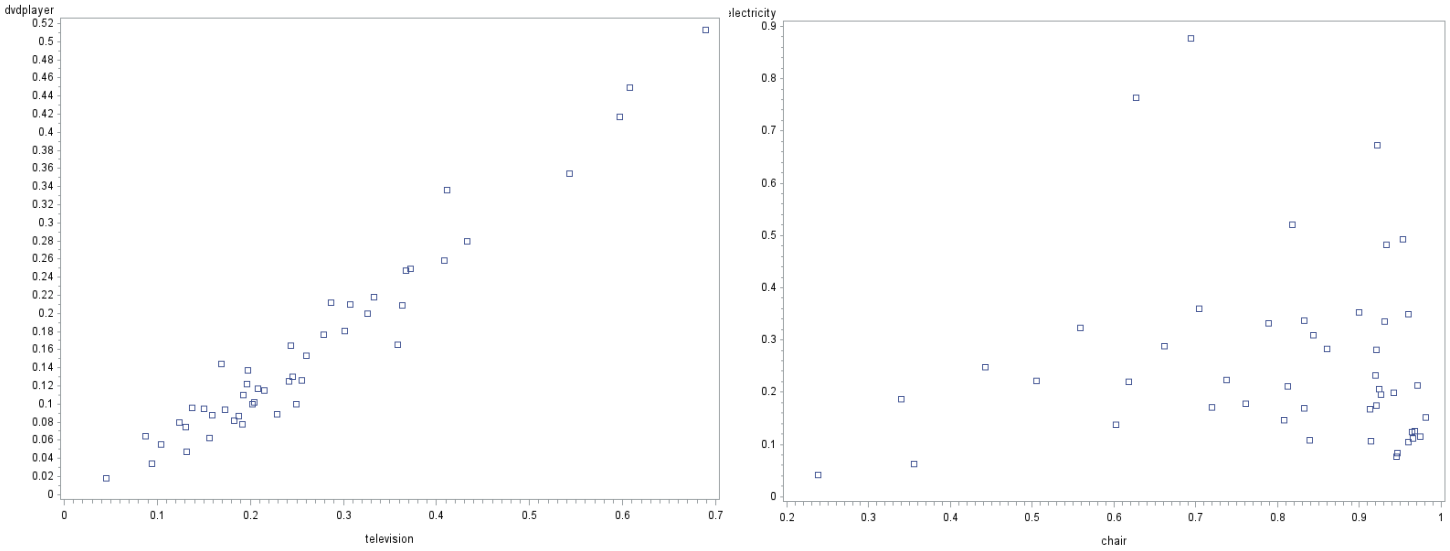
	electricity	radio	television	refrigerator	landlinephone	mobilephone	solarpanel	table	chair	sofa	bed	cupboard	clock	microwave	dvdplayer	cdplayer
electricity	1.00000	0.40908 0.0043	0.94218 <.0001	0.89569 <.0001	0.40817 0.0044	0.53183 0.0001	-0.26087 0.0765	0.23014 0.1197	0.00624 0.9668	0.49874 0.0004	0.25831 0.0796	0.51249 0.0002	0.64786 <.0001	0.92274 <.0001	0.95676 <.0001	0.80320 <.0001
radio	0.40908 0.0043	1.00000	0.62783 <.0001	0.20855 0.1595	0.02236 0.8814	0.88997 <.0001	0.54324 <.0001	0.90548 <.0001	0.74154 <.0001	0.85541 <.0001	0.66691 <.0001	0.90041 <.0001	0.78916 <.0001	0.37029 0.0104	0.56525 <.0001	0.58954 <.0001
television	0.94218 <.0001	0.62783 <.0001	1.00000	0.78196 <.0001	0.36128 0.0126	0.68070 <.0001	-0.01927 0.8977	0.45534 0.0013	0.18755 0.2068	0.71968 <.0001	0.36015 0.0129	0.73416 <.0001	0.79910 <.0001	0.86576 <.0001	0.97218 <.0001	0.84345 <.0001
refrigerator	0.89569 <.0001	0.20855 0.1595	0.78196 <.0001	1.00000	0.41130 0.0041	0.42917 0.0026	-0.38399 0.0077	0.03644 0.8079	-0.12681 0.3957	0.22534 0.1278	0.16343 0.2724	0.26001 0.0776	0.54079 <.0001	0.88800 <.0001	0.84225 <.0001	0.72929 <.0001
landlinephone	0.40817 0.0044	0.02236 0.8814	0.36128 0.0126	0.41130 0.0041	1.00000	0.05700 0.7036	-0.12778 0.3920	-0.01985 0.8946	-0.10895 0.4660	0.11590 0.4379	-0.05145 0.7312	0.04192 0.7796	0.25417 0.0847	0.44424 0.0018	0.34284 0.0183	0.20336 0.1704
mobilephone	0.53183 0.0001	0.88997 <.0001	0.68070 <.0001	0.42917 0.0026	0.05700 0.7036	1.00000	0.42347 0.0030	0.79433 <.0001	0.62410 <.0001	0.71679 <.0001	0.72012 <.0001	0.78346 <.0001	0.76865 <.0001	0.49285 0.0004	0.64860 <.0001	0.64259 <.0001
solarpanel	-0.26087 0.0765	0.54324 <.0001	-0.01927 0.8977	-0.38399 0.0077	-0.12778 0.3920	0.42347 0.0030	1.00000	0.51466 0.0002	0.44811 0.0016	0.36983 0.0105	0.32907 0.0239	0.45776 0.0012	0.24113 0.1025	-0.28475 0.0524	-0.09672 0.5178	-0.09032 0.5460
table	0.23014 0.1197	0.90548 <.0001	0.45534 0.0013	0.03644 0.8079	-0.01985 0.8946	0.79433 <.0001	0.51466 0.0002	1.00000	0.86397 <.0001	0.82278 <.0001	0.71688 <.0001	0.79948 <.0001	0.70412 <.0001	0.21966 0.1379	0.37298 0.0098	0.44445 0.0017
chair	0.00624 0.9668	0.74154 <.0001	0.18755 0.2068	-0.12681 0.3957	-0.10895 0.4660	0.62410 <.0001	0.44811 0.0016	0.86397 <.0001	1.00000	0.58339 <.0001	0.65919 <.0001	0.59010 <.0001	0.51129 0.0002	0.03443 0.8183	0.13502 0.3655	0.28117 0.0556
sofa	0.49874 0.0004	0.85541 <.0001	0.71968 <.0001	0.22534 0.1278	0.11590 0.4379	0.71679 <.0001	0.36983 0.0105	0.82278 <.0001	0.58339 <.0001	1.00000	0.50323 0.0003	0.90310 <.0001	0.79159 <.0001	0.45709 0.0012	0.62715 <.0001	0.59013 <.0001
bed	0.25831 0.0796	0.66691 <.0001	0.36015 0.0129	0.16343 0.2724	-0.05145 0.7312	0.72012 <.0001	0.32907 0.0239	0.71688 <.0001	0.65919 <.0001	0.50323 0.0003	1.00000	0.52760 0.0001	0.45369 0.0014	0.19848 0.1811	0.31360 0.0318	0.36202 0.0124
cupboard	0.51249 0.0002	0.90041 <.0001	0.73416 <.0001	0.26001 0.0776	0.04192 0.7796	0.78346 <.0001	0.45776 0.0012	0.79948 <.0001	0.59010 <.0001	0.90310 <.0001	0.52760 0.0001	1.00000	0.86711 <.0001	0.43989 0.0020	0.64882 <.0001	0.61041 <.0001
clock	0.64786 <.0001	0.78916 <.0001	0.79910 <.0001	0.54079 <.0001	0.25417 0.0847	0.76865 <.0001	0.24113 0.1025	0.70412 <.0001	0.51129 0.0002	0.79159 <.0001	0.45369 0.0014	0.86711 <.0001	1.00000	0.62641 <.0001	0.74812 <.0001	0.77325 <.0001
microwave	0.92274 <.0001	0.37029 0.0104	0.86576 <.0001	0.88800 <.0001	0.44424 0.0018	0.49285 0.0004	-0.28475 0.0524	0.21966 0.1379	0.03443 0.8183	0.45709 0.0012	0.19848 0.1811	0.43989 0.0020	0.62641 <.0001	1.00000	0.91213 <.0001	0.73890 <.0001
dvdplayer	0.95676 <.0001	0.56525 <.0001	0.97218 <.0001	0.84225 <.0001	0.34284 0.0183	0.64860 <.0001	-0.09672 0.5178	0.37298 0.0098	0.13502 0.3655	0.62715 <.0001	0.31360 0.0318	0.64882 <.0001	0.74812 <.0001	0.91213 <.0001	1.00000	0.86631 <.0001
cdplayer	0.80320 <.0001	0.58954 <.0001	0.84345 <.0001	0.72929 <.0001	0.20336 0.1704	0.64259 <.0001	-0.09032 0.5460	0.44445 0.0017	0.28117 0.0556	0.59013 <.0001	0.36202 0.0124	0.61041 <.0001	0.77325 <.0001	0.73890 <.0001	0.86631 <.0001	1.00000

