Projet de Statistiques Descriptives

Groupe 3 MIC B : Etheve Eva - Gonzalez Julie - Roig Lila

1. Description du jeu de données

```
spotify <- read.table("spotify-3MIC.txt", header=TRUE)
nb_ind = nrow(spotify)
nb_var = ncol(spotify)
```

Le jeu de données comporte 11 variables statistiques et porte sur un échantillon de 10 000 morceaux de musique extraits de la plateforme musicale *spotify*. On peut d'ores et déjà classer les variables en différentes catégories :

Variables qualitatives ordinales :
- *year* : année de sortie du morceau de 1921 à 2020 (possède donc 100 modalités)
- *pop.class* : popularité du morceau (avec 4 modalités : de "A" pour très populaire, à "D" pour "pas populaire")

Variables qualitatives nominales :
- *explicit* : avec 2 modalités "1" si le morceau contient des vulgarités, et "0" sinon
- *key* : tonalité en début de morceau. Cette variable comporte 12 modalités
- *mode* : mode du morceau (avec 2 modalités : "0" si la tonalité est mineure, et "1" si la tonalité est majeure)

Variables quantitatives continues :
- *acousticness* : métrique relative interne de l'acoustique du morceau
- *duration* : durée du morceau en millisecondes (ms)
- *energy* : métrique relative interne de l'intensité, des rythmes du morceau (rapide, fort, bruyant)
- *liveness* : proportion du morceau où l'on entend un public (en live)
- *loudness* : mesure relative du volume du morceau (en décibels dB)
- *tempo* : le tempo du morceau, en battements par minute (bpm)

Dans ce projet nous tenterons de décrire de façon synthétique les données à notre disposition afin de mieux les analyser et mettre en lumière les liens qui peuvent exister entre nos différentes variables.

2. Etude statistique unidimensionnelle

2.1. Pour les variables qualitatives nominales

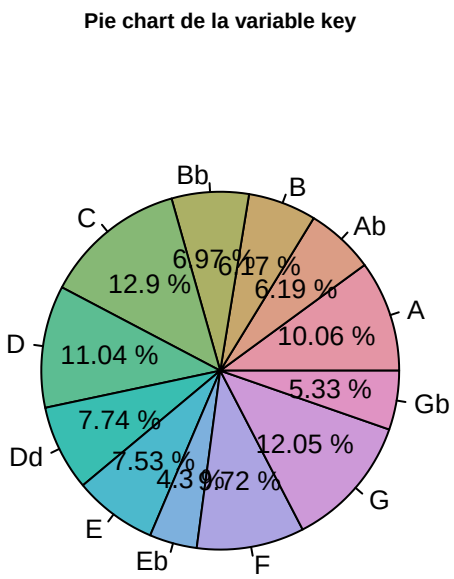
```
# Fonction permettant l'affichage des pourcentages dans les pie charts
text_pie = function(vector,labels=c(),cex=1) {
     vector = vector/sum(vector)*2*pi; temp = c()
     j = 0; l = 0
     for (i in 1:length(vector)) {
          k = vector[i]/2; j =  j+l+k; l = k
          text(cos(j)/2,sin(j)/2,paste(labels[i],"%"),cex = 1)}
     vector = temp }
```

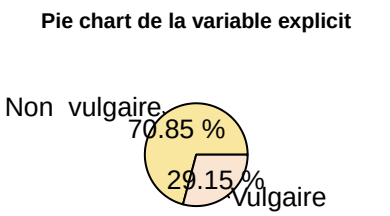
Nous pouvons représenter graphiquement les variables *explicit*, *key* et *mode* avec les pie charts suivants

