Rapport de projet: Analyse d'un jeu de Donnée de l'Application Spotify

Lakhdar Othmane & Linon Romuald

19 Novembre 2019

Contents


```r
library(cluster)
library("gplots")
library("FactoMineR")
library("factoextra")
library(ade4)
library(ggplot2)
library(cowplot)
library(readr)
library(dplyr)
library(lattice)
library(caret)
library(dendextend)
```

1 Introduction

Tous les ans, Spotify, une application de service de piste audio, sort le top 100 des pistes audios les plus écoutées sur l'application. Le jeu de données que nous étudierons ici à pour individus le top 100 des pistes audio de l'année 2018. Ce jeu de données (disponible sur le site kaggle.com), comporte 3 variables qualitatives (auteur, mode, clé) ainsi que 11 variables quantitatives.
Il est important de préciser que l'auteur du jeu de données à uniquement regroupé les observations des individus sur les variables. En d'autres termes, les variables proviennent toutes de la platforme offcielle de Spotify web API. Le but de ce rapport est de montrer, les liens entre les variables ainsi que les indépendances que l'on peut trouver. On pourra ainsi determiner ce qui rassemblent ces Pistes audios, mais aussi ce qui les différencient. Ce rapport est structuré de la façon suivante: Nous commencerons d'abord par de l'analyse univariée avec la transformation d'une variable qualitative permettant ainsi une analyse plus digeste, puis nous enchainerons sur de l'analyses multivariées(ACP, classification, AFC), puis après avoir filtrer les informations concernant les liens entre les variables, et éliminer de potentielles redondances qui alourdiraient les interprétations, nous procederons à de l'analyse bivarée (Test de Fisher/χš, représentations graphiques, modélisation linéaire).

2 Remarques

2.1 Concernant les méthodes d'analyses

Dû au nombre de variables dont nous disposons ainsi que du nombre maximal de pages autorisées nous ne pouvons pas nous permettre d'appliquer toutes les méthodes d'analyse de données sur toutes les combinaisons de variables. Ainsi nous "filtrerons" ces dernières celons une certaine logique qui sera explicité lors des premières parties.
Du fait que spotify produit aussi des pistes audio de type podcast, donc pas nécessairement des musiques, nous devrions considérer, dans une optique formelle et inclusive, les individus comme des pistes audios. Cependant après avoir parcourus les individus, il est clair que chacun d'entre eux sont des musiques, nous les qualiferons alors de la sorte.
Á noter, certaines variables comme liveness et duration_ms(qui ne figure pas dans ce rapport mais sont présent sur le jeu de données) rendent l'ACP trop dense à analyser. En leur présence, nous passons de 3 cercles à 10 cercles de corrélations toutes contenant des informations au moins faible certe mais toute fois présente. Ainsi ces variables peuvent être vu comme des "bruits" qui nuise à la synthése de l'ACP, nous avons donc décidé de les supprimer du jeu de données et donc de notre champ d'étude.

2.2 Clarfication de certaines variables

Bien que certaines variables soient des mesures standards du domaine musical (tempo, mode), d'autres sont inhérantes à l'application spotify.
Par exemple d'après spotify, les variables Danceability, Valence, Energy representent l'ambiance générale que dégage la musique.
De plus speechness, loudness, instrumentalness traitent la présence ou non de moyens musicaux techniques.

Plus d'informations sont à votre disposition dans le tableau ci-dessous:
NB:Ces définitions ont recours à des termes techniques de la musique. Bien que nous ne rentrerons pas dans les détails de chacuns des termes, il reste necessaire d'avoir de brèves notions de ces derniers, avant de lire le tableau:

Sonie ou bruyance (loudness): valeur numérique qui représente le volume sonore tel que perçu par l'être humain. Sur une chaîne de haute-fidélité, un poste de télévision, un smartphone, et autres appareils électroacoustiques, le volume sonore est la sonie du son produit.
acoustique: dans le cas d'un instrument de musique, qui n'est pas amplifié par des equipements électroniques.
Tempo:la vitesse d'exécution d'une œuvre ou plus exactement la vitesse de la pulsation, ce battement régulier « qui sert d'étalon pour construire les différentes valeurs rythmiques »(par un métronome par exemple). Son unité de mesure est le Battement par minute (BPM).

Mode:est la mélodie appartenant aux mode dit-majeur ou au mode dit-mineur utilisé dans une œuvre
Beat: terme anglais désignant une pulsation se répétant régulièrement, fournissant la base d'un pattern (d'un chemin) musical.
Mesure: la mesure est une segmentation de la durée du discours musical. En d'autres termes, la mesure est la division d'un morceau de musique en parties d'égales durées.
Key: (clé en français) Dans la notation musicale, une clef ou clé est un signe graphique placé au début de la (partition) qui indique !la hauteur des notes associées à chaque ligne.

Nom de la variable	Nom dans le code.R	valeur pris par les variables	Définitions
danceability	da	$[0; 1]$	Décrit le fait qu'une musique soit adaptée pour la dance (une valeur de 1/0 indiquant que la chanson l'est/l'est pas)
energy	en	$[0; 1]$	Represente la "forte et intense activité" d'une musique, exemple de style de musique: le heavy metal. Une valeur de 1 indique une musique au sonorité s'enchainant rapidement avec un volume élevé, et 0 dans le cas contraire
key	key	$A, .., G, +A*, C*, D*, F*, G*$	clef de la chanson.
loudness	loud	$[-60, 0]$	sonie ou bruyance moyenne de la musique En décibel (dB).
mode	mode	$mineur; majeur$	indique le mode (majeur/mineur) de la musique.
speechiness	spee	$[0; 1]$	Detecte la présence de mots parlés dans la chanson. Une valeur entre 1 et 0.66 indique une musique probablement composée uniquement de parole, entre 0.66 et 0.33 cela correspondrait à un mélange de sons et de paroles,et en dessous de 0.33 representerait plutôt à une musique sans langage.
acousticness	ac	$[0; 1]$	mesure de confiance permettant de determiner si le son de la musique est acoustique/ou non (se rapprochant de 1/0)
instrumentalness	inst	$[0; 1]$	indique si la musique comporte des voix. Plus la valeur se rapproche de 1 plus il est probable que la chanson ne contienne PAS de sons vocals.
valence	val	$[0; 1]$	décrit la positivité d'une musique. Plus la valeur se rapproche de 1/0 plus la musique à des sonorités positives/négatives (joyeux, euphoriques, ...)/(triste, deppresive, colérique).
tempo	temp	$[0; +\infty]$	donne le tempo,"la rapidité", en battement par minute de la musique.
time-signature	ts	$\mathbb{N}$	*signature rythmique* en français, c'est une notation qui indique le nombre de battement dans une mesure

Voici donc les variables que l'ont va étudier tout au long de ce rapport

```
colnames(data)
```

```
## [1] "author" "da"     "en"     "key"    "loud"   "mode"   "spee"   "ac"
## [9] "inst"   "val"    "temp"   "ts"
```

3 librairies utilisées

```
library(cluster)
library("gplots")
library("FactoMineR")
library("factoextra")
library(ade4)
library(ggplot2)
library(cowplot)
library(readr)
library(dplyr)
library(lattice)
library(caret)
library(dendextend)
```

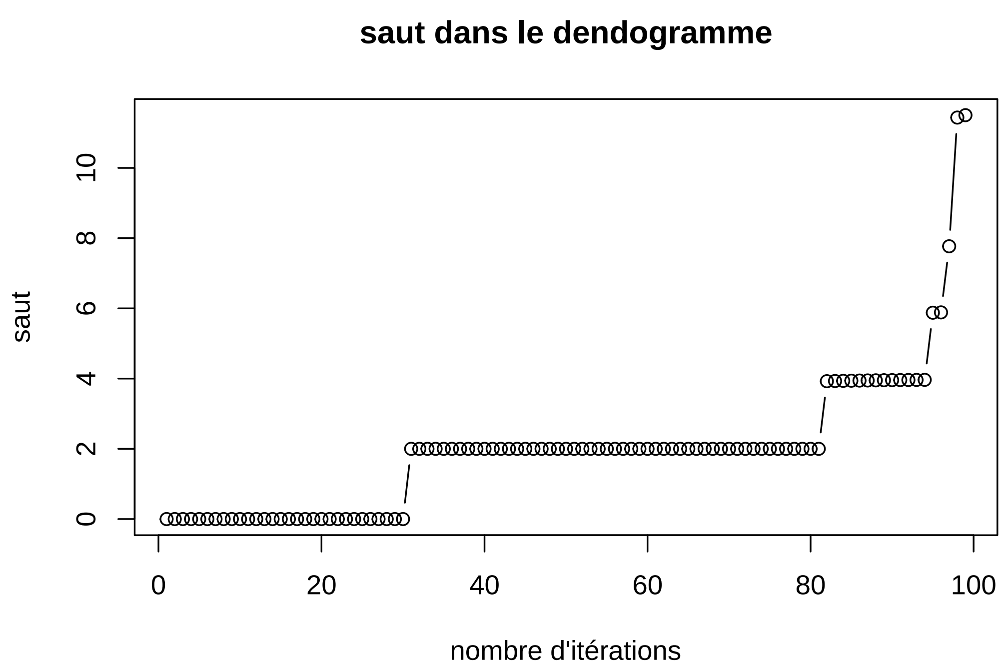
4 Analyses univariées

Dans notre jeu de données, la variable **author** présente beaucoup trop de modalités pour être étudiée :

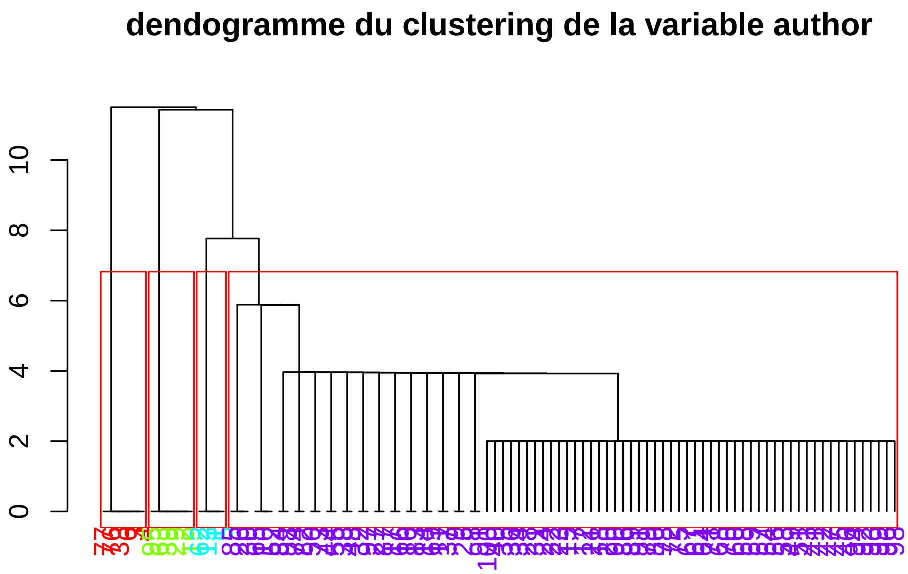
```
length(levels(author))
```

```
## [1] 70
```

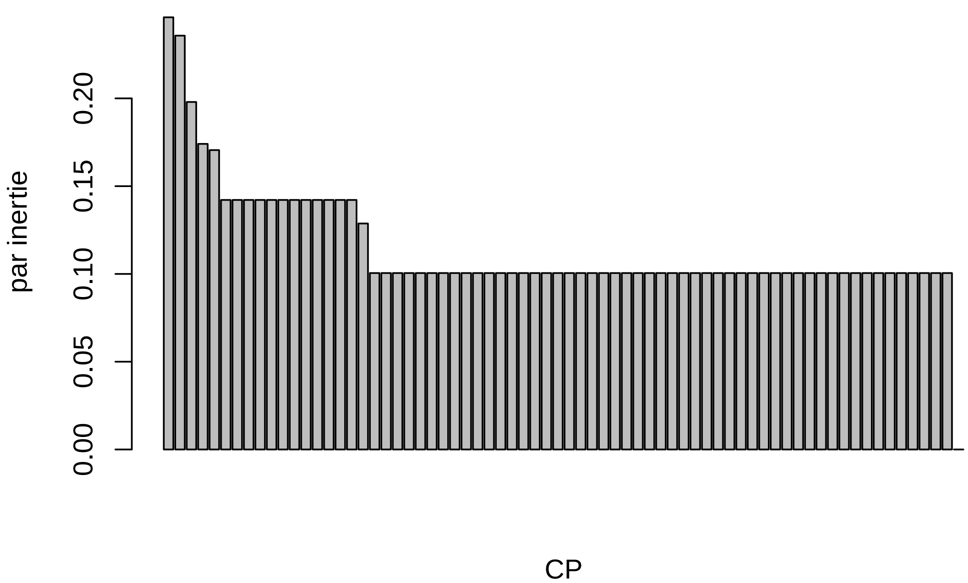
Nous entamons alors dans un premier temps une classification hierarchique des individus par rapport à cette variable qualitative, dans le but de réduire de façon optimale le nombre de modalité (i.e en perdant le moins d'inertie).Nous auront recours par ailleurs à "transformer" cette variable qualitative en une variable quantitative via la méthode du *One-hot encoding* (vû aussi en cours).