Ci-dessus est représentée la matrice de corrélation. Cette matrice est composée des coefficients de corrélation entre deux variables, ainsi que la p valeur associée.

Cette p valeur permettra de rejeter (ou pas) l'hypothèse h0 : les deux variables sont indépendantes. Si on rejette h0, on choisira alors h1 : les variables ne sont pas indépendantes, avec un niveau de confiance de 95%. Par exemple, nous pouvons remarquer une forte corrélation entre la variable electricity et les variables television ou microwave (corrélation > 0.9), et la pvaleur indiquée en dessous de ces coefficients permet d'affirmer que nous sommes sûr à 95% que electricity et television sont corrélées car cette p valeur est inférieure à 5%.

En analysant le tableau complet, on s'aperçoit que certaines variables sont fortement corrélées avec la majorité des variables. Nous avons entouré en rouge les corrélations supérieures à 0.8 dans la partie supérieure de la matrice de corrélation, et on s'aperçoit que les variables radio et electricity sont très corrélées avec les autres variables. Cependant, d'autres variables ne présentent peu (ou pas) de corrélations fortes (mobilephone).

- Les variables les plus fortement corrélées et celles qui sont faiblement corrélées :



Le graphique à gauche correspond à la variable dvdplayer en fonction de la variable television, on remarque que la dispersion des points ressemble à une droite linéaire de type y=x, on conclue qu'il y a une forte corrélation positive entre ces deux variables et ce sont les variables les plus corrélées positivement par rapport aux autres d'après la matrice ci-dessus.