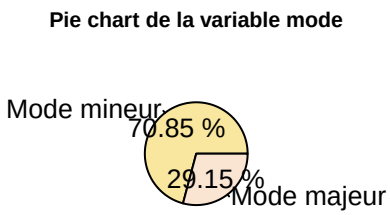
```
###Variable Explicit
Explicit = factor(x=spotify$explicit, labels=c("NV", "V"))
#nombre de morceaux présentant une vulgarité
nb_Vexplicit = sum(spotify$explicit)
#vecteur pourcentages
```

```
pourcent_explicit = c(100*(nb_ind-nb_Vexplicit)/nb_ind, 100*nb_Vexplicit/nb_ind)
###Variable Key
Key = factor(x=spotify$key)
B = table(Key)
table_key = data.frame(Effectif= c(B), Fréquence=c(B)/sum(B), Angle= c(B)/sum(B)*360)
pourcent_key = table_key[,2]*100 #vecteur pourcentages
###Variable Mode
Mode = factor(x=spotify$mode, labels=c("mineur","majeur"))
nb_Majmode = sum(spotify$mode)#nombre de morceaux présentant un mode majeur
#vecteur pourcentages
pourcent_mode = c(100*(nb_ind-nb_Majmode)/nb_ind, 100*nb_Majmode/nb_ind)
```

```
layout(matrix(c(1,1,2,3), nrow=2),  widths=c(3,2.5), heights=c(2,2))
### pie chart de la variable key
pie(table(Key), labels=c("A", "Ab", "B", "Bb", "C", "D", "Dd", "E", "Eb", "F" , "G", "Gb"),
main ="Pie chart de la variable key", cex.main = .8, col= rainbow_hcl(12))
text_pie(pourcent_key, strsplit(toString(pourcent_key), ", ")[[1]], cex=0.9)
### pie chart de la variable Explicit
pie(table(Explicit), col=c("#F9E79F", "#FAE5D3"), labels=c("Non  vulgaire", "Vulgaire"),
main="Pie chart de la variable explicit", cex.main =.8)
text_pie(pourcent_explicit, c(pourcent_explicit[1],pourcent_explicit[2]), cex=1.1)#affichage pourcentages
### pie chart de la variable mode
pie(table(Mode), labels=c("Mode mineur", "Mode majeur"), main ="Pie chart de la variable mode",
    cex.main = .8, col=c("#F9E79F","#FAE5D3"))
text_pie(pourcent_mode, c(pourcent_mode[1],pourcent_mode[2]), cex=.8) #affichage pourcentages
```







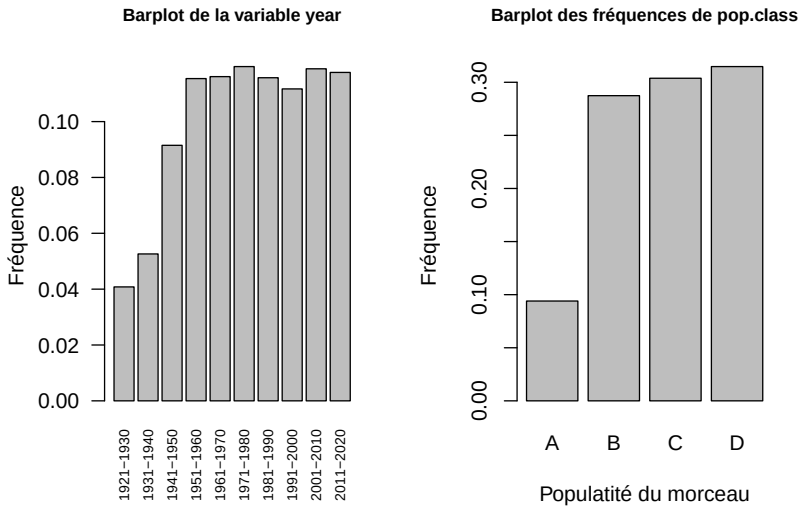
Interprétation des résultats :
On remarque que les variables *mode* et *explicit* ont exactement la même répartition. En effet, 70.85% des morceaux ne comportent pas de vulgarités et sont de tonalité mineure, tandis que 29.15% des morceaux contiennent des vulgarités et sont de tonalité majeure. On peut donc conjecturer pour la suite, lors de l'analyse bidimentionnelle, que ces variables

sont liées. Pour la variable *key*, nous pouvons relever que les clés A, C, D et G sont les plus représentées dans l'échantillon mais que globalement il existe une répartition équitable entre les différentes tonalités.