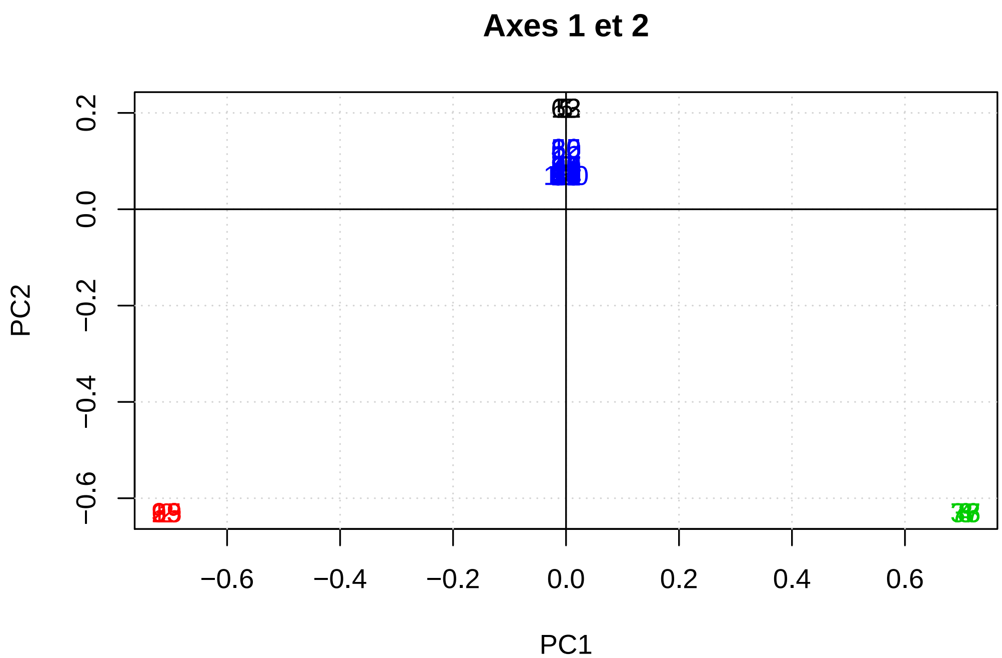


D'après le graphique le dernier saut sgnificatif se fait après la 97 ème itération. Nous choisissons alors de décomposer notre dendogramme en 4 parties, celle que crée cette interation (qui constituerons les 4 futures modalités de la variable).



Maintenant, une question se pose : est-il légitime de choisir une telle décomposition ? Comme dans le TP n°7, nous répondrons par une ACP.





D'après le graphe représentant les distances des coordonnées sur le plan factoriel des axes 1 et 2 de l'ACP, qui sont les plus signifcatifs, le clustering est clairement satisfaisant, Nous remarquons aussi que les agglomerations sont extrement éloignées (sauf pour les agglomerations bleu et noir) ce qui temoigne d'une différence notable de chaque modalité . Nous gardons donc un tel clustering et décomposerons donc **author** et les 4 modalités associées

```
levels(author)
```

```
## [1] "A1" "A2" "A3" "A4"
```

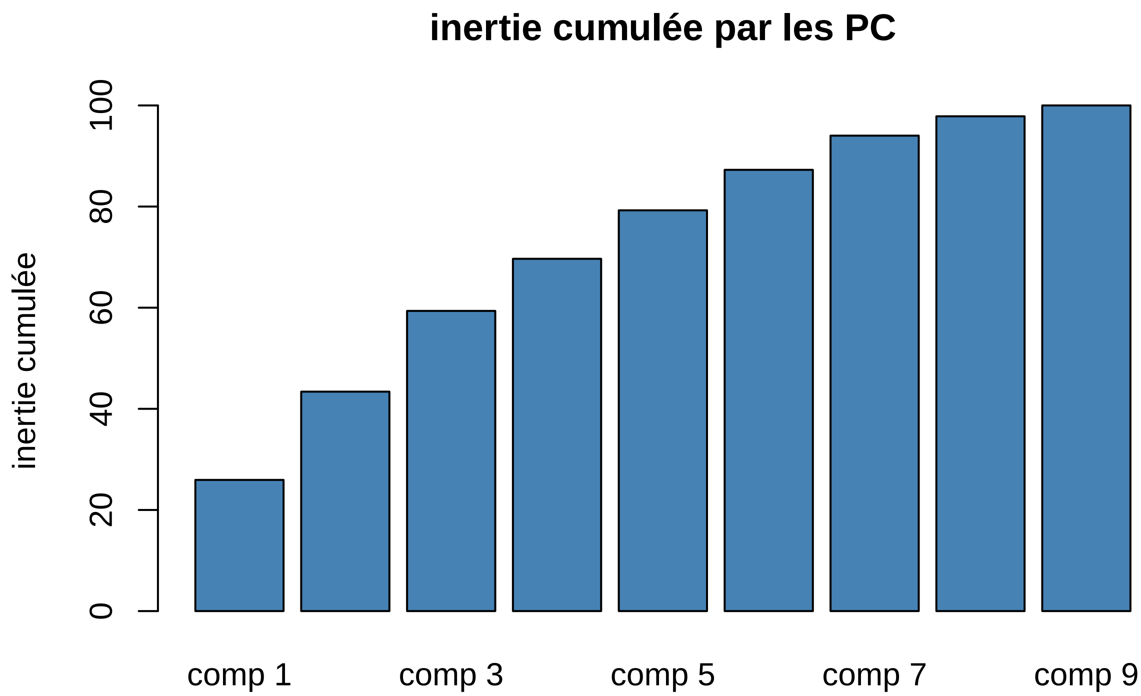
5 Analyses multivariées

Dans cette section, le but est d'analyser les jeux de données dans son ensemble, afin d'avoir une vision globale des liens entre les variables.

5.1 Analyse en composante principal(ACP)

5.1.1 Mise en place de l'ACP et détermination du nombre de composantes à garder

```
#Réalisation de l'ACP
quanti_data = data[,-c(1,4,6)] #on extrait les variables qualitatives uniquement
acp = PCA(quanti_data, scale = TRUE, graph = FALSE) #acp centrée réduite
info = get_pca_var(acp) #apporte plusieurs info comme le cos2 et les contributions des variables

  #Choix du nombre d'axes :traitement des variances(cumulées)/vp
c_variance = acp$eig[,3]
barplot(c_variance,col = "steelblue", main  = "inertie cumulée par les PC",ylab = "inertie cumulée")
```



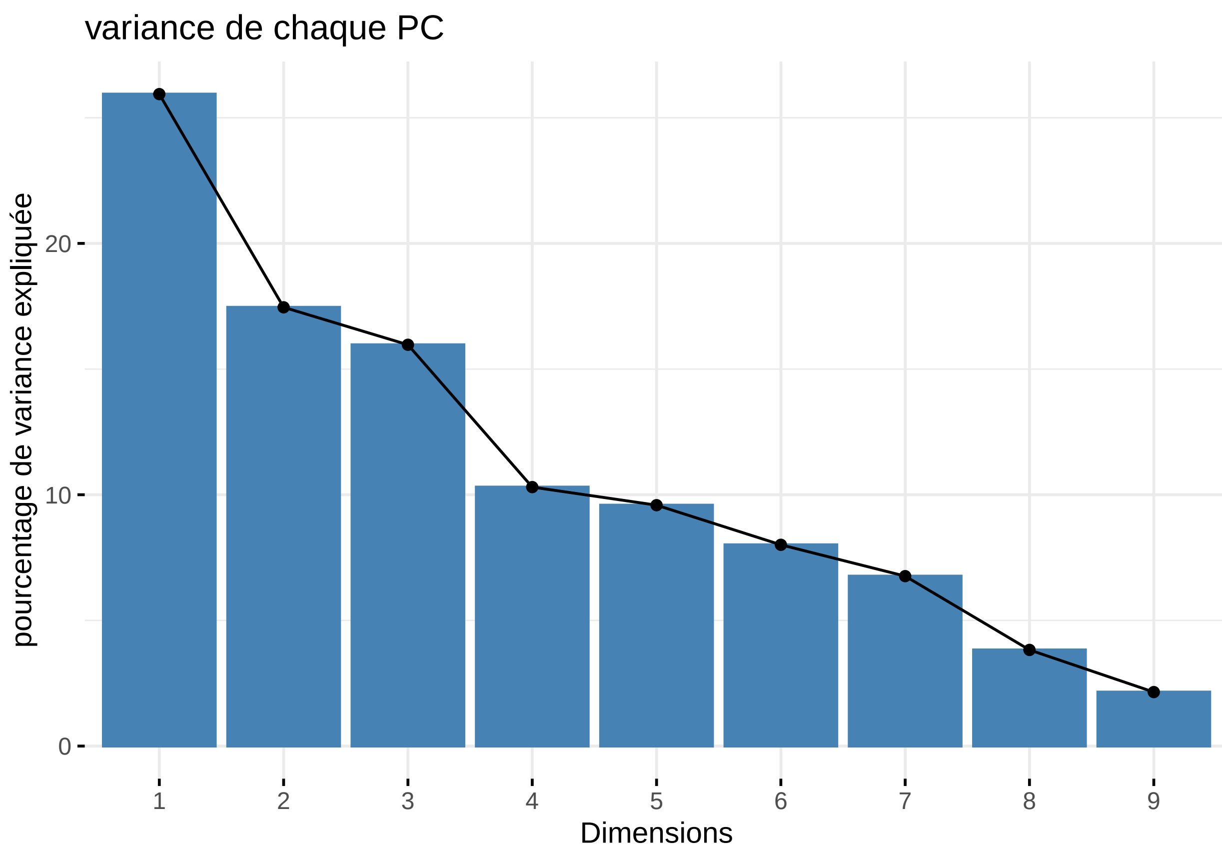
Le *summary* de l'ACP et les graphes nous indiquent clairement une absence de saut significatif ce qui rend difficile la determination du nombre de composantes principales à garder.
Après des recherches personnelles, nous avons trouvé d'autres méthodes afin de determiner le nombre optimal de PC à choisir. La règle de kaiser-Guttman mais aussi la règle KSS (pour Karlis-Saporta-Spinaki).
La règle de Kaiser-Guttman repose sur le fait que la moyenne des valeur propres vaut 1. Ainsi nous ne prennons que les CP ayant des valeurs propres supérieures à 1 (i.e dans notre cas les composantes participant à plus de 11.1% de la inertie totale). Si nous considèrons cette régle comme adéquate pour notre cas, nous prendrons les 3 premières CP.
la régle KSS, en revanche, est plus exigente en terme de proportion de l'inertie total, cette exigence, sur les valeurs propres, dépend notamment du nombre n d'individus (ici 100), et du nombre de variables p considéré (ici 9). la régle KSS ne prend que les composantes principales dont la valeur propre est supérieure à $c = 1 + 2\sqrt{\frac{p-1}{n-1}}$ en faisant l'application numérique on obitent $c = 1,57$ (ie une participation d'au moins 17,4% à l'inertie totale).

```
  #information numérique des graphiques
acp$eig
```

```
##        eigenvalue percentage of variance cumulative percentage of variance
## comp 1  2.3345949              25.939943                          25.93994
## comp 2  1.5711978              17.457754                          43.39770
## comp 3  1.4370616              15.967351                          59.36505
## comp 4  0.9273923              10.304358                          69.66941
## comp 5  0.8626832               9.585369                          79.25477
## comp 6  0.7207578               8.008420                          87.26319
## comp 7  0.6085753               6.761947                          94.02514
```

```
## comp 8  0.3443431                3.826035                          97.85118
## comp 9  0.1933941                2.148823                         100.00000
```

```
fviz_eig(acp,ncp = 9,main ="variance de chaque PC", ylab ="pourcentage de variance expliquée")
```



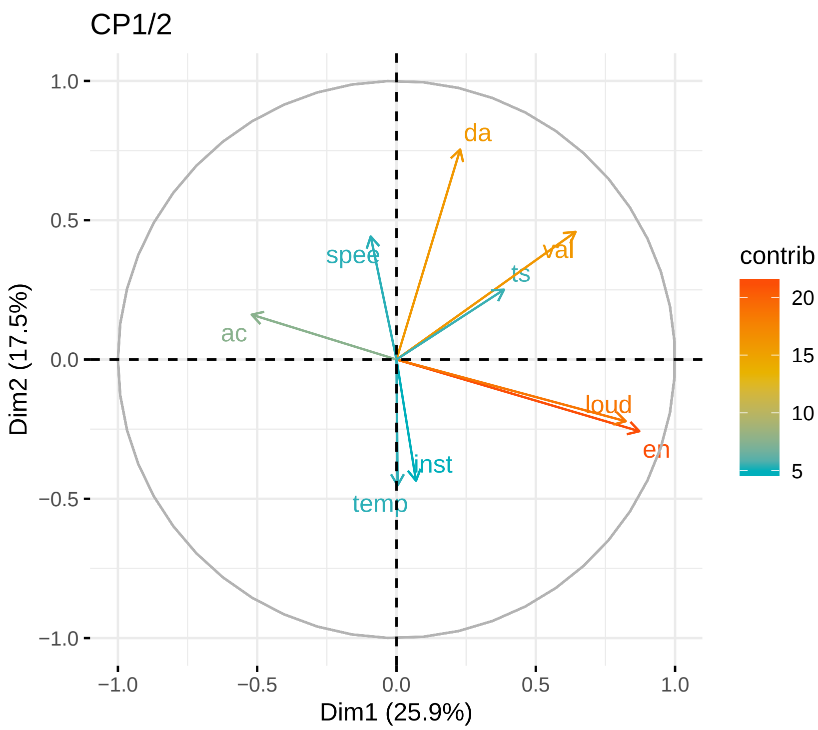
Ainsi, avec la régle KSS, nous prennons les 2 premiers axes. Neamoins le couple (PC1, PC2) n'explique que 43.40% de l'inertie totale ce qui est insuffisant. De ce fait nous optons pour la régle de Kaiser-Guttman et choisissons donc les 3 premiers axes (qui couvre 60% de l'inertie totale).