Concernant le graphique de la variable electricity en fonction de la variable chair, le nuage de point n'a pas de forme particulière. Ce sont les variables les moins corrélées.

Ces résultats confirment les analyses de la matrice de corrélation.

V. Etude multivariée

A. Comparaison entre ACP normé et ACP canonique

L'ACP canonique est basée sur la matrice de covariances, elle est caractérisée par une même unité de mesure pour toutes les variables, ce qui est le cas dans notre étude (possession). En revanche, lorsque les variables ont des variances différentes ou éloignées, celles qui ont une forte variance vont écraser les autres variables ayant une faible variance en contribuant le plus à la formation des valeurs propres des axes, ce qui pose un problème pour la représentation. C'est pour cela qu'on fait une ACP normée basée sur la matrice de corrélation afin de prendre en compte les variables ayant une faible variance.

Dans le cas où les variables ont une unité de mesure différente, on effectue également une ACP normée.

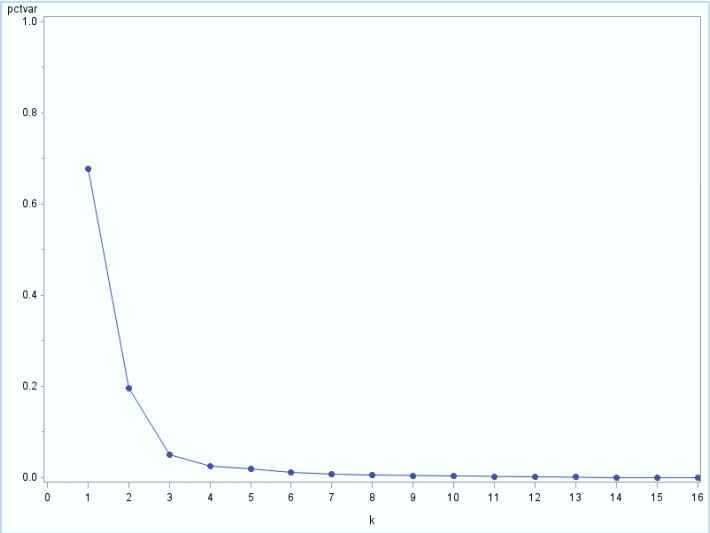
B. ACP

B.1. Détermination du nombre d'axes

Le but de l'ACP consiste à réduire la dimension de chaque observation parmi les 47 sans perte d'informations. En effet, dans notre cas, on a 16 variables quantitatives, donc chaque observation est dans R^{16}, ce qui justifie l'utilité de l'ACP en utilisant la macro variable « %acp ». Nous ne sommes pas

capables de visualiser les données dans cette dimension, on fait donc une projection de nos données sur un espace de 2D ou de 3D selon le nombre d'axes qui maximisent l'apparition de l'information.

Dans un premier temps, on cherche à déterminer le nombre d'axes nécessaires pour notre projection.



Valeurs propres, variances expliquees

k	lambda	pctvar	cumpct
1	9.00	0.56	0.56
2	3.81	0.24	0.80
3	0.87	0.05	0.86
4	0.73	0.05	0.90
5	0.51	0.03	0.93
6	0.31	0.02	0.95
7	0.20	0.01	0.97
8	0.17	0.01	0.98
9	0.12	0.01	0.98
10	0.09	0.01	0.99
11	0.06	0.00	0.99
12	0.05	0.00	1.00
13	0.04	0.00	1.00
14	0.02	0.00	1.00
15	0.01	0.00	1.00
16	0.01	0.00	1.00

Dans le cas particulier de l'ACP normée, un critère basé sur le taux d'inertie restituée, appelé critère de Kaiser, conduit à ne retenir que les valeurs propres supérieures à 1, en se basant sur l'idée que chaque axe devrait restituer en moyenne une inertie de 1 (inertie moyenne).

Dans notre cas, le critère de Kaiser nous conduit à sélectionner deux axes (2 valeurs propres supérieures à 1). On voit que l'axe 1 restitue plus de la moitié (56%) de l'inertie. Avec deux axes, on restitue 80 % de l'inertie totale, ce qui est suffisant pour prendre en compte que le premier plan factoriel.

On peut le voir également à partir du graphique à gauche correspondant à l'éboulis des valeurs propres. D'après le critère du coude, on remarque une forte diminution de k = 2 à k= 3 (apparition d'un coude).

La colonne lambda représente les valeurs propres, on rappelle que l'inertie totale est définie par la somme des valeurs propres ie $\sum_{k=1}^{p} \lambda_k = I = tr(R) = p$. Ici, $\sum_{k=1}^{16} \lambda_k$= 16 avec R la matrice des corrélations.

B.2. Représentation des individus sur les axes

Rappelons qu'on a décidé de prendre en compte les deux axes de projection u1 et u2 correspondant aux vecteurs propres 1 et 2 (respectivement) associés aux valeurs propres 1,2(respectivement) de la matrice de corrélation « R ».

On définit les composantes principales C1 et C2 comme coordonnées de projection des individus sur l'axe 1 et 2 (respectivement). En effet chaque composante principale est une combinaison linéaire des variables initiales.

Le tableau ci-dessous contient les coordonnées de projections de quelques individus sur l'axe 1 (prin1) et sur l'axe 2 (prin2) et le tableau complet se trouve en annexe (Annexe 1).