2.2. Pour les variables qualitatives ordinales

```
###prend en argument un vecteur V1 et regroupe les valeurs par plages de "pas" ans.
f_regroup_vecteur <- function (V1,pas) {
  nb_Annees = nrow(V1)
  year_plage = seq (1,nb_Annees/pas)
  j = 1;
  for (i in seq(1,(nb_Annees-pas+1),by=pas)) {
    year_plage[j] = sum(V1[i:(i+pas-1)])
    j = j + 1}
  return(year_plage)}
###prend en argument une matrice M1 et regroupe les valeurs par plages de "pas" ans.
f_regroup_matrice <- function (M1,pas) {
  nb_colonnes = ncol(M1)
  nb_Annees = nrow(M1) #nb_Annees=100 ans
  year_plage = matrix(0,ncol=nb_colonnes, nrow=nb_Annees/pas)
  for (n in 1:nb_colonnes){
    j = 1;
    for (i in seq(1,(nb_Annees-pas+1),by=pas)) {
      year_plage[j,n] = sum(M1[i:(i+pas-1),n])
      j = j + 1} }
  return(year_plage) }
#affiche le nom des colonnes d'une matrice ou vecteur regroupées par plages de "pas" ans
f_affiche <- function(M1,a1,a2,pas) {
  tabYears1 = seq(a1, 2020, by = pas)
  tabYears2 = seq(a2, 2020, by = pas)
  abs_names = seq(1, (nrow(M1)/pas))
  for (n in 1:(nrow(M1)/pas)){
    abs_names[n] = paste(tabYears1[n], tabYears2[n], sep = "-")}
  return(abs_names)}
#regroupe les annees en pas ans pour un meilleur affichage du biplot (cf. ACP)
f_year_biplot <- function (V,pas) {
  l = length(V) ; V2 = seq(1,l)
  for (i in seq(1,l)){nb = V[i]
    if (nb%%pas == 0){a = nb-pas+1 ; b = nb}
    else {a = (nb%/%pas)*pas + 1 ; b = (nb%/%pas)*pas + pas}
    V2[i] = paste(toString(a),toString(b),sep = "-")}
  return (V2)}
```

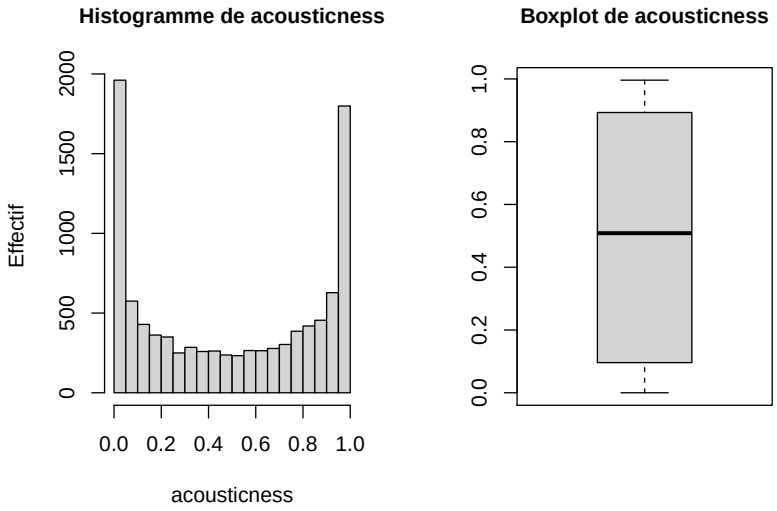
```
Year = spotify$year; par(mfrow = c(1,2))
###Variable Year
V = f_regroup_vecteur (table(Year)/nb_ind,10)
abs_names = f_affiche (table(Year),1921,1930,10)
barplot(V, main="Barplot de la variable year", xlab = "", ylab = "Fréquence",
        names.arg = abs_names, las=2, cex.names=0.7, cex.main =.8)
###Variable pop.Class
PopClass=spotify$pop.class; fq_pop <- table(PopClass)/nb_ind
barplot(fq_pop, main="Barplot des fréquences de pop.class", xlab =
          "Populatité du morceau", ylab = "Fréquence", cex.main =.8)
```



Interprétation des résultats :
Pour la variable *year*, on voit que les utilisateurs écoutent peu de morceaux datant d'avant 1950. En revanche, ils écoutent en proportions quasiment égales des morceaux datant des années 1960 jusqu'a nos jours. On remarque que les chansons récoltant le plus grand succès sont celles datant des années 70 et les chansons actuelles. Pour la variable *pop.class* représentant la popularité du morceau, on remarque que le jeu de données contient peu de morceaux populaires (10% environ) i.e classés A. Concernant les autres notations, elles sont bien réparties et chacune d'entre elles représente près d'un tiers du jeu de données.