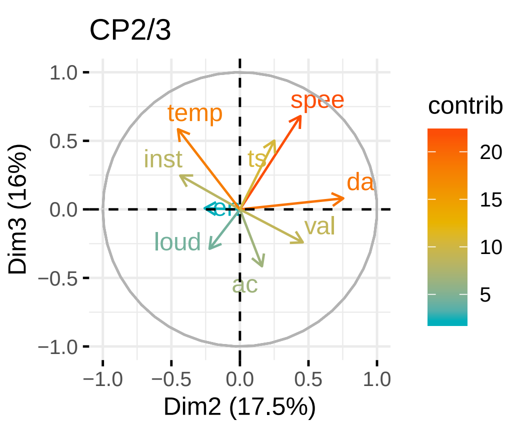
5.1.2 Cercles de correlation, contributions des variables et interpretations

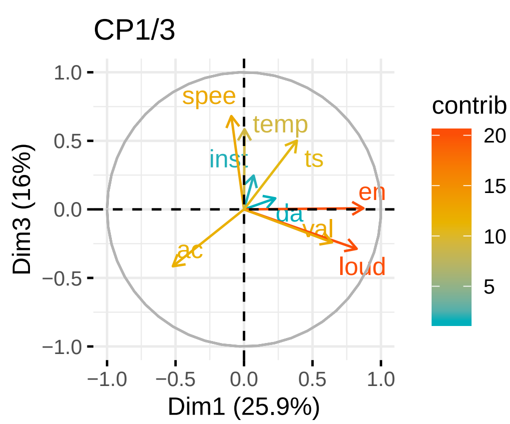
Notre professeur de TP nous a conseiller d'analyser, parmis, les axes gardés, l'ensemble des cercles de corrélations et de retirer les cercles qui présentent une pauvre contribution des variables, ou qui contiennent des d'informations (indépendance entre variables , corrélation positive/négative) peu sgnificatives. Cependant dans chaque cercles nous trouvons au moins deux variables avec des contributions acceptable (au dessus de 10%).

Nous générons ainsi les cercles de corrélation à l'aide la fonction suivant de de plot_grid :

```
Corc = function(i,j, titre){
  fviz_pca_var(acp,axes = c(i,j), col.var = "contrib",gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07")
}
```








```
info$contrib[,1:3]#plus explicite à mon sens
```

```
##              Dim.1     Dim.2       Dim.3
## da    2.237073e+00 36.171429  0.450576746
## en    3.253708e+01  4.219231  0.004788051
## loud  2.899958e+01  3.125771  5.758872404
## spee  3.675754e-01 12.413362 32.245917590
```

```
## ac   1.157532e+01  1.648929 11.987416138
## inst 2.098339e-01 12.056245  4.235782804
## val  1.769897e+01 13.365651  4.027433233
## temp 6.586585e-04 12.998876 23.785693455
## ts   6.373917e+00  4.000506 17.503519579
```

5.1.2.1 Descriptions des CP par les variables d'origines, et des CP

Nous commençons d'abord par une représentation des composantes principales, à partir des variables utilisées pour l'ACP, $C_{(i,j)}$ sera la notation du cercle de corrélaton de (CPi, CPj).
Dans le cercle $C_{(1,2)}$ et nous constatons que les variables **en**, **loud**, **val** et **da** ont de fortes contributions (supérieures à 10%). Dans ce cercle, nous étudierons uniquement ces variables.
Les variables **en**, **loud** et **val** sont fortement et négativement corrélées à CP1, tandis que CP2 est fortement et positivement corrélée à **da** et légerement à **val**.
Dans le cercle $C_{(1,3)}$, les variables **loud**, **en**, **val**, **ac**, **spee**, **ts**, et **temp** ont des contributions supérieures à 10%.
Concernant CP1, en plus des informations dans le cercle $C_{(1,2)}$, **ac** est positivement corrélée à CP1. CP3 quant à lui est négativement corrélées à **ts**, **temp**, **spee** et est corrélée à **ac** positivement .
Pour le cercle $C_{(2,3)}$ **temp**, **spee**, **da** ont des contributions supérieures à 10%.
CP2 est négativement corrélée à **temp** , tandis qu'en plus, CP2 est positivement corrélée à **spee**. CP3 quant à lui contient les mêmes informations dans ce cercle que $C_{(1,3)}$.

<u>En résumé :</u>
légende : +/-(**.**): compte tenu d'une CP, la liste des variables qui lui sont positivement/négatvement corrélées
~: suivie d'une variable; décrit un rapport moins prononcé de la corrélation (positive/négative).
CP1: +(~**ac**), -(**loud**,**en**,**val**)
CP2: +(**da**, ~**val**, ~**spee**), -(~**temp**)
CP3: +(~**ac**) -(**spee**,**temp**,**ts**)

Ainsi, nous pouvons conclure que CP1 peut être vu comme étant un axe indiquant une musique à l'intensité sonore faible, et avec des sonorités exprimant la tristesse; CP2 en revanche se comporte comme un indicateur de fiabilité d'une musique potentiellement chantée, comme étant propre à la dance et à la bonne humeur avec tempo bas . Enfin, CP3 peut se représenter comme un indicateur d'une musique au rythme bas, probablement non parlée avec une acoustique relativement prononcée.

5.1.2.2 Indépendance, corrélations positives et négatives entre les variables d'origine

independance: Concernant les corrélations positives nous avons, pour chaque cercles les couples :
legende: (**.**,**.**) := couple qui sont décorrélée
~ := suivie par un couple, décorrélations les moins prononcées

$C_{1,2}$: (**da**, **loud**), (**da**, **en**),
$C_{1,3}$: (**temp**, **en**)
$C_{2,3}$: (néant)

positive: Concernant les corrélations positives nous avons, pour chaque cercles les couples :
legende: (**.**, **.**) := couple ayant une corrélation positive
~ := suivie par un couple, corrélation positive les moins prononcées

$C_{1,2}$: (**en**, **loud**) , ~(**da**, **val**)

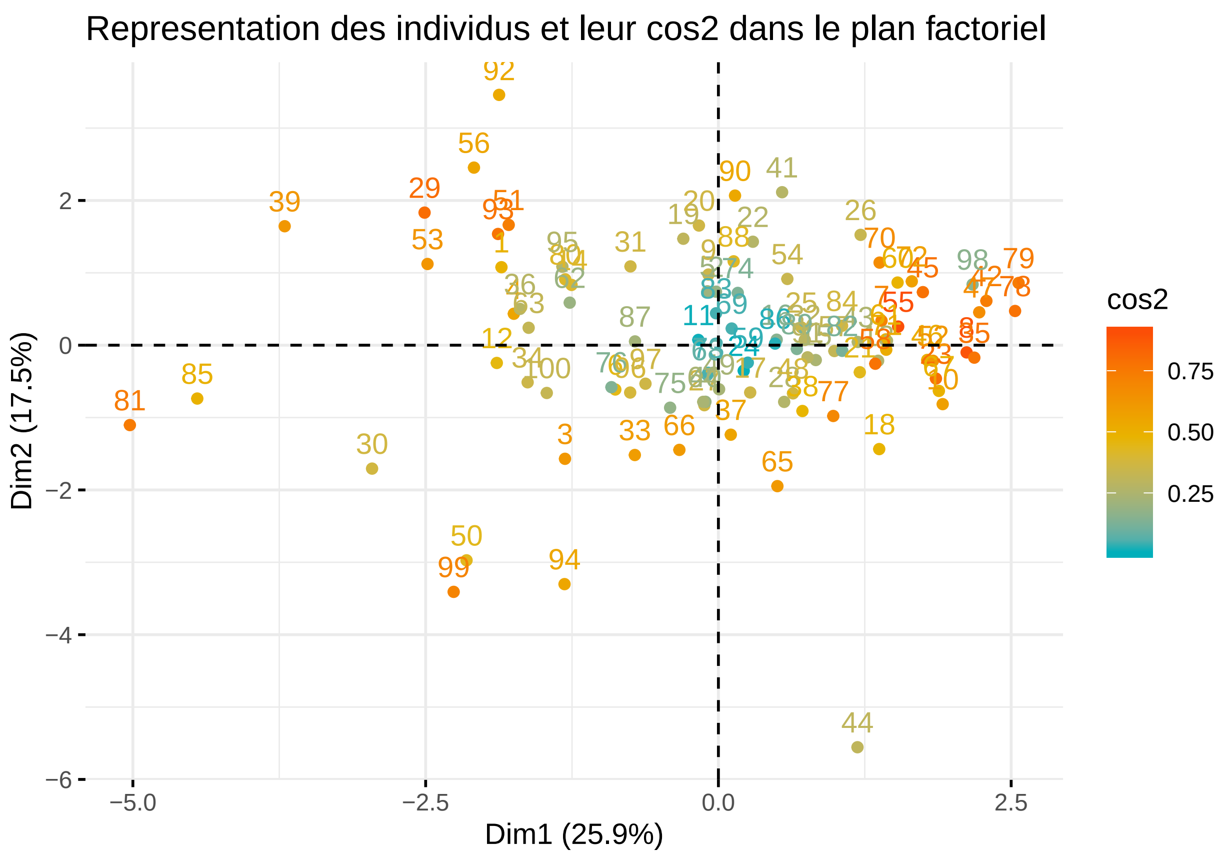
$C_{1,3}$: (**loud**, **val**),(**spee**, **temp**), ~(**loud**, **val**,**en**), ~(**spee**, **temp**,**ts**)
$C_{2,3}$: ~(**da**,**spee**)

négative: Concernant les corrélations négative nous avons, pour chaque cercles les couples :
legende: (**.**,**.**) := couple ayant une corrélation négatives
~ := suivie par un couple, corrélation positive les moins prononcées

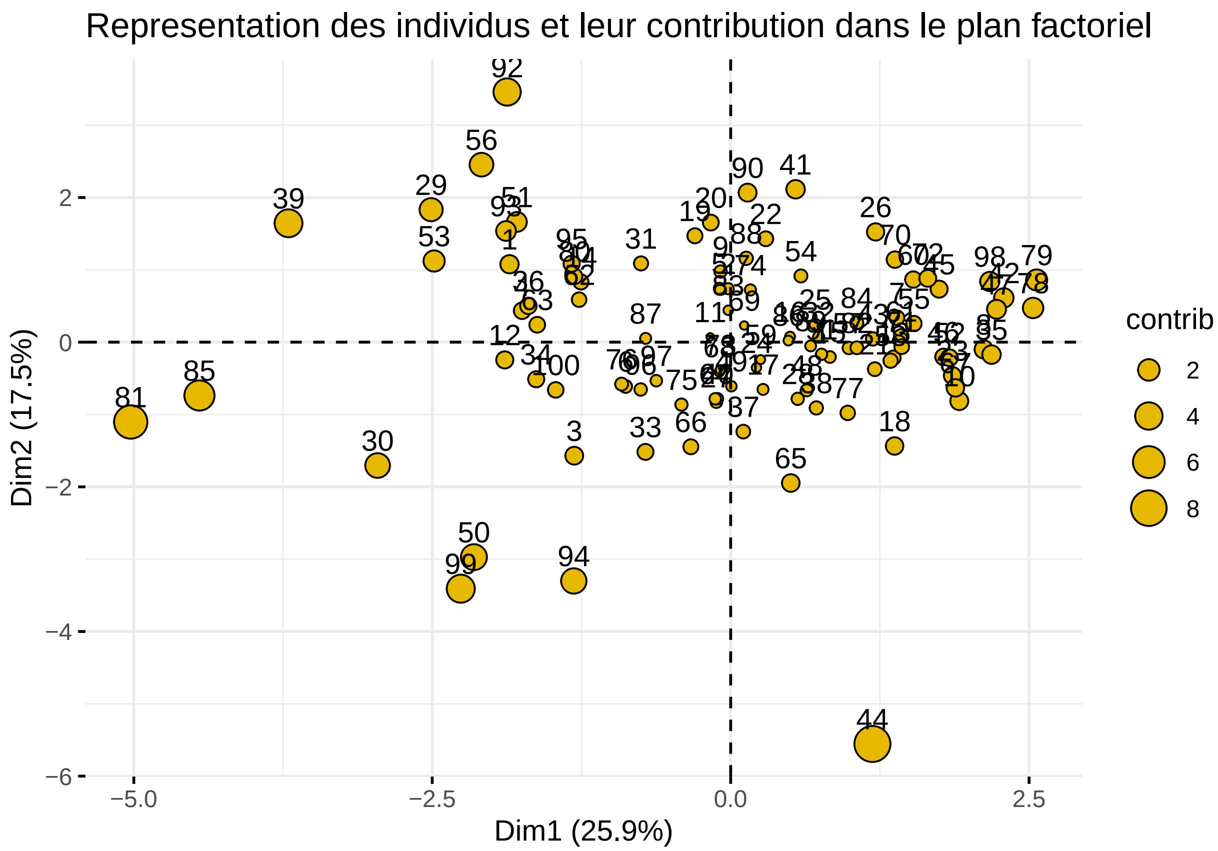
$C_{1,2}$: (néant) , (**ac**,**loud**)
$C_{1,3}$: (**ac**, **ts**)
$C_{2,3}$: (néant)