Concernant la colonne ``contg'', elle correspond à la contribution générale à l'inertie totale tandis que la colonne ``cont1'' et ``cont2'' représentent la contribution des individus à l'inertie expliquée par l'axe 1 et 2 respectivement.

Coordonnees des individus contributions et cosinus carres

ident	poids	Prin1	Prin2	contg	cont1	cont2	cosca1	cosca2
Nairobi	0.02	7.04	-5.49	10.76	11.71	16.84	0.61	0.37
Nyandarua	0.02	2.36	2.81	2.38	1.32	4.41	0.31	0.44
Nyeri	0.02	5.43	-0.15	4.32	6.98	0.01	0.91	0.00
Kirinyaga	0.02	2.56	0.62	1.22	1.55	0.22	0.72	0.04
Murang'a	0.02	2.32	1.56	1.42	1.28	1.36	0.50	0.23
Kiambu	0.02	6.46	-2.23	6.50	9.86	2.77	0.85	0.10

Pour étudier la qualité de représentation des individus sur les axes, on calcule le $\cos^2$ défini par cette formule :

$$\frac{(c_i^k)^2}{\sum_{j=1}^{p}(c_i^j)^2}$$

Plus cette valeur est proche de 1 plus la qualité de représentation est meilleure. En revanche, dans le cas où le $\cos^2$ est faible, on doit calculer sa distance à l'origine pour pouvoir conclure que cet individu est mal représenté.

Prenons l'exemple des 6 premiers individus, on voit que Nairobi contribue à environ 12% de l'inertie de l'axe 1 et à environ 17% de l'inertie de l'axe 2. On remarque également que ce comté est bien représenté sur l'axe 1 avec un $\cos^2$ de 0.61.

Les individus Nyeri, Kirinyaga, Murang'a et Kiambu sont bien représentés sur l'axe 1 et ils sont moins bien représentés sur l'axe 2.

Concernant la qualité de représentation sur le premier plan principal 1,2, il faut additionner les deux qualités de représentation sur l'axe 1 et 2. Avec ces 6 individus, on obtient les valeurs suivantes : (0.98, 0.75, 0.91, 0.76, 0.73, 0.95), on en conclut que l'ensemble de ces individus sont bien représentés sur le plan principal.

B.3. Représentation des variables sur les axes

Correlations variables x facteurs

NAME	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	v16
electricity	0.78	-0.58	-0.05	-0.01	0.06	-0.11	0.02	0.04	0.05	-0.08	0.08	-0.03	-0.03	-0.08	-0.01	0.04
radio	0.86	0.45	0.03	-0.05	0.03	0.02	0.04	0.06	-0.10	-0.07	-0.10	-0.08	0.10	-0.03	-0.02	-0.01
television	0.91	-0.35	0.04	-0.12	0.04	-0.11	-0.02	0.03	0.02	-0.05	0.07	-0.01	-0.02	0.00	0.02	-0.08
refrigerator	0.62	-0.71	-0.13	0.12	0.16	0.13	0.05	-0.10	-0.01	0.05	0.08	-0.03	0.10	0.04	0.04	0.01
landlinephone	0.25	-0.45	0.74	0.42	-0.05	0.01	-0.07	0.03	-0.04	-0.03	-0.01	0.01	0.00	0.01	0.00	0.00
mobilephone	0.88	0.27	-0.10	0.10	0.25	0.09	0.10	-0.02	-0.22	-0.04	0.01	0.09	-0.05	0.00	0.00	0.00
solarpanel	0.22	0.75	0.35	-0.19	0.44	0.10	0.02	0.07	0.13	0.06	0.03	0.00	0.00	-0.01	0.00	0.00
table	0.75	0.60	0.01	0.12	-0.16	-0.02	0.05	0.02	-0.07	0.10	0.06	-0.13	-0.07	0.02	0.01	0.01
chair	0.53	0.68	-0.09	0.29	-0.28	0.16	0.15	0.06	0.14	-0.08	0.05	0.05	0.02	0.00	0.00	0.00
sofa	0.85	0.28	0.14	-0.20	-0.21	-0.25	-0.02	0.05	-0.05	0.13	0.04	0.09	0.06	-0.01	0.00	0.01
bed	0.60	0.43	-0.29	0.49	0.21	-0.22	-0.17	-0.05	0.08	0.03	-0.04	0.01	0.01	0.00	-0.01	0.00
cupboard	0.88	0.31	0.09	-0.27	-0.05	-0.09	-0.05	-0.13	0.05	-0.13	-0.05	0.00	-0.02	0.04	0.05	0.02
clock	0.91	0.04	0.11	-0.10	-0.12	0.20	-0.09	-0.26	0.03	0.05	0.00	0.01	-0.01	-0.03	-0.04	-0.01
microwave	0.74	-0.59	-0.01	0.05	-0.01	-0.04	0.24	0.00	0.09	0.10	-0.13	0.02	-0.04	-0.02	0.02	0.00
dvdplayer	0.88	-0.44	-0.04	-0.10	0.06	-0.04	0.03	0.09	0.04	-0.03	-0.01	0.00	-0.01	0.08	-0.08	0.01
cdplayer	0.84	-0.29	-0.19	-0.07	-0.08	0.25	-0.25	0.20	0.00	0.04	-0.04	0.01	-0.02	0.00	0.03	0.01

Le tableau ci-dessus correspond aux corrélations variables-facteurs. En effet, pour chaque variable initiale, il donne le coefficient de sa corrélation linéaire avec les facteurs (composantes principales C1 jusqu'à C16).