2.3. Pour les variables quantitatives continues

```
###Variable acousticness
acousticness = spotify$acousticness
par(mfrow = c(1,2))
A = hist(acousticness,freq = TRUE, ylab = "Effectif", main = "Histogramme de acousticness", cex.main = 1)
boxplot(acousticness, main = "Boxplot de acousticness",cex.main = 1)
```



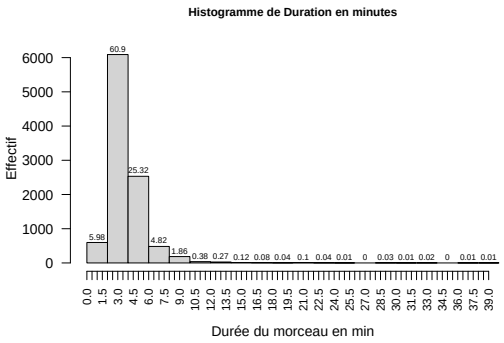

```
summary(acousticness)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0961  0.5085  0.4990  0.8930  0.9960
```

Interprétation des résultats :

Avec l'histogramme de la variable **acousticness**, on remarque que près de 2000 morceaux ont une valeur de acousticness minimale (0.0) et près de 2000 morceaux ont une valeur de acousticness maximale (0.9960). Avec le boxplot et l'histogramme, on voit que les morceaux ne prenant pas de valeurs d'acousticness extrêmes sont identiquement répartis entre le premier quantile (0.0961), la médiane (0.5085) et le troisième quantile (0.8930). Ainsi, 50% des morceaux ont un taux d'acousticness entre 0.0961 et 0.8930 et 50% des morceaux prennent des valeurs extrêmes.

```
Duration = spotify$duration; par(mfrow = c(1,1))
Duration_min = round(Duration/60000, 2) #on affiche en minutes
D = hist(Duration_min,freq = TRUE, ylab = "Effectif", xlab = "Durée du morceau en min",
main = "Histogramme de Duration en minutes", breaks = 20, las = 2,xaxt="n", cex.main = .8)
axis (side = 1, at = seq (floor(min(Duration_min)),ceiling(max(Duration_min)),0.5), las = 2, cex.axis = .8)
text(D$mids,D$counts,labels=(D$counts/100), adj=c(0.5, -0.5), cex = .6)
```

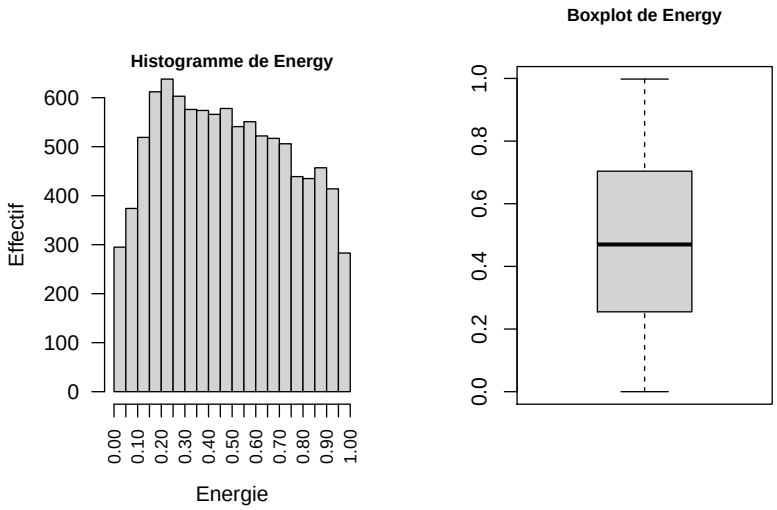



```
#boxplot pas pertinent car trop de outliers
sd(Duration)
```

```
## [1] 119049.1
```