5.1.3 Plan factoriel

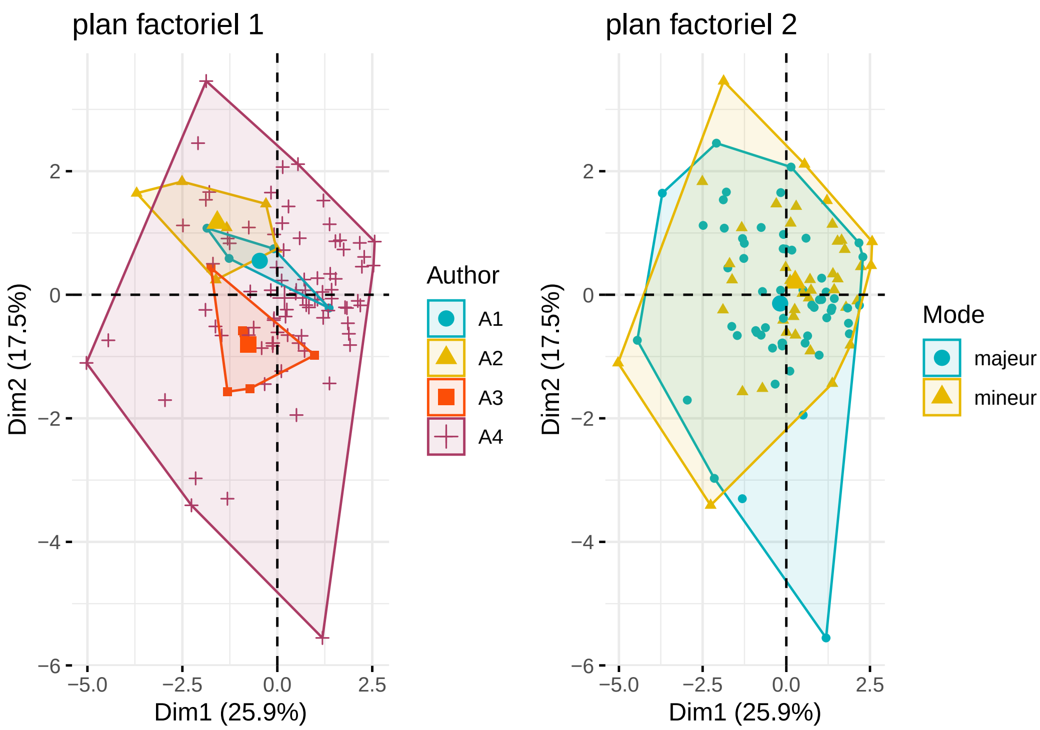
Dans cette sous section, nous allons, par souci de page restante, étudier le plan factoriel engendré par CP1 et CP2 uniquement.



Dans ce graphe (ou plan factoriel), on constate une concentration non négligeable dans la partie droit, avec un cos2 très important(autour de 0.75 dans la plus part des cas),ces derniers sont fortement (et positivement) représenté par l'axe 1. En revanche des individus tels que 81 et 85 qui sont à la partie gauche du graphique, sont peu nombreux et moins denses, leur cos2 en revanche reste "décent" (autour de 0.50).
Une telle observation est reproductible pour l'axe 2 en parcourant le graphe de haut en bas(faible concentration dans les 2 cas). Evidemment, on constate aussi une forte concentration au centre du graphe, mais avec un cos2 de plus en plus bas lorsque l'ont se rapproche du centre, ce qui rend peu exploitable cette zone.
Dans le prochain graphique nous traiterons les contributions de ces zones pour une analyse plus complète.



En reprennant les observations pour le graphe précedent, les contributions aux axes ne sont pas étonnants. On constate des contributions très faibles au centre du plan factoriel (au alentour de 2% en général), car proche du centre et donc avec des valeurs proche de la moyenne. À l'inverse, en peripherie de cette agglomération les contributions sont beaucoups plus élevées. C'est notamment le cas pour les individus 81, 85.



Sur ces deux graphes, les répresentations du plan factoriel nous prennons en compte, par un marquage de formes et de couleurs des points, les modalités respectives des variables **author** et **mode**, prise par chaque individu. Dans le graphe 1 on remarque les points ayant la modalité A4 est d'une part étendue, et d'autre part est plutot concentrée dans la partie superieure droite du graphe, elle à donc une forte dispertion de "ses" (en terme de modalité) individus. Á l'inverse, les individus des modalités A1,A2 et A3 sont agglomérées au centre respectivement dans la partie gauche, haut-gauche, basse .
notez d'ailleurs que les individus de la modalité A1 présente une variance plutôt faible de ses individus pour l'axe 2.
En se rappelant des définitions de ces deux axes nous pouvons caractériser **globalement** les indvidus de chaque groupe. Par exemple, nous pouvons dire que le groupe des individus marqués par A1 sembleraient se caractériser par une musique probablement chantée incitant à dancer.les individus de A2 disposent des mêmes caractéristiques que ce de A1 avec la particularité d'être **encore plus** probablement chantée. En revanche les individus de A3 sont des musiques aux sonorités plutôt fortes et joyeuses, avec une certaine acoustique et un tempo plutot élevée. De par l'étendu de A4 en revanche nous pouvons étendre aucune caractéristique au global.
Le graphe 2 en revanche présente peu d'information à extraire, on constate peu de différences entre les individus des modalités de la variable, le brassage entre individus des modalités est trop important pour une quelconque analyse des individus des modalités.

5.2 Classification hierarchique

Nous allons Procéder à la classification hierarchique des individus. Pour se faire,nous allons nous servir de l'ACP faite ulterieurement et utiliser la fonction HCPC (disponible dans la librarie *factoMineR*).

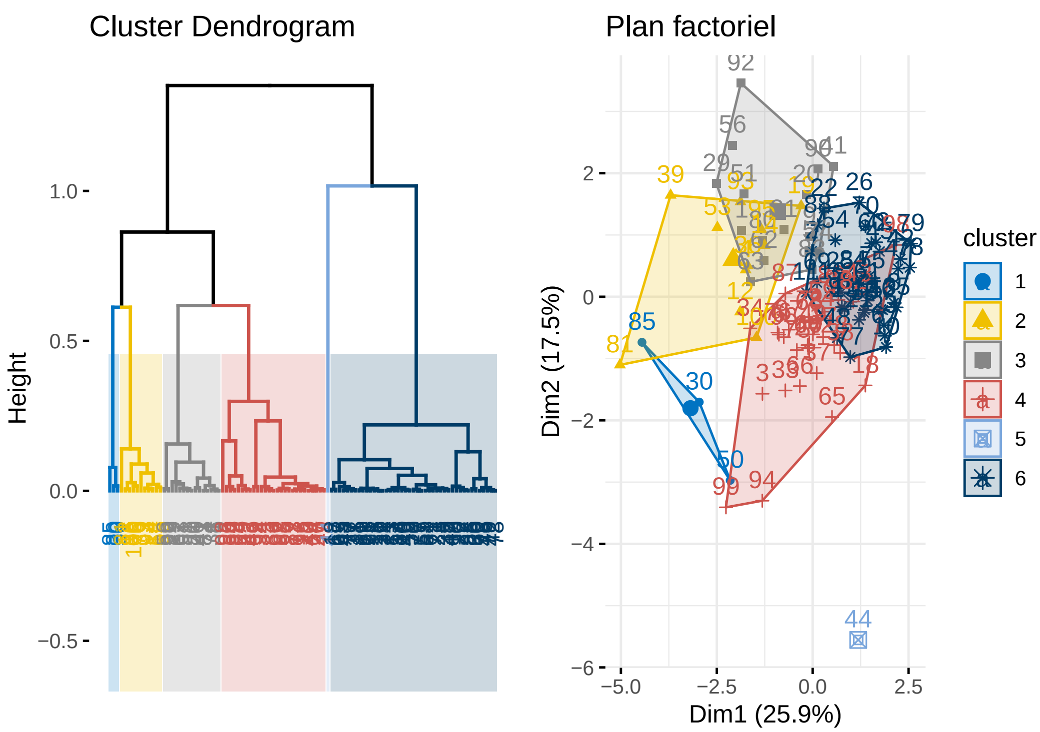
5.2.1 Mise en Place de la classification

```
clust = HCPC(acp, graph = FALSE)
head(clust$data.clust)
```

```
##      da    en   loud   spee     ac     inst    val     temp ts clust
```

```
## 1 0.754 0.449 -9.211 0.1090 0.0332 8.29e-05 0.357  77.169  4     3
## 2 0.740 0.613 -4.880 0.1450 0.2580 3.72e-03 0.473  75.023  4     6
## 3 0.587 0.535 -6.090 0.0898 0.1170 6.56e-05 0.140 159.847  4     4
## 4 0.739 0.559 -8.011 0.1170 0.5800 0.00e+00 0.439 140.124  4     2
## 5 0.835 0.626 -5.833 0.1250 0.0589 6.00e-05 0.350  91.030  4     3
## 6 0.680 0.563 -5.843 0.0454 0.3540 0.00e+00 0.374 145.028  4     4
```

5.2.2 Dendogramme et plan factoriel: interpretations



La première figure est le dendogramme associé à la hierarchisation appliquée précédemment. Il complète le graphique par la hauteur de chaque groupe, mais dans ce cas, les hauteurs sont relativement identiques. Ce graphe est le plan factoriel de l'ACP réalisée par la séction précedente. Le marquage observé correspond à chaque groupe produit par la hierarchisation (clustering), que nous nommerons $C_1, ..., C_6$. Ce clustering semble être conculant, on ne constate aucun brassage trop important (qui serait manifesté par un chevauchement trop important des zones colorées).les groupes entourent le centre du plan. Ainsi les différents groupes ont une relation particulière avec CP1 et CP2, typiquement C_1 est négativement lié au deux axes tandis que C_3 est positivement liées à PC1 mais la liaison est mitigé avec CP2 (le cluster est à la fois sur la partie positive et négative de CP2).

5.3 Analyse Factorielle des correspondances (AFC)

Dans cette section nous entammerons une AFC avec les variables **author** et **key**, Pour se faire nous utiliserons la fonction CA et d'autres fonctions associées, des librairies factoextra et factoMineR.

5.3.1 Mise en place de l'AFC

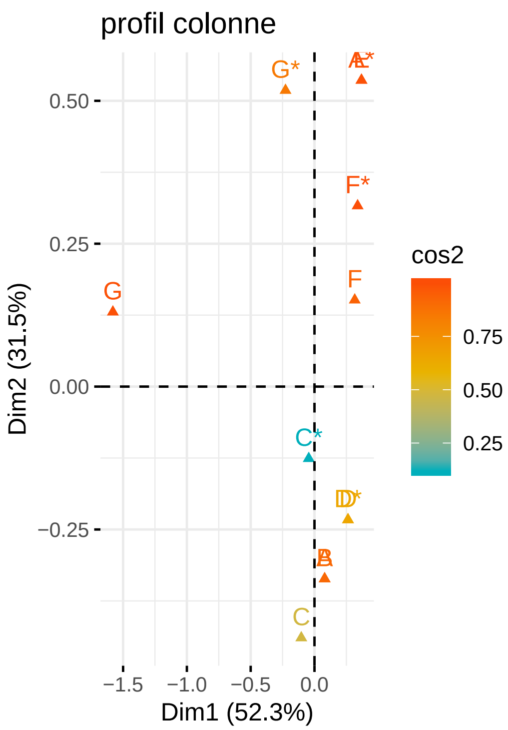
```
##        key
## author  G G*  F C* A*  A  D F*  C  B  E D*
##     A1  2  1  0  1  0  0  0  0  0  0  0  0
##     A2  1  1  0  0  0  1  0  0  2  1  0  0
```

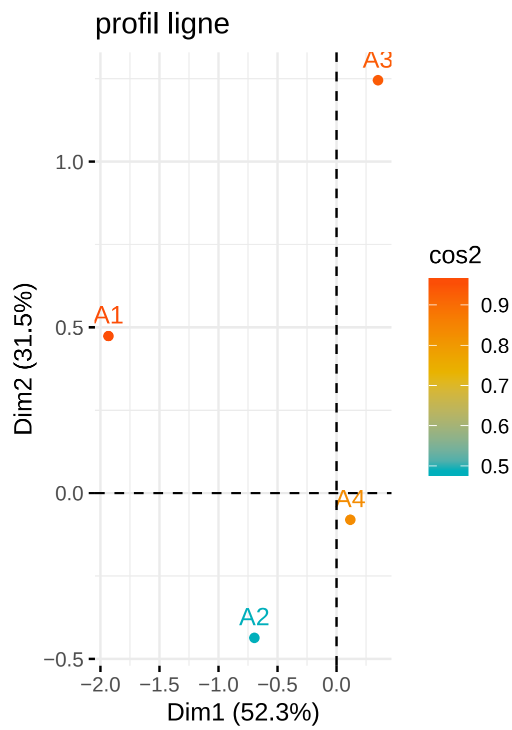
```
##     A3  0  2  1  0  1  0  0  1  0  0  1  0
##     A4  3  7  9 14  4  9  7  6  8  9  4  4
```