Le premier facteur est fortement corrélé avec l'ensemble des variables à part landlinephone et solarpanel, avec un coefficient de corrélation supérieur ou égal à 0.5. On en déduit que l'axen1 permet déjà de représenter presque la plupart des variables.

Concernant l'axe 2, on trouve des corrélations négatives comme le cas des variables (electricity, refrigerator, television, landlinephone, dvdplayer microwave et cdplayer) et des corrélations positives telles que : radio, mobilephone, solarpanel, table, chair et bed. Cet axe oppose donc les variables ayant une corrélation négatives aux variables qui ont une corrélation positive données comme exemple précédemment.

On en conclut que l'axe 2 est un axe d'opposition entre les biens éléctriques et les biens représentant des possessions objets mobiliers dans un ménage. Cela nous confirme que l'accès à l'éléctricité peut être une variables importante pour séparer entre les ménages riches et pauvre au Kenya.

Obs	_NAME_	v1	v2	Qualite_axe1	Qualite_axe2	Qualite_plan
1	electricity	0.78384	-0.58388	0.61440	0.34091	0.95531
2	radio	0.86239	0.45162	0.74372	0.20396	0.94768
3	television	0.91136	-0.35085	0.83057	0.12310	0.95367
4	refrigerator	0.62121	-0.71043	0.38590	0.50471	0.89061
5	landlinephone	0.25126	-0.44592	0.06313	0.19884	0.26198
6	mobilephone	0.87590	0.26891	0.76720	0.07232	0.83951
7	solarpanel	0.21655	0.74549	0.04689	0.55575	0.60264
8	table	0.74951	0.59594	0.56177	0.35515	0.91692
9	chair	0.53196	0.67983	0.28298	0.46217	0.74515
10	sofa	0.84858	0.28269	0.72009	0.07991	0.80001
11	bed	0.60169	0.42972	0.36204	0.18466	0.54670
12	cupboard	0.87512	0.30910	0.76583	0.09554	0.86137
13	clock	0.91419	0.04056	0.83574	0.00164	0.83738
14	microwave	0.74142	-0.59435	0.54971	0.35325	0.90296
15	dvdplayer	0.87585	-0.43848	0.76711	0.19226	0.95937
16	cdplayer	0.83796	-0.29273	0.70218	0.08569	0.78787

La macro variable « %acp » ne permet pas de calculer les contributions et les qualités de représentation des variables sur les deux axes et sur le plan factoriel, mais on peut les calculer à l'aide d'une étape DATA en utilisant la matrice de corrélation entre les facteurs et les variables initiales. Il suffit donc d'appliquer les formules suivantes avec d_k^j est la coordonnée de la variable centrée réduite sur l'axe factoriel k.

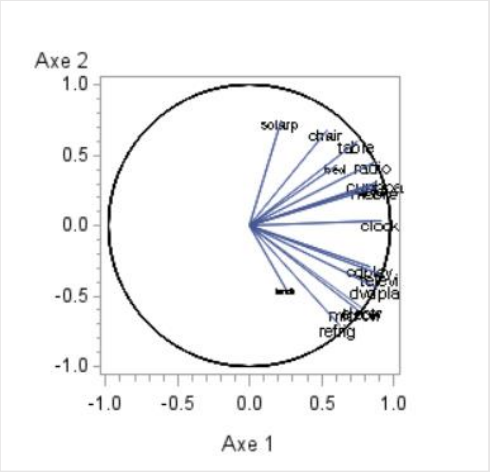
$$cos^2_{axe.k} = (d_k^j)^2 \qquad Contrib_k = \frac{(d_k^j)^2}{\lambda_k} \qquad cos^2_{plan} = (d_2^j)^2 + (d_1^j)^2$$

De même, pour la contribution sur le plan factoriel, il faut additionner les contributions sur les axes 1 et 2.

D'après le tableau, on voit que les variables electricity, radio et television sont très bien représentées sur le premier plan factoriel tandis que la variable landlinephone est mal représentée avec un cos^2 de 0.26.

Concernant la contribution (Annexe 2). Ce sont les variables electricity, réfrigérateur, solarpanel, table, chair, et microwave qui contribuent le plus à l'inertie du plan factoriel. Refrigerator et solarpanel contribuent à la formation de l'axe 2 tandis que radio, television, mobilephone et clock contribuant le plus à l'inertie de l'axe 1.