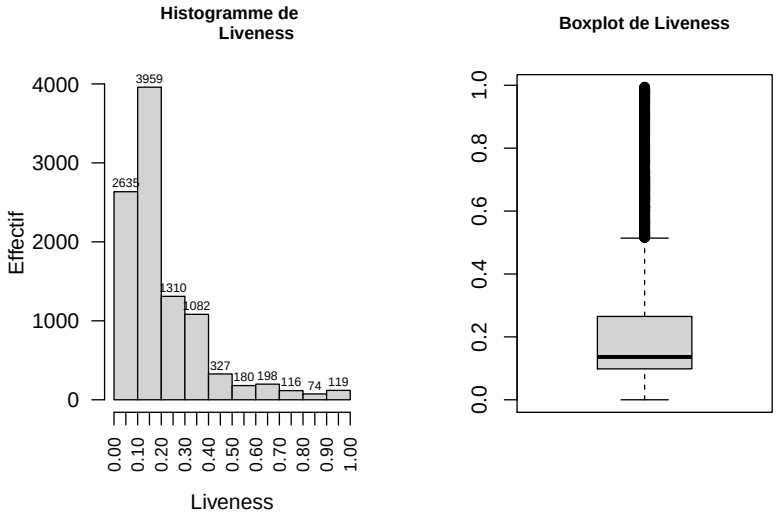
Interprétation des résultats :
L'histogramme de la variable **duration**, montre que 60% des morceaux ont une durée comprise entre 1.5 minutes et 4 minutes. C'est en effet la durée classique d'un morceau de musique. 25% des morceaux ont une durée comprise entre 4 minutes et 6 minutes. Seulement 5% des morceaux ont une durée inférieure à 1.5 min et environ 8% des morceaux durent plus de 6 minutes. Ce graphe souligne les grandes disparités qu'il peut y avoir dans ce jeu de données car nous voyons que la variable peut prendre des valeurs très différentes. Cela est vérifié par la valeur de l'écart-type qui est extrêmement élevé : 119049.

```
Energy = spotify$energy; par(mfrow = c(1,2))
E = hist(Energy,freq = TRUE, ylab = "Effectif", xlab = "Energie", breaks = 20, las = 2, xaxt="n",
         main = "") ; title(main="Histogramme de Energy",cex.main = .8, line=0)
axis (side = 1, at = seq (floor(min(Energy)),ceiling(max(Energy)),0.05), las = 2,
      cex.axis = .8)
boxplot(Energy, main = "Boxplot de Energy",cex.main = .8)
```



Interprétation des résultats :
On voit avec la variable **energy** que l'énergie des morceaux est équitablement répartie entre 0 et 1.

```
Liveness = spotify$liveness; par(mfrow = c(1,2))
Li = hist(Liveness,freq = TRUE, ylab = "Effectif", xlab = "Liveness", main = "Histogramme de
          Liveness", breaks = 10, las = 2, xaxt="n", cex.main = .8)
text(Li$mids,Li$counts,labels=Li$counts, adj=c(0.5, -0.5), cex = .6)
axis (side = 1, at = seq (floor(min(Liveness)),ceiling(max(Liveness)),0.05), las = 2, cex.axis = .8)
boxplot(Liveness, main = "Boxplot de Liveness",cex.main = .8); summary(Liveness)
```

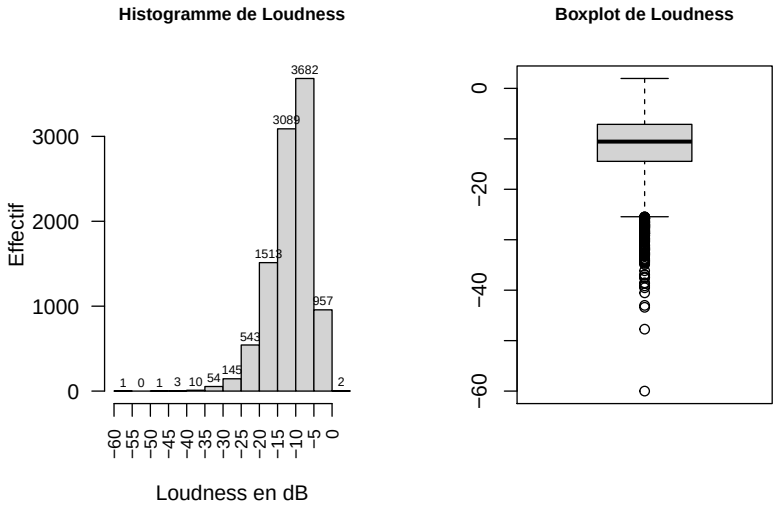



```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.0983  0.1360  0.2065  0.2650  0.9940
```