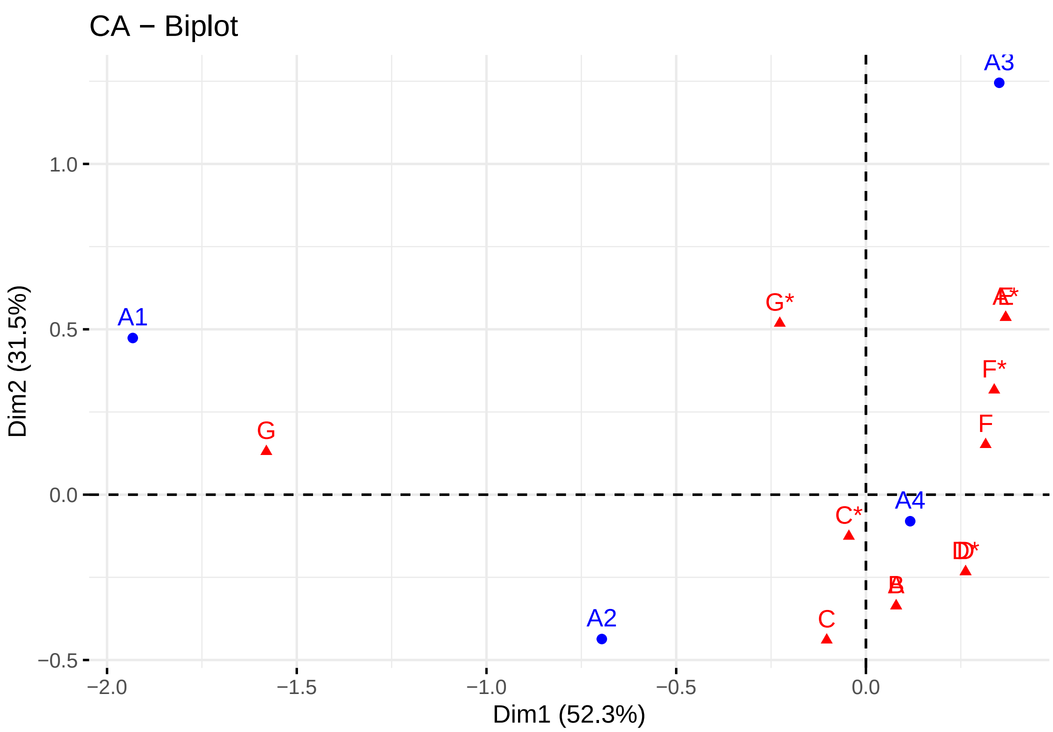
```
##
##  Pearson's Chi-squared test
##
## data:  table
## X-squared = 37.703, df = 33, p-value = 0.2628
```

```
##       eigenvalue percentage of variance cumulative percentage of variance
## dim 1 0.19722440               52.31065                          52.31065
## dim 2 0.11887943               31.53088                          83.84153
## dim 3 0.06092153               16.15847                         100.00000
```

5.3.2 Plan factoriel: cos2, et interpretations







Le dernier graphe est une généralisation des deux suivants. Ils mêlent les modalités de la variable **author** et de **key**. Analyser les 2 premiers et les comparer revient à analyser ce graphe.
le premier graphe présente les modalités de la variable **key** dans le plan factoriel,c'est donc le profil colonne. Nous remarquons d'abord que certains points se chevauchent sur ce plan. Donc, les auteurs qui appartiennent respectivement au deux modalités sont très proches. De même concernant les différences, l'axe 1 oppose G à

la majeure partie des autres modalités. Ainsi, la modalité G est composée d'un profil d'individus différents des individus des autres clefs. Concernant le cos2 des modalités, elles ne sont pas choquantes, les modalités les plus au centre ont un cos2 faible (autour de 0.25), tandis que les plus éloignées ont un cos2 élevé (0.75 ou plus).
Quant à l'axe 2, il oppose les clefs C, B,A et dans une moindre mesure D,D#(groupe 1) aux clefs F,F#,A# et E(groupe2). Ainsi les individus des modalités du groupe 1 sont proches entre elles et se différencent des individus des modalités du groupe 2 ,qui sont eux (le groupe) aussi proches.
Concernant le second graphe, on constate une parabole passant par le nuages de points. Ceci témoigne d'un effet Gultman: **key** et **author** semblent être redondante. Les différences constatées sont au niveau des modalités (A1,A4),(A2,A3). Néanmois A2 présente un cos2 bien moins faible que A2 alors que cette dernière est plus proche du centre.
En confrontons les deux axes, la clef G est le plus souvent utilisée par le groupe d'artiste A1, tandis que A4 est le plus proche d'une variété de clefs. Ceci est expliqué par le fait que lors de la premiere hierarchisation, fait pour regrouper les modalités de la variable **author**, le 4ème groupe comportait le plus d'individus, et ce, de manière significative. Enfin A3 et A2 sont les plus isolés de toutes les clefs, de ce fait on peut supposer que ces groupes d'artistes varient beaucoups leurs choix de clefs dans leur musique.

6 Analyses bivariées

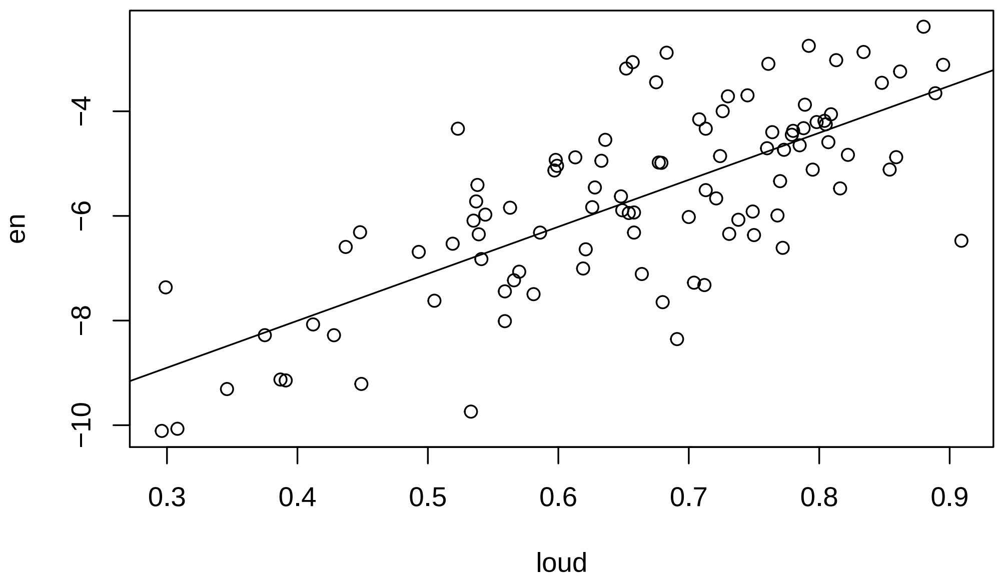
6.1 Couple variables quantitatives

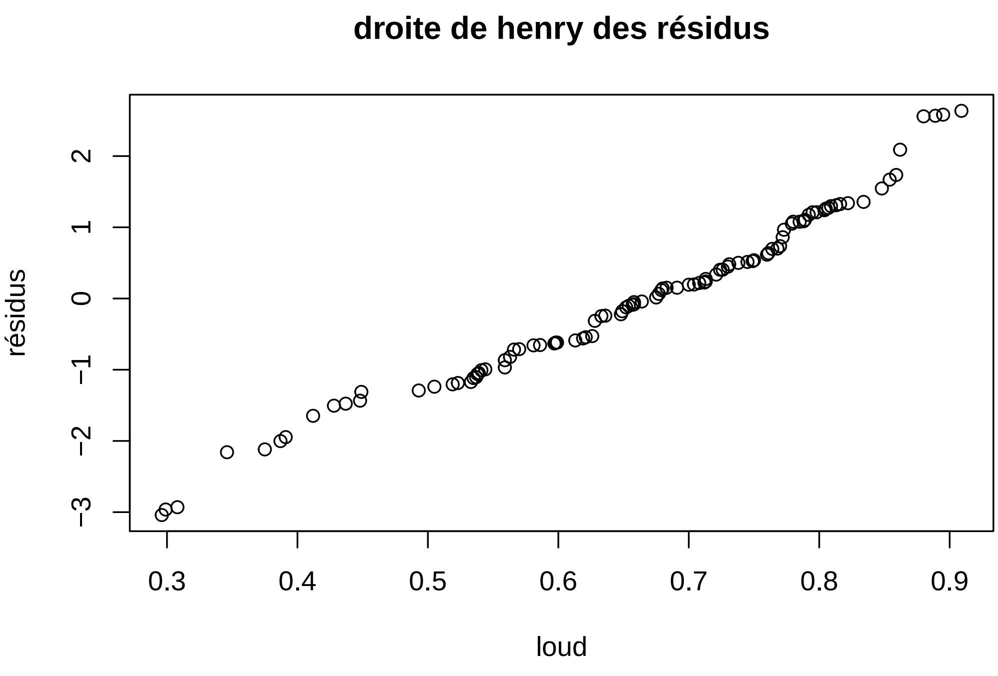
Dans cet sous-section, nous étudierons les liens entre les couples de variables quantitatives repéré notamment lors de l'ACP. Pour chaque rubrique, nous noterons la variance des résidus $var(e)$ pour générer les plots nous avons utilisé cette fonction :

```
#Test de fisher:
linear_mod = function(x,y,xnames, ynames){
  d = lm(y~x)
  plot_grid({plot(x,y,xlab = xnames,ylab = ynames)
    abline(d)},
            qqplot(x,d$residuals,xlab = xnames,ylab = "résidus", main ="droite de henry des résidus"))
  return(summary(d))
}
```

6.1.0.1 Variables Loud/en

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.04041 -0.83194  0.09185  0.88666  2.63514
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.5949     0.5683  -20.40   <2e-16 ***
## x             8.9784     0.8424   10.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.216 on 98 degrees of freedom
## Multiple R-squared:  0.5369, Adjusted R-squared:  0.5322
## F-statistic: 113.6 on 1 and 98 DF,  p-value: < 2.2e-16
```

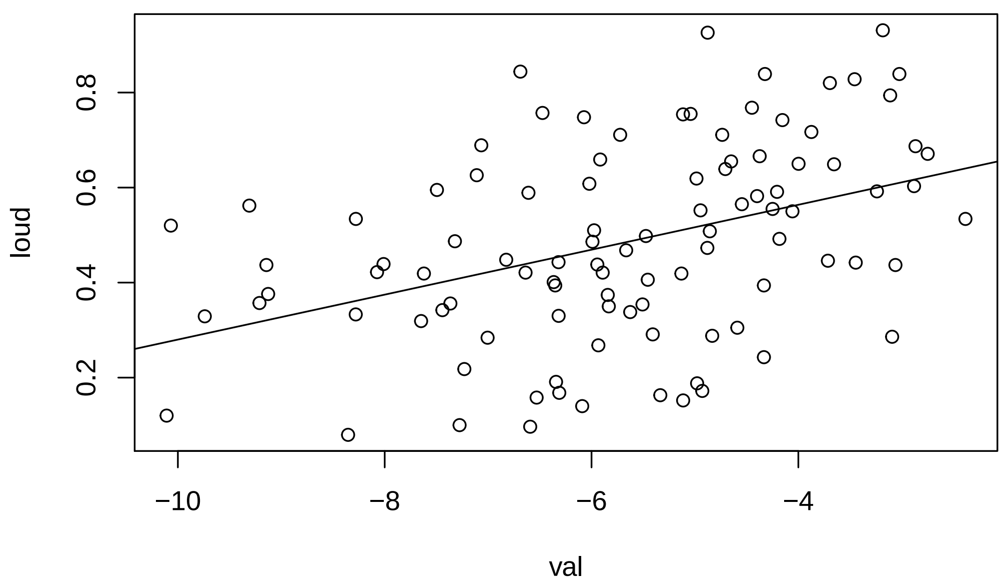


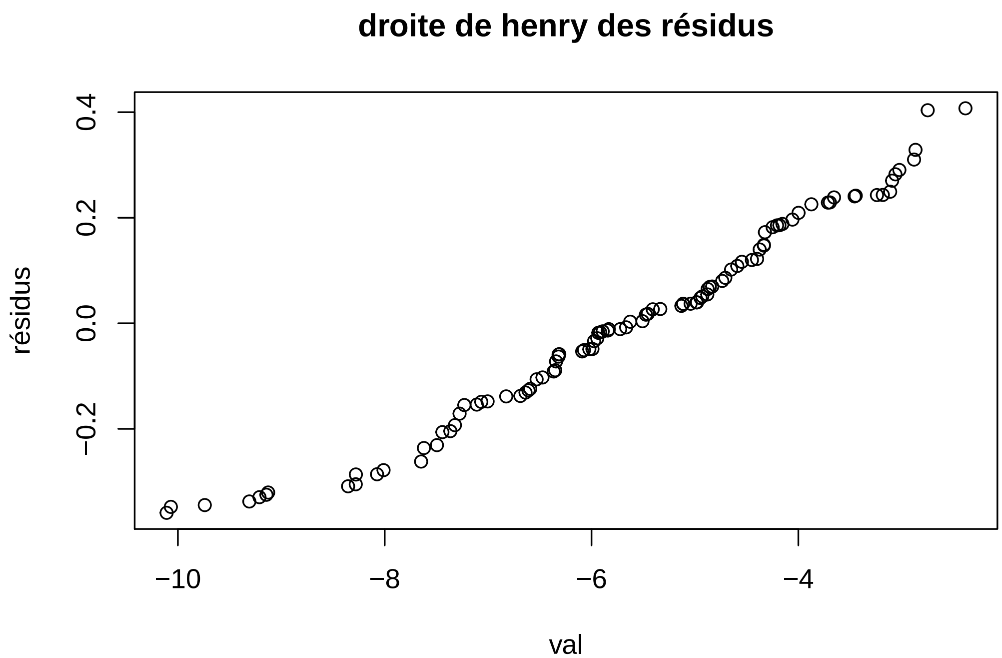


Au vu de la valeur de F-statistic et de la p-valeur, le test de fisher montre un lien linéaire entre les 2 variables. Par la droite de henry des résidus qui montre une distribution gausienne de ces derniers par rapport aux valeurs de **loud** le test de fisher semble valable. le plot du modéle linéaire et de la dispertion de **en** en fonction de **loud** est donc concluant

6.1.0.2 Variables Loud/val

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35910 -0.13289  0.00346  0.14166  0.40733
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.75293    0.06361  11.837  < 2e-16 ***
## x            0.04729    0.01070   4.421 2.54e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1892 on 98 degrees of freedom
## Multiple R-squared:  0.1663, Adjusted R-squared:  0.1578
## F-statistic: 19.54 on 1 and 98 DF,  p-value: 2.544e-05
```