B.4. Cercle de corrélation



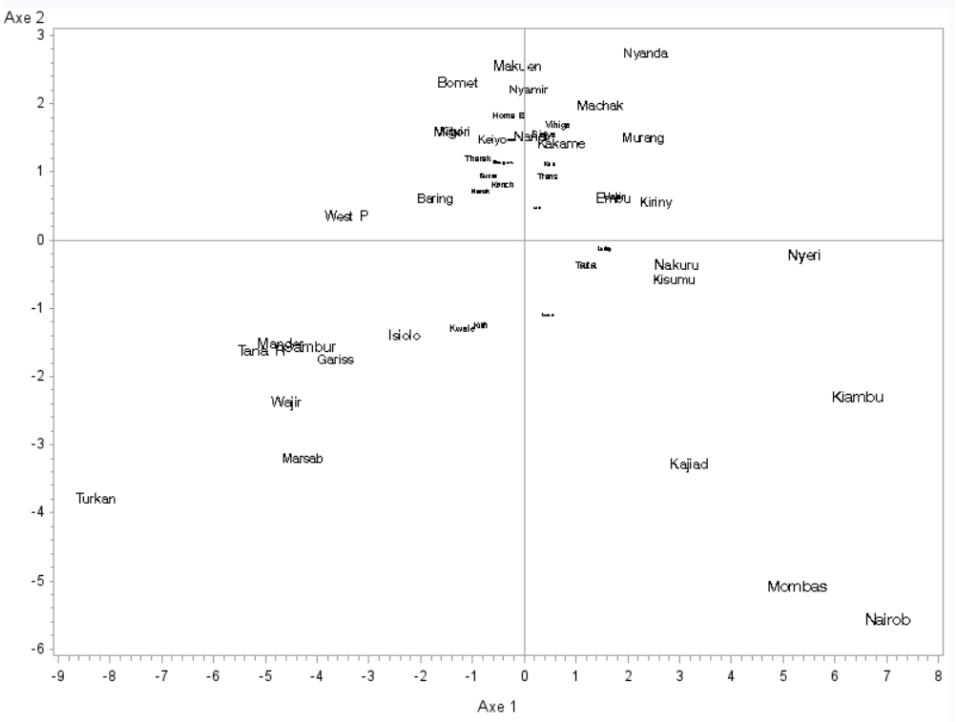
Voici la représentation des variables dans le premier plan factoriel. Nous arrivons à distinguer quelques noms de variables mais pour le reste de l'interprétation nous utiliserons également le tableau des coordonnées des variables présent en annexe (Annexe 3).

Nous remarquons que la majorité des variables sont proches du cercle à part la variable landlinephone et solarpanel et qu'aucune variable n'a de coordonnées négatives en abscisses.

De plus, en regardant l'axe 2, il apparait 2 classes différentes.

Une première pour laquelle les variables ont une coordonnée négative, et une deuxième où la coordonnée est positive. A partir du tableau en annexe, nous déduisons que la première classe correspond aux variables qui ont besoin d'une installation électrique comme television, dvdplayer ou bien refrigirator, marquant une certaine aisance financière et sociale, autrement dit ce sont des variables caractérisant des ménages plutôt riches, tandis que la deuxième classe correspond aux variables reliées aux biens matériels utilitaire, comme chair, table ou bed qu'on trouve généralement dans l'ensemble des ménages interrogés. A partir de ces informations, il est possible d'analyser la carte des individus.

B.5. Carte des individus



Avant d'analyser la carte des individus, on rappelle que l'axe 2 permet de séparer les comtés selon l'accès ou le non-accès à l'électricité tandis qu'une coordonnée élevée sur l'axe 1 signifie qu'une population interrogée a plus de biens, qu'ils soient électriques ou mobiliers.

On voit donc apparaître 4 principales catégories. La première catégorie est située dans la partie inférieure droite du plan, elle contient les comtés qu'on qualifie comme riches, ayant accès à l'électricité et beaucoup de biens électriques, c'est l'exemple de Nairobi, Kiambu et Mombasa. Tandis qu'en bas à gauche on retrouve la deuxième catégorie correspondant aux comtés ayant tout de même accès à l'électricité, mais possédant moins de biens électriques (Turkana, Wajir etc.).

La troisième catégorie dans la partie supérieure droite, correspond aux comtés ayant beaucoup de biens mobiliers mais qui généralement n'ont pas d'accès à l'électricité. On y retrouve Nyandarua et Machakos.

Revenons maintenant à la partie en haut à gauche, ce sont les comtés qui ont un nombre faible de biens matériels et n'ont pas accès à l'électricité tels que Baringo et Bomet.

En comparant la représentation des individus sur le premier plan factoriel, à la carte du Kenya, on remarque des similarités entre ces deux cartes. Les comtés situés en bas à droite de la représentation correspondent à la capitale et ses alentours (ex : Kiambu). Les comtés situés en haut de la représentation correspondent aux comtés moyennement proches de la capitale (ex : Bomet). Enfin les comtés situés en bas à gauche de la représentation correspondent aux comtés très éloignés de la capitale (ex : Turkana).

VI. Classification

A. But de la classification

La classification est l'une des approches les plus importantes pour l'exploration des données multivariées. L'objectif est d'identifier des groupes d'objets similaires dans un jeu de données selon des similarités. Nous utiliserons la classification ascendante hiérarchique (CAH) qui consiste à fournir un ensemble de partitions de moins en moins fines obtenues par regroupement successifs de parties en considérant la distance comme critère de ressemblance.

Nous allons voir, par la suite, comment regrouper les différents comtés du Kenya afin de visualiser les comtés où la pauvreté est la plus élevée, et les comtés où elle est la plus faible.

B. Procédure CLUSTER

Pour faire une classification ascendante hiérarchique des districts, on utilise la procédure CLUSTER de SAS, qui prend comme options la procédure utilisée (ici WARD), ainsi que l'ensemble des variables à prendre en compte pour la classification. Les différentes étapes de la classification sont disponibles en annexe (Annexe 4). Avec la méthode de Ward, on agrège à chaque itération les classes dont l'agrégation fait perdre le moins d'inertie interclasse.