Interprétation des résultats :
Avec la variable **liveness** on voit qu'il y a 50% des chansons pour lesquelles on entend le public sur moins de 14% du morceau (la mediane vaut environ 14%). On remarque qu'il y a très peu de chansons (2.4%) pour lesquelles les extraits de live représentent 75% du morceau.

```
Loudness = spotify$loudness; par(mfrow = c(1,2))
Lo = hist(Loudness,freq = TRUE, ylab = "Effectif", xlab = "Loudness en dB", main =
"Histogramme de Loudness", breaks = 20, las = 2, xaxt="n", cex.main = .8)
axis (side = 1, at = seq (floor(min(Loudness)),ceiling(max(Loudness)),5), las = 2, cex.axis = .8)
```

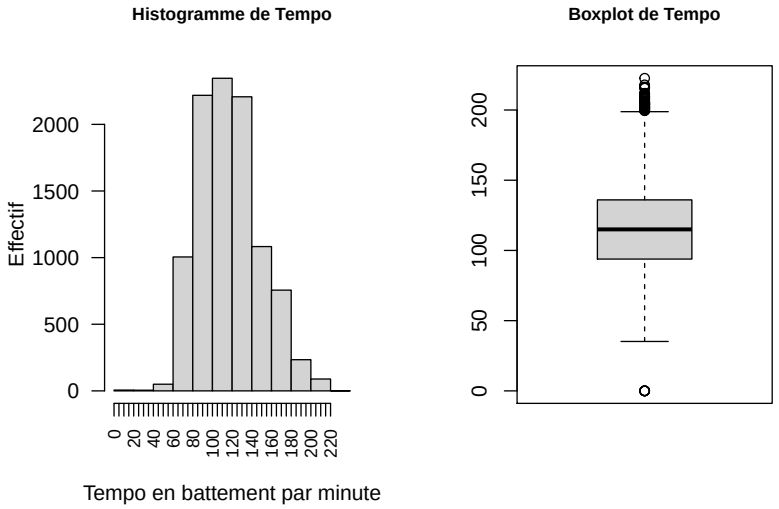
```
text(Lo$mids,Lo$counts,labels=Lo$counts, adj=c(0.5, -0.5), cex = .6)
boxplot(Loudness, main = "Boxplot de Loudness", cex.main = .8) ; summary(Loudness)
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -60.000 -14.466 -10.550 -11.371  -7.130   1.963
```

Interprétation des résultats :
Avec la variable **loudness** on voit que 50% des morceaux sont au dessus de -10.5 dB. On remarque une grande quantité de morceaux autour de -10 dB.

```
Tempo = spotify$tempo; par(mfrow = c(1,2))
T = hist(Tempo,freq = TRUE, ylab = "Effectif", xlab = "Tempo en battement par minute", main =
"Histogramme de Tempo", breaks = 10, las = 2, xaxt="n", cex.main = 0.8)
axis (side = 1, at = seq (floor(min(Tempo)),ceiling(max(Tempo)),5), las = 2, cex.axis = .8)
boxplot(abs(Tempo), main = "Boxplot de Tempo",cex.main = .8) ; summary(Tempo)
```




```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   93.85  115.00  117.28  136.00  222.60
```