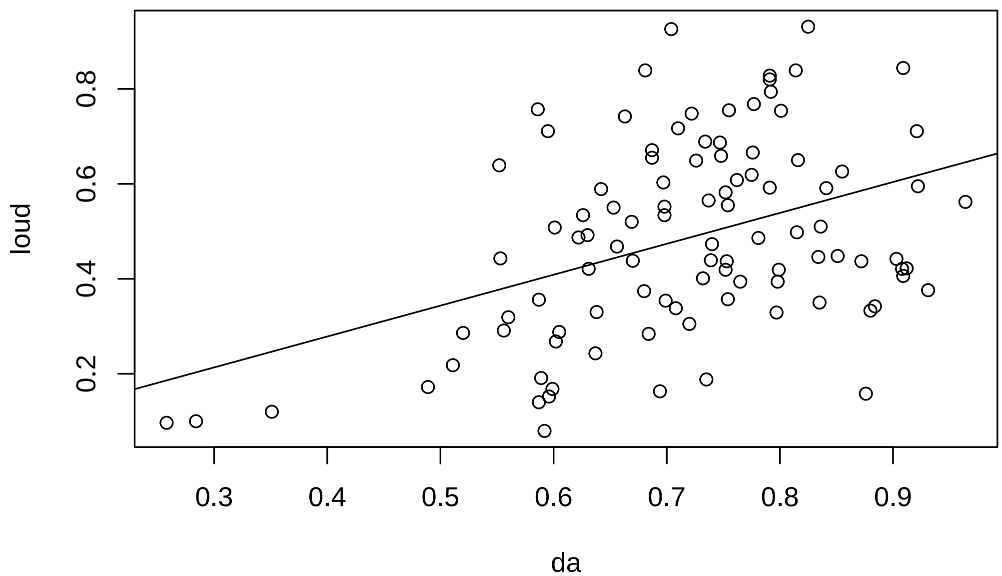


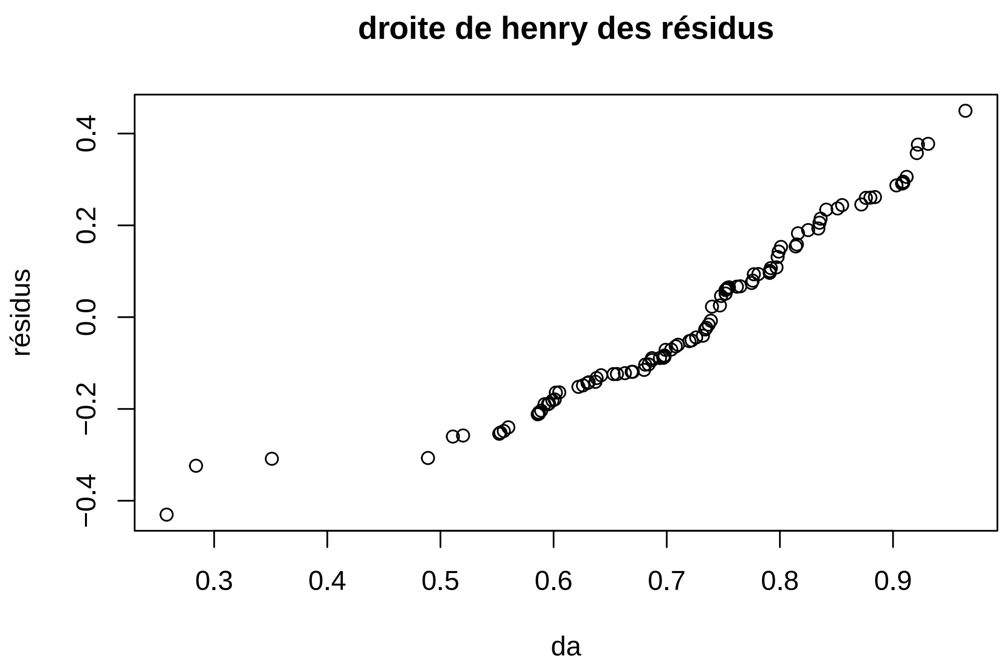


Au vu de la valeur de F-statistic et de la p-valeur, le test de fisher montre un lien linéaire entre les 2. Cependant, Par la droite de henry des résidus qui montre une distribution non-gausienne de ces derniers par rapport aux valeurs de **val**, et par le fait que les valeurs des résidus ne soient pas inclus dans l'intervalle $[-2var(e), +2var(e)]$,le test de fisher n'est pas valable et la modélisation linéaire non satisfaisante.

6.1.0.3 Variables da/val

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.43029 -0.14119 -0.03361  0.13412  0.44967
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.0181     0.1053   0.172    0.864
## x             0.6509     0.1446   4.500 1.87e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1886 on 98 degrees of freedom
## Multiple R-squared:  0.1713, Adjusted R-squared:  0.1628
## F-statistic: 20.25 on 1 and 98 DF,  p-value: 1.868e-05
```

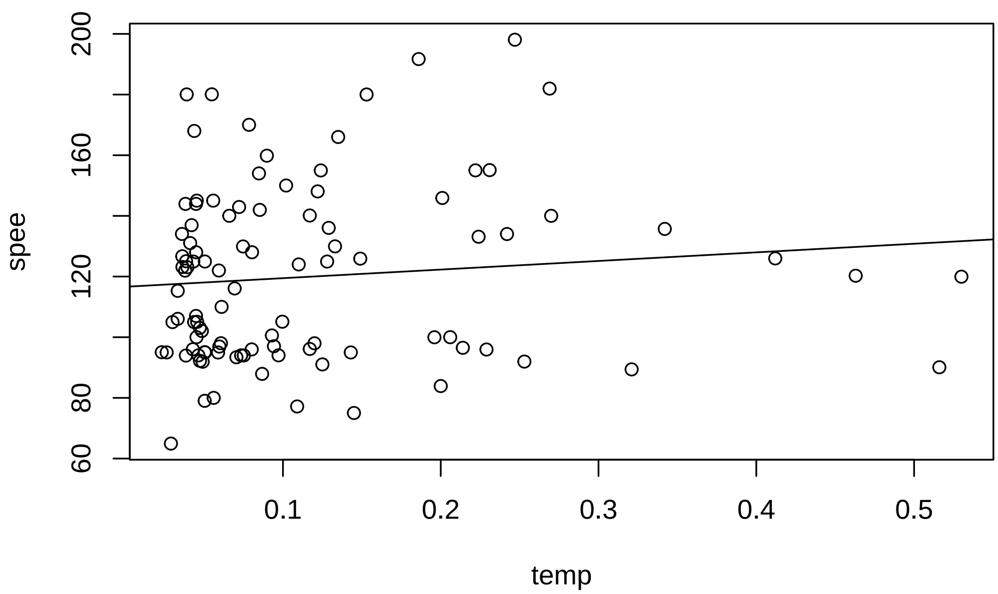




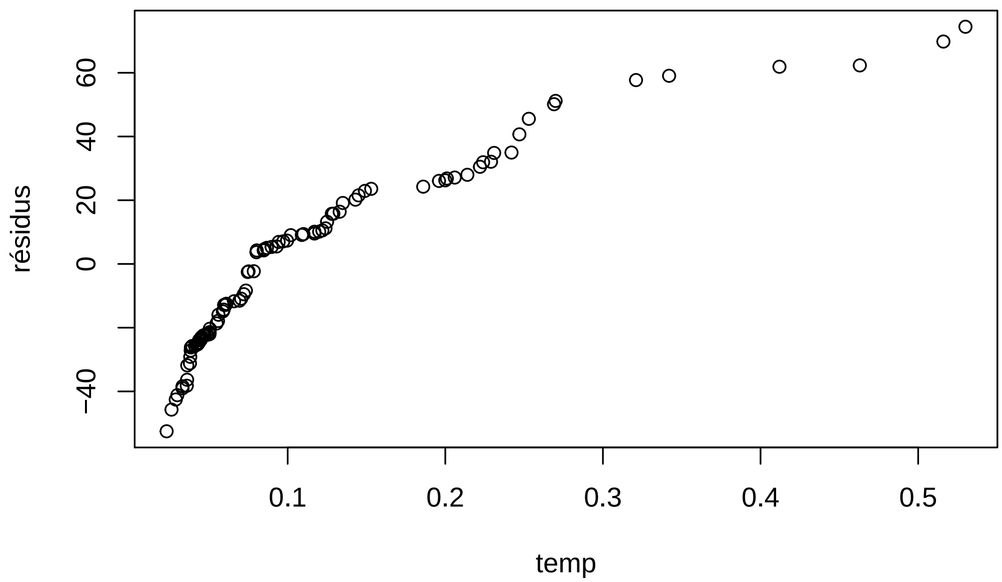
Au vu de la valeur de F-statistic et de la p-valeur, le test de fisher montre un lien linéaire entre les 2 variables. Cependant, Par la droite de henry des résidus qui montre une distribution non-gausienne de ces derniers par rapport aux valeurs de **da**, et par le fait que les valeurs des résidus ne soient pas inclus dans l'intervalle $[-2var(e), +2var(e)]$,le test de fisher n'est pas valable et la modélisation linéaire non satisfaisante.

6.1.0.4 Variable spee/temp

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -52.514 -22.997  -2.429  19.392  74.441
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  116.625      4.304  27.099   <2e-16 ***
## x             28.375     27.681   1.025    0.308
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 28.79 on 98 degrees of freedom
## Multiple R-squared:  0.01061,    Adjusted R-squared:  0.0005131
## F-statistic: 1.051 on 1 and 98 DF,  p-value: 0.3078
```



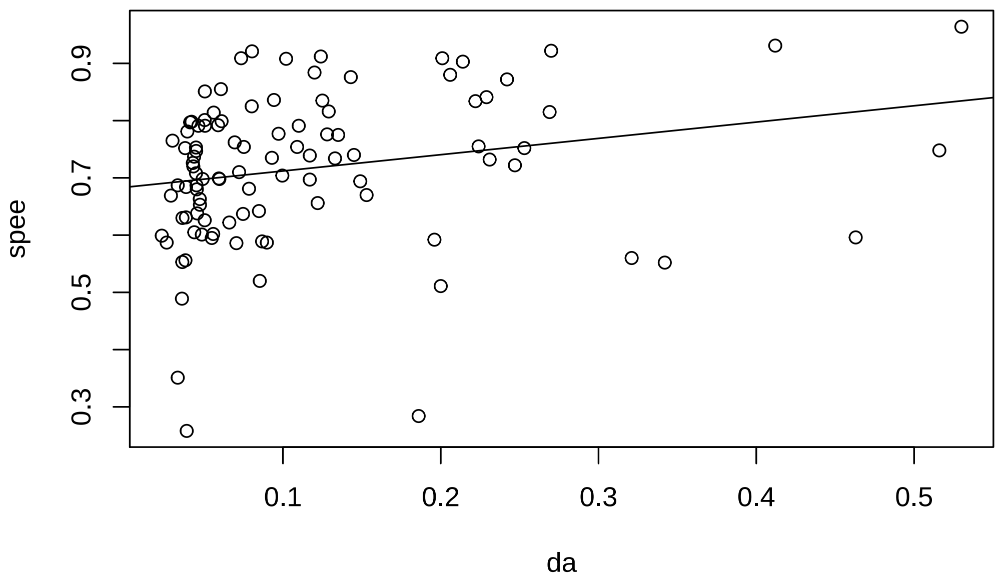
droite de henry des résidus

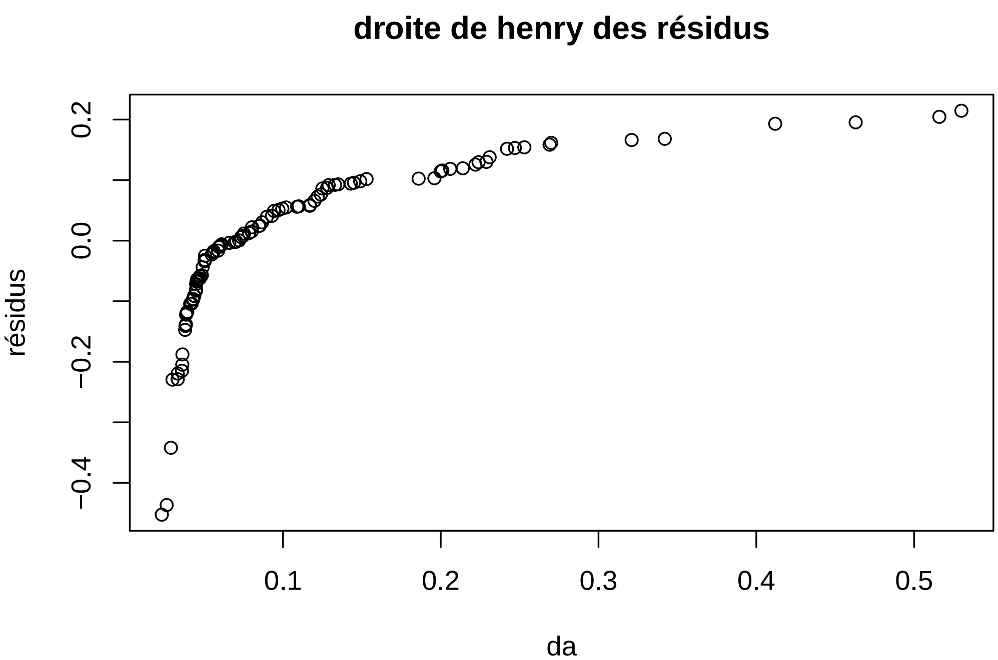


Au vu de la valeur de F-statistic et de la p-valeur, le test de fisher montre un lien linéaire entre les 2 variables. Cependant, Par la droite de henry des résidus qui montre une distribution non-gausienne de ces derniers par rapport aux valeurs de **temp**, et par le fait que les valeurs des résidus ne soient pas inclus dans l'intervalle $[-2var(e), +2var(e)]$,le test de fisher n'est pas valable et la modélisation linéaire non satisfaisante.