En général, on définit l'inertie d'une classe q avec q allant de 1 jusqu'à k comme suit :

$$I_q := \sum_{i \in C^q} \frac{p_i}{p^q} \cdot \| x_i - g \|_M^2$$

p^q est la somme des p_i et g le centre de gravité. En effet, il y a deux types d'inertie, l'inertie interclasse qui permet de mesurer la séparation des classes, définie par :

$$I_{interclasse} := \sum_{q=1}^{k} p^q \cdot \| g^q - g \|_M^2$$

Et l'inertie intra classe mesurant l'hétérogénéité des classes tel que :

$$I_{intraclasse} := \sum_{q=1}^{k} p^q \cdot I_q$$

Enfin l'inertie total est définie par la somme de ces deux types d'inerties. Notre objectif consiste donc à avoir une inertie intra petite autrement dit une inertie inter plutôt élevée.

En utilisant l'option simple de la procédure CLUSTER (Annexe 5), on remarque que la majorité des moyennes des variables qui ont un rapport avec l'électricité sont faibles. En effet, la moyenne pour l'électricité est de 25,68%. Cette valeur est petite par rapport à des pays développés. Concernant l'écart-type, il est environ égal à 0.18. L'écart type étant élevé, cela montre une inégalité de répartition de l'électricité au sein du pays.

Concernant l'option standard, elle permet de normaliser les variables afin que les variables aient le même poids lors du calcul des distances entre les individus.

On voit qu'au début, l'inertie interclasse contenue dans la colonne R carré est élevée (0.999) car chaque individu forme une classe mais il diminue au cours des étapes jusqu'à devenir nul à la dernière étape de classification où il n'y a qu'une seule classe.
Concernant la colonne R carré semi-partiel, elle correspond à la perte d'inertie interclasse à chaque regroupement de deux classes. En effet, lors de la formation du cluster Keyo et Nandi, il y a eu une perte de 0.0007 d'inertie.

Rappelons que notre objectif est de maximiser l'inertie interclasse et de minimiser l'inertie intra classe, c'est pour cela qu'on mesure le rapport entre les deux dans la colonne pseudo F carré. En effet, plus le rapport est élevé, plus l'inertie interclasse est maximale, d'une façon équivalente, l'inertie intra classe est minimale. On voit qu'au début, le pseudo F carré vaut 31.5 et après il déminue. Quant à la colonne pseudo T carré, elle mesure la séparation entre les deux clusters les plus récemment fusionnés.

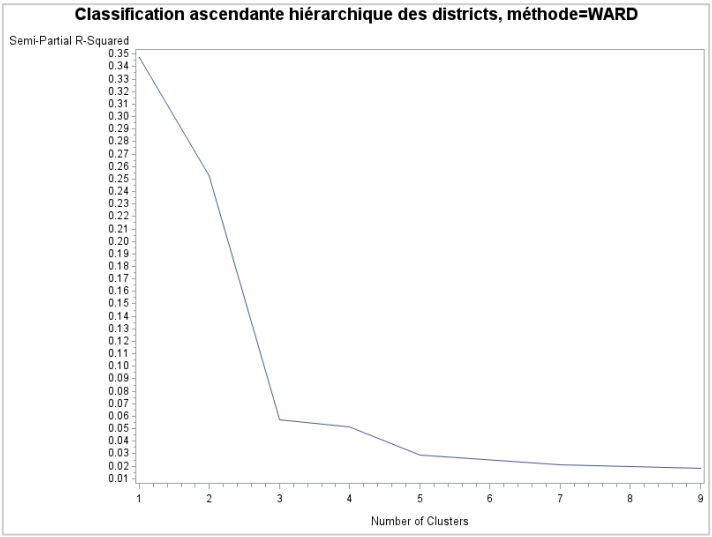
C. Choix du nombre de classes

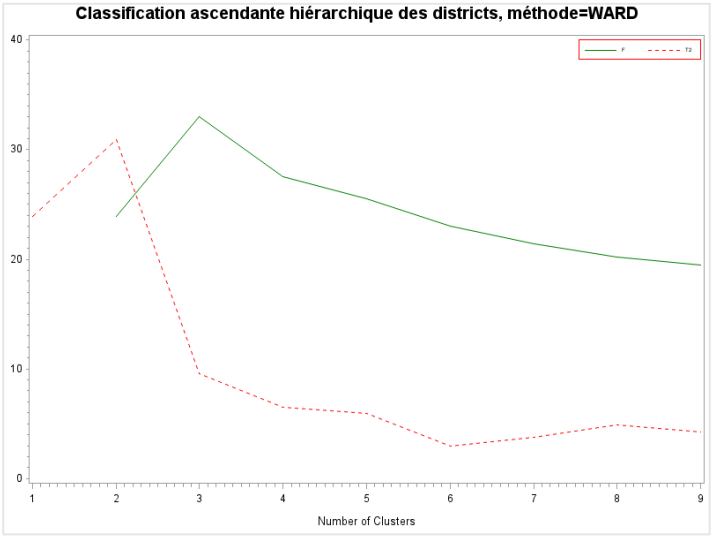
Une grande valeur de pseudo F indique que les individus au sein d'une classe se ressemblent, alors que les classes entre elles sont hétérogènes. En particulier, un pic lors de la représentation du pseudo F en fonction du nombre de classes est un bon indicateur pour la séparation des classes.

Un pic pour un nombre de clusters égal à k sur le graphe de pseudo T^2 indique que les clusters fusionnés afin d'obtenir k clusters au total sont très différents. On prendra donc k+1 classes pour notre étude.