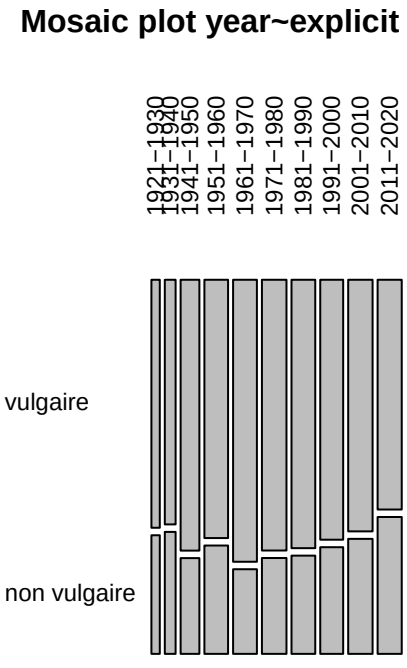
Interprétation des résultats :
Avec la variable **tempo**, on voit que le tempo moyen est autour de 117 bpm. Ce qui traduit un rythme dansant. Globalement, tous les morceaux ont un tempo entre 60 bpm (coeur au repos) et 180 bpm (sprint).

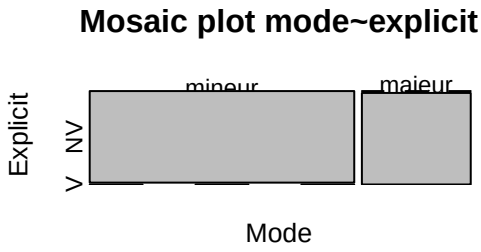
3. Etude statistique bidimensionnelle

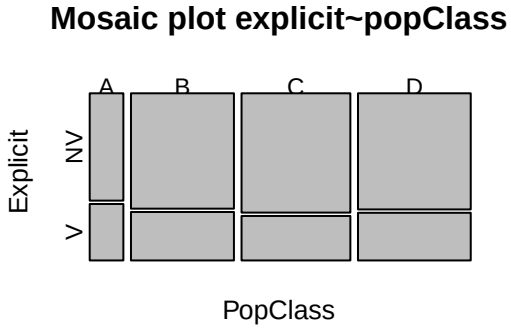
Il s'agira ici de présenter les études statistiques bidimentionnelles pour plusieurs couples de variables qu'il semble pertinent d'étudier ensemble.

3.1. Entre deux variables qualitatives

```
layout(matrix(c(1,1,2,3), nrow=2),  widths=c(1,1), heights=c(2.5,3))
### year et explicit
table.cont = table(Year, Explicit); prop.cont = prop.table(table.cont)
NewTable2 = f_regroup_matrice (table.cont,10)
rownames(NewTable2) <- f_affiche(table.cont,1921,1930,10)
colnames(NewTable2) <- c("vulgaire","non vulgaire")
NewProp2 =  f_regroup_matrice (prop.cont,10)
rownames(NewProp2) <- f_affiche(prop.cont,1921,1930,10)
colnames(NewProp2) <- c("vulgaire","non vulgaire")
mosaicplot(NewTable2, main="Mosaic plot year~explicit",las = 2, cex.axis = 0.9, cex.main = .8)
### mode et explicit
table.cont = table(Mode, Explicit); prop.cont = prop.table(table.cont)
mosaicplot(table.cont, main ="Mosaic plot mode~explicit", cex.main = .8, cex.axis = 0.9)
### popularite et explicit
table.cont = table(PopClass, Explicit); prop.cont = prop.table(table.cont)
mosaicplot(table.cont, main="Mosaic plot explicit~popClass", cex.main = .8, cex.axis = 0.9)
```

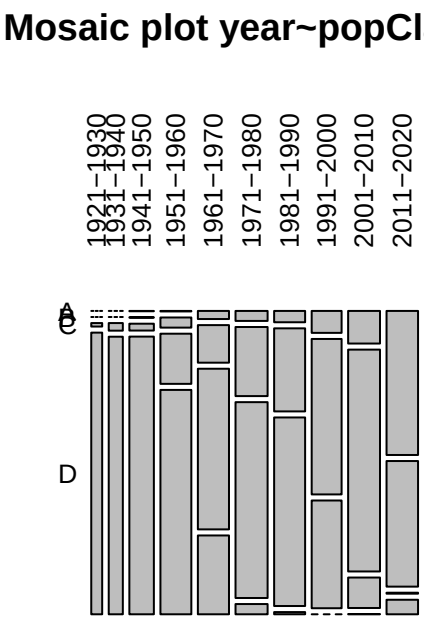