6.1.0.5 Variable spee/da

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.45251 -0.06623  0.00966  0.09334  0.21455
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.68355    0.01918  35.640   <2e-16 ***
## x            0.28474    0.12336   2.308   0.0231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1283 on 98 degrees of freedom
## Multiple R-squared:  0.05156,    Adjusted R-squared:  0.04189
## F-statistic: 5.328 on 1 and 98 DF,  p-value: 0.02309
```





Au vu de la valeur de F-statistic et de la p-valeur, le test de fisher montre un lien linéaire entre les 2 variables. Cependant, Par la droite de henry des résidus qui montre une distribution non-gausienne de ces derniers par rapport aux valeurs de **da**, et par le fait que les valeurs des résidus ne soient pas inclus dans l'intervalle $[-2var(e), +2var(e)]$,le test de fisher n'est pas valable et la modélisation linéaire non satisfaisante.

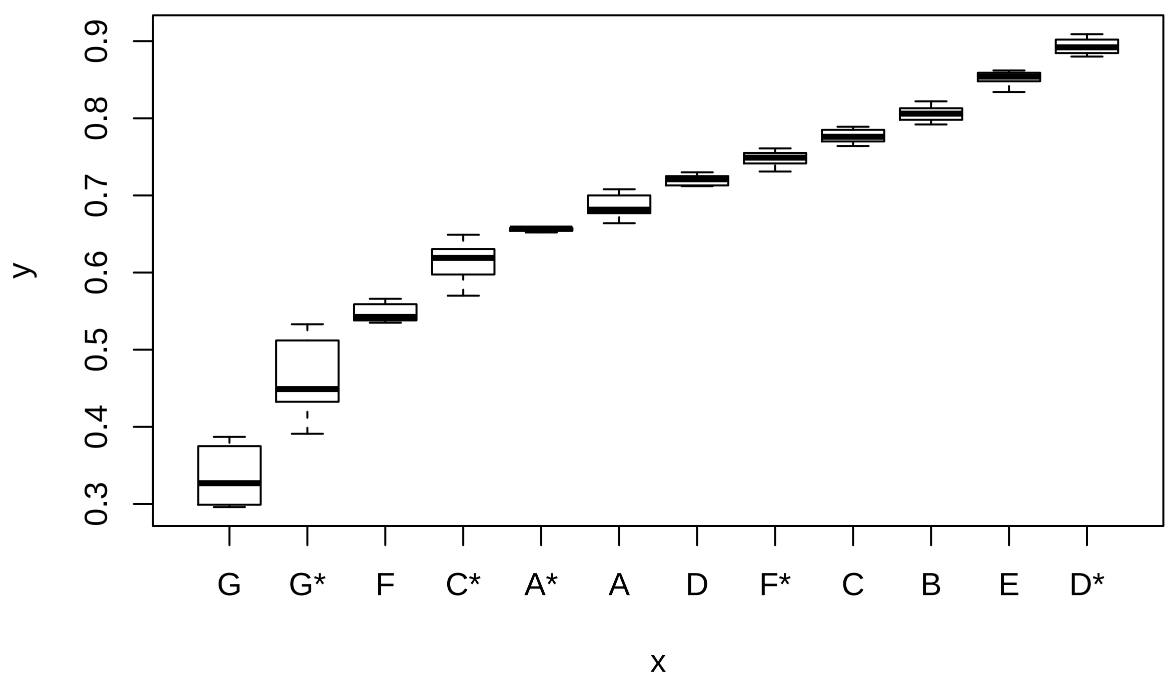
6.2 Couple de variable qualitative/quantitative

Dans cet sous section nous étudierons les couples de variables qualitative/quantitative. Nous procéderons à différents tests de fisher. De plus sachant que le nombre d'individus est de 100 nous étudierons uniquement l'hypothése d'homoscédasticité du couple. Nous validerons ce point par un plot adéquat. Pour ce faire, nous utiliserons la fonction suivante :

```r
fisher.quali = function(x,y,p){
  qqplot(x,y)
  #x est la variable quali
  #p est le nombre de classe pour y
  yquali = cut(y,p)
  table = table(x,yquali)
  fisher.test(table, simulate.p.value = TRUE, B = 1e6)
}
```

6.2.0.1 variables key/en

```r
fisher.quali(key,en,10)
```

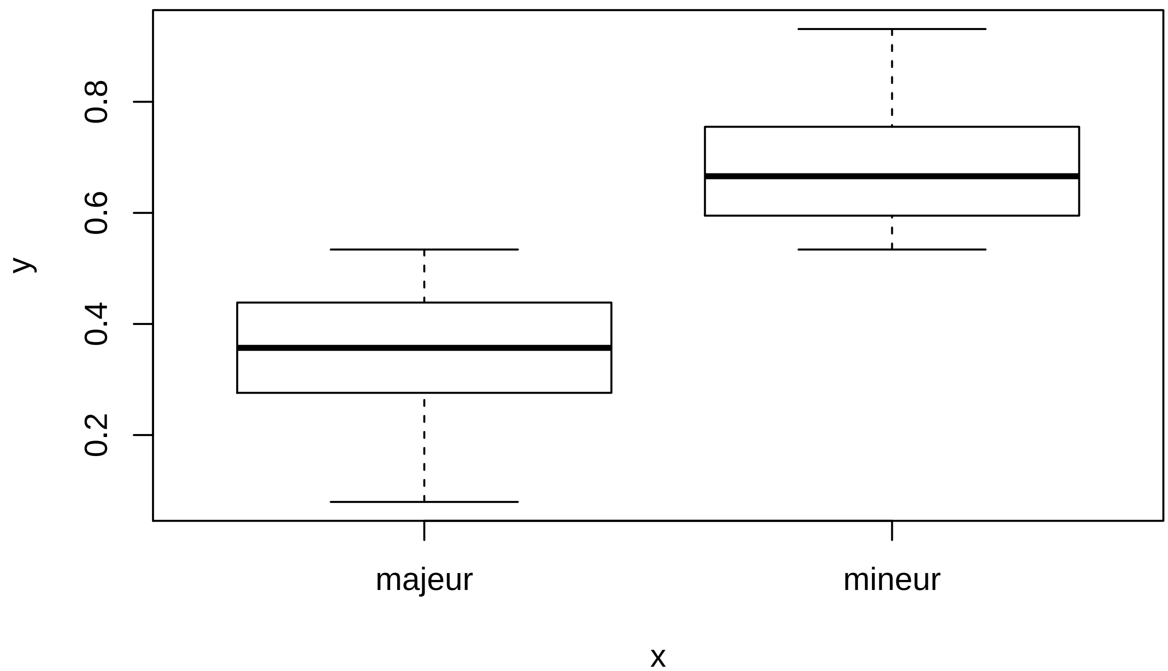



```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  table
## p-value = 0.9961
## alternative hypothesis: two.sided
```

Le graphique montre des boîtes à moustaches qui réfutent l'état d'homoscédasticité. Ainsi le test de fisher n'est pas valide.

6.2.0.2 variable mod/val

```
fisher.quali(mod,val,2)
```

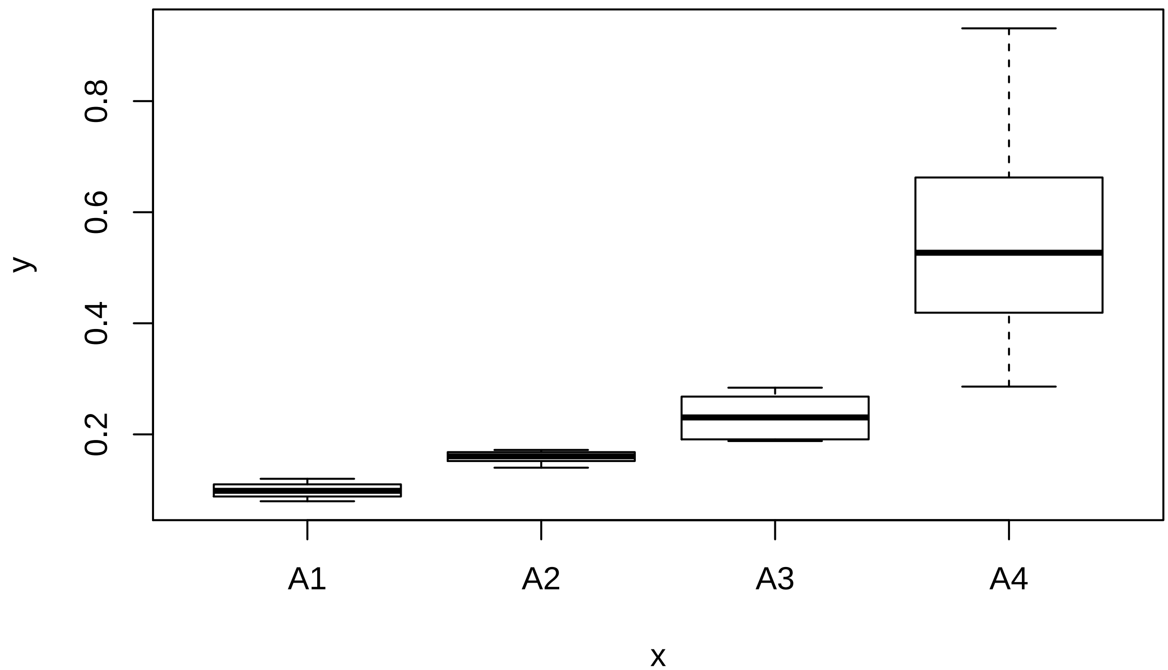



```
##
##  Fisher's Exact Test for Count Data
##
## data:  table
## p-value = 0.02654
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  1.043908 6.285309
## sample estimates:
## odds ratio
##   2.529777
```

Par le premier graphique, nous reconnaissons un état d'homoscédasticité des boites à moustaches ainsi le test est valide. Le test de Fisher nous rapporte avec une faible p-valeur que l'hypothèse d'indépendance(H0) est réfutée.

6.2.0.3 variables author/val

```
fisher.quali(author,val,10)
```

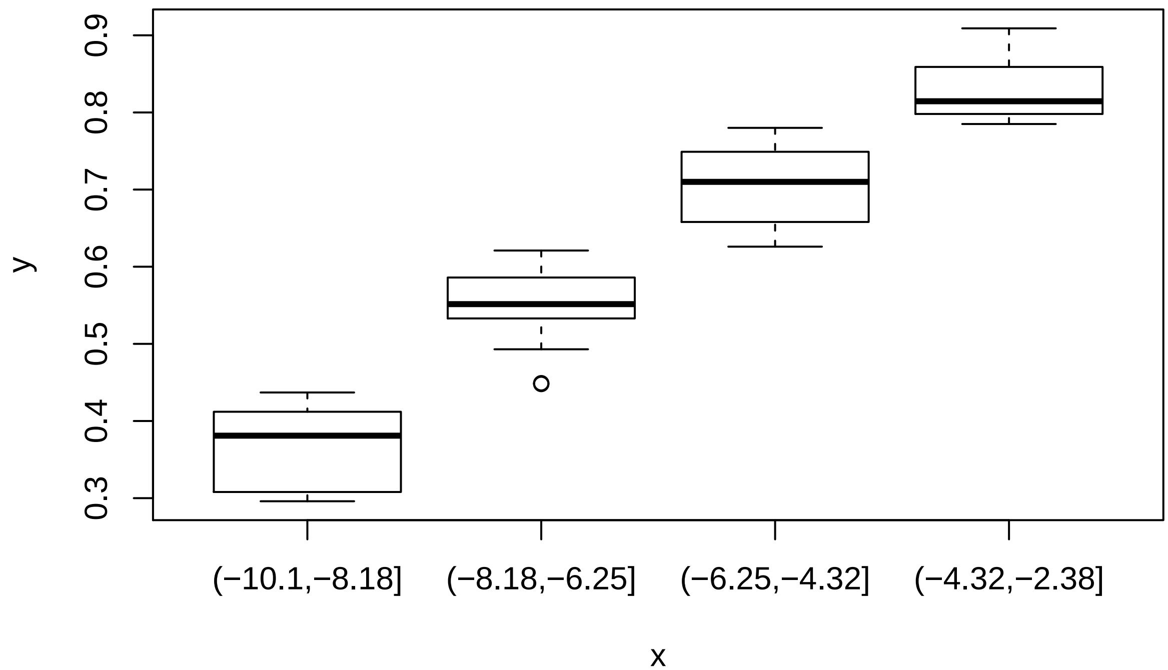



```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  table
## p-value = 0.6631
## alternative hypothesis: two.sided
```

Le graphique montre des boîtes à moustaches qui réfutent l'état d'homoscédasticité. Ainsi le test de fisher n'est pas valide.

6.2.0.4 variables loud(vu comme qualitative)/en