Enfin, le critère R^2 semi partiel mesure la perte d'inertie interclasse provoquée en regroupant deux classes. On recherche une forme de coude, indiquant un changement brutal d'inertie interclasse, tout en indiquant un nombre de classes le plus faible possible afin de synthétiser nos données.

Ci-dessous sont affichés les graphes des trois critères mentionnés précédemment.





En utilisant le critère du pseudo F, on constate un pic pour un nombre de clusters égal à 3. Concernant le critère du pseudo T^2, le pseudo T^2 prend des valeurs faibles à partir de k=3, alors que nous constatons un pic pour k=2. Ce critère nous indique également qu'on doit choisir 3 classes. Enfin pour le critère R^2 semi partiel, le coude apparait pour k=3. On retient donc 3 classes.

Les trois classes sont réparties selon le graphe d'analyse de classification disponible en annexe (Annexe 6). Ce graphe est donné par la procédure CLUSTER, mais il est également possible d'obtenir un graphe similaire avec la procédure TREE.

Il est possible d'obtenir la liste des comtés contenus dans chaque classe avec la procédure TREE. On remarque alors que la classe 1 contient 33 comtés, la classe 2 contient quant à elle 9 comtés, et la classe 3 contient 5 comtés.

D. Interprétation des classes

Dans un premier temps, nous pouvons regarder les moyennes des variables selon les classes (Annexe 8). Ces moyennes représentent une première approche afin de comprendre à quoi correspond chaque classe.

La classe 3 semble avoir des moyennes supérieures à la moyenne des autres classes pour la plupart des variables notamment l'électricité, mobilephone, table et bed. Cette classe possède donc beaucoup de biens et le niveau de vie est plutôt élevé. En comparant ce résultat avec la carte de Kenya obtenue par la procédure GMAP (Annexe 7), on remarque que cette classe correspond à la petite zone géographique verte située dans le sud-ouest du pays.

Concernant la classe 2, les moyennes sont nettement plus faibles. Cela signifie que les foyers sont plutôt pauvres. Elle correspond à la zone géographique rouge qui est majoritaire sur la carte. Quant à la classe 1, les moyennes des variables sont globalement proches de la moyenne. Les foyers habitants dans les comtés de la classe 1 sont des foyers qui ne sont ni pauvre ni riches, possédant des biens utilitaires. Cette classe est représentée en bleu sur la carte du Kenya.

E. Caractérisation des classes

Nous pouvons également utiliser le tableau des valeurs tests (Annexe 9). Il s'agit d'une statistique de test de comparaison de paramètres calculés dans le sous échantillon associé au groupe et dans la totalité de l'échantillon : Pour analyser ce tableau, on rappelle qu'une valeur positive pour une variable indique une forte proportion du bien correspondant au sein d'un échantillon. A contrario, une valeur négative indique une faible proportion de bien.

Commençons par la classe 3. Elle contient le comté Nairobi, ainsi que des comtés relativement proches. Ces comptés possèdent des biens de confort comme l'électricité, la télévision, le refrigirateur, le micro-onde, le DVD Player et le CD Player avec des proportions supérieures à la moyenne.

Concernant la classe 1, celle-ci correspond aux comtés en marge de la capitale, mais pas totalement éloignées. On remarque la présence de certains bien utilitaires, principalement des tables, lits et chaises, et également une haute valeur pour les panneaux solaires, indiquant que ces comtés ont tout de même un accès à l'électricité et donc ne sont pas des comtés extrêmement pauvres.

Enfin, la classe 2 correspond aux comtés très éloignées de la capitale. Ces comtés sont très pauvres. Les foyers vivant dans ces comtés ne sont en général pas du tout équipés, même pour des biens les plus basiques, tel que des lits ou des tables.

VII. Synthèse

Cette analyse nous permet de donner une première réponse à la question « Est-ce que la localisation géographique d'un comté a une influence sur la pauvreté ? ».

Après avoir analysé les variables de manières individuelle, l'analyse par composantes principales a montré que certains comtés étaient plus riches que d'autres. Les foyers au sein de ces comtés riches sont mieux équipés. De plus, en regardant s'il était déjà possible d'identifier un potentiel lien entre la pauvreté d'un foyer et le comté dans lequel il était, on s'aperçoit que les comtés les plus riches sont réparties autour de la capitale, tandis que les plus pauvres sont éloignés. Nous avons alors décidé de regrouper les comtés en classes, en fonction de la pauvreté des foyers interrogés. Nous avons obtenu 3 classes, chacune correspondant à un niveau de richesse et de confort différent. La visualisation de ces classes sur la carte du Kenya a pu bel et bien confirmer l'hypothèse du départ, à savoir que la pauvreté d'un foyer dépend de sa situation géographique.

Cependant, un facteur important n'a pas été pris en compte. Le poids accordé aux comtés n'est pas intervenu dans nos calculs, et nous avons supposé que le nombre de personnes interrogées fût le même dans tous les comtés. Or, ce n'est pas le cas, et il est donc difficile de se rendre compte de la situation dans les comtés les plus riches, et également dans les comtés les plus pauvres. Une étude plus poussée aurait pris en compte ce paramètre, afin d'avoir une meilleure précision au niveau des comtés. De plus, au sein de notre étude, certaines variables et individus étaient mal représentés, il serait donc intéressant de refaire une étude en changeant nos axes de représentations afin d'obtenir de meilleurs résultats pour ces variables et individus.