```
fisher.quali(cut(loud,4),en,4)
```




```
##
##  Fisher's Exact Test for Count Data with simulated p-value (based on
##  1e+06 replicates)
##
## data:  table
## p-value = 1e-06
## alternative hypothesis: two.sided
```

Par le premier graphique, nous reconnaissons un état d'homoscédasticité des boites à moustaches ainsi le test est valide. Le test de Fisher nous rapporte avec un intervalle de confiance de 98% que l'hypothèse d'indépendance (H0) est réfutée.

6.3 couple de variables qualitatives

Dans cette sous section, nous étudierons les couples de variables qualitatives. Nous procédrons à différent tests du $\tilde{\chi}^2$. Par un souci de page restante, nous ne regrouperons pas les modalités des couples de variables pour lesquels la condition de validité n'est pas respectée (à savoir que chaque case du tableau de contingence doit contenir au moins 6 individus). Pour appliquer ce test, et réaliser quelques profiles (lignes et colonnes) nous utiliserons la fonction suivante:

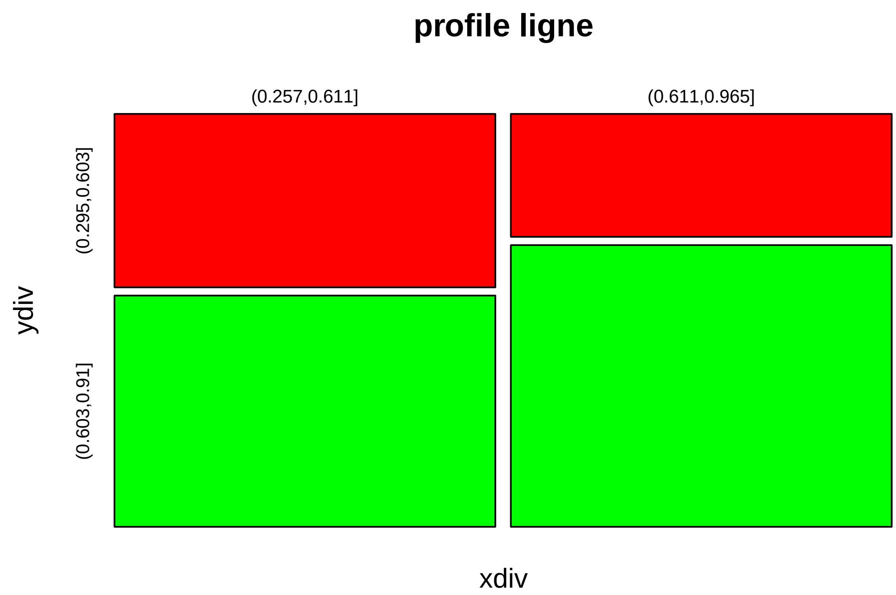
```
chi.quanti = function(x,p,y,q){
  xquali = cut(x,p)
  yquali = cut(y,q)
  table = table(xquali,yquali)
  for (i in 1:dim(table)[1]){
    for (j in 1:dim(table)[2]){
      if (table[i,j]<=5){
```

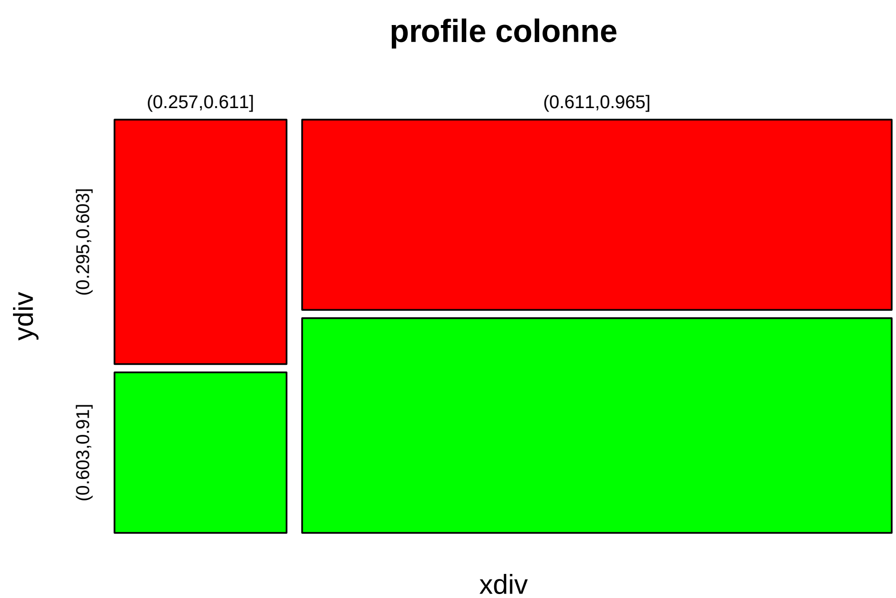
```r
        print("case n°")
        print(i)
        print(j)
        print(table[i,j])
        return(1)
      }
    }
  }
  chisq.test(table)
}

profiles = function(x,y,division,profil){
  #on suppose qu'au moins une est qualitative ici c'estx
  ydiv = (cut(y,division))
  #sinon
  if (is.null(levels(x))){
    xdiv = (cut(x,division))
  }
  table = table(xdiv,ydiv)
  if (division <=4){
    if (profil == 1){
      plot((prop.table(table,profil)),col =c('red','green','blue','grey') , main = "profile ligne")
      }
    else{
      plot((prop.table(table,profil)),col =c('red','green','blue','grey') , main = "profile colonne")
    }
  }
}
```

```r
#Test du Chi 2
chi.quanti(da,2,en,2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 0.67199, df = 1, p-value = 0.4124
```

```r
profiles(da,en,2,1)
profiles(da,en,2,2)
```

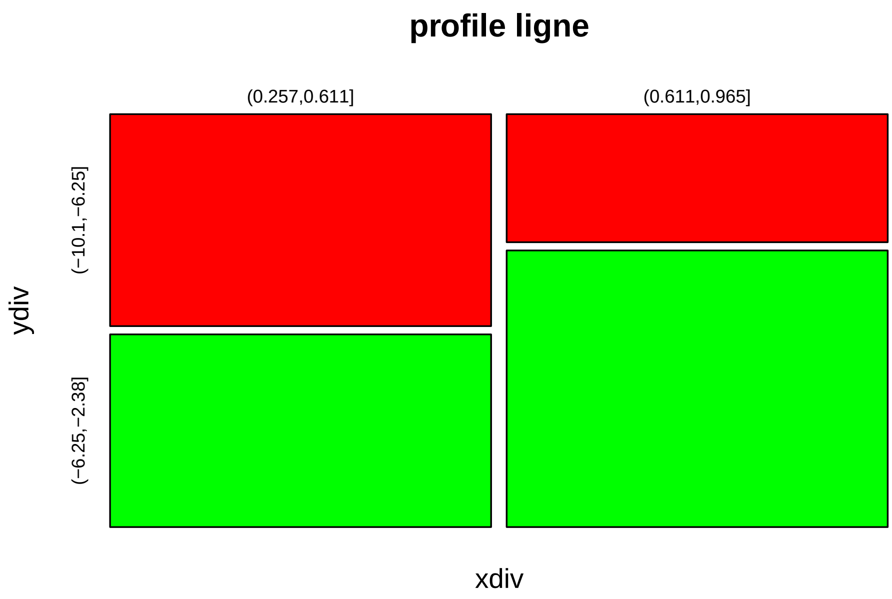


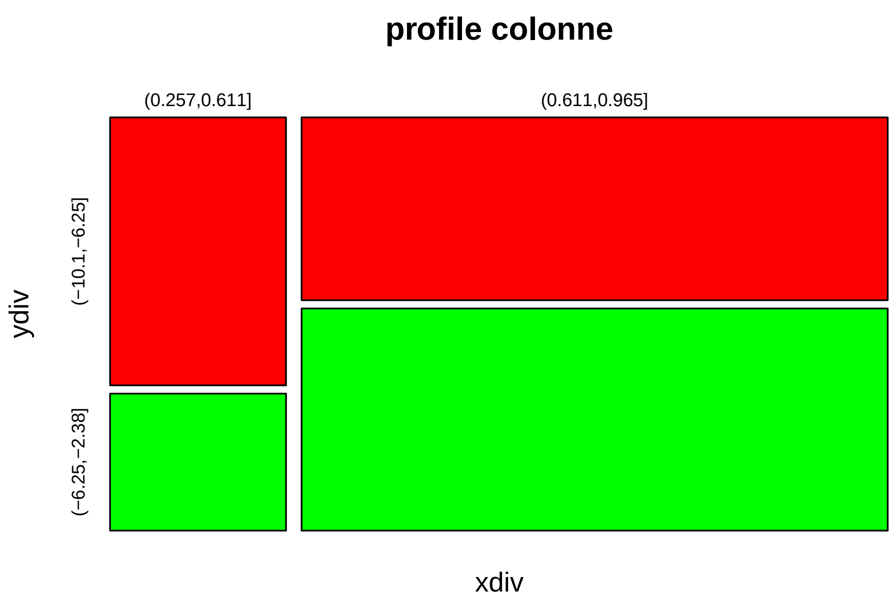



```r
chi.quanti(da,2,loud,2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table
## X-squared = 2.2613, df = 1, p-value = 0.1326
```

```r
profiles(da,loud,2,1)
profiles(da,loud,2,2)
```



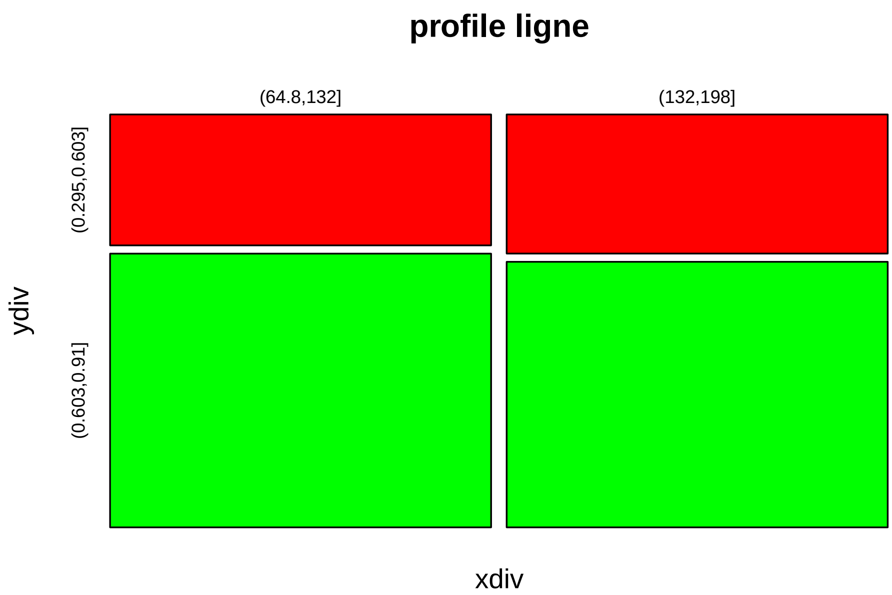


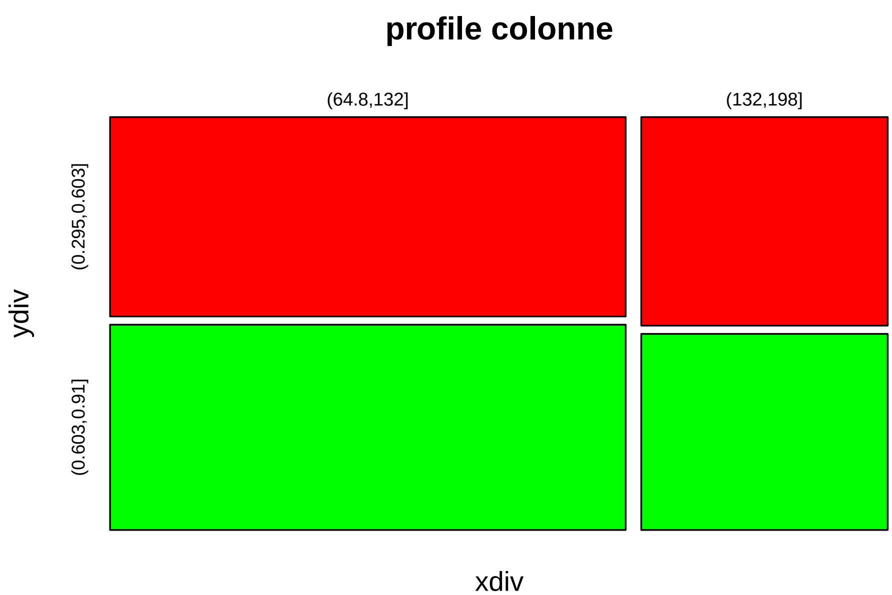

```r
chi.quanti(temp,2,en,2)
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  table
## X-squared = 1.45e-30, df = 1, p-value = 1
```

```r
profiles(temp,en,2,1)
profiles(temp,en,2,2)
```





7 Conclusion

Pour conclure, malgrès le fait que ce jeu de données contienne des variables peu nets quant aux sens et aux liens qu'elles engendrent avec le reste des variables, les différentes méthodes d'analyses nous ont permi de les mettre en lumière et de montrer les structures saillantes que compose ce jeu de données.
Nous faisons notamment allusion aux analyses multivariées qui nous ont permi de comprendre les informations globales qu'impliquent les variables ainsi que les individus via l'ACP et l'AFC, et qui nous a aussi permi de simplifier le jeu de donnée comme lorsque l'on a regroupé un nombre conséquent de modalités d'une variable qualitative via une classification hierarchique, cette même méthode qui nous a permi d'analyser les individus via les variables quantitatives du jeu de données.
Nous avons d'ailleurs aussi pû établir des analyses plus précise en prenant des couples de variables et de les soumettre à des tests(d'indépendance/linéaire ou non), ainsi que d'établir des modéles linéaires prédictifs d'une variable par rapport à une autre, ces modéles sont d'ailleurs eux même testés par la suite.
Enfin ce projet à aussi été l'occasion de se familliariser d'avantage que via les TPs avec le langage R par la documentation supplémentaire sur le net ainsi que par l'élaboration de fonctions qui pour la plus part ont fonctionnées et ont été utilisées pour ce projet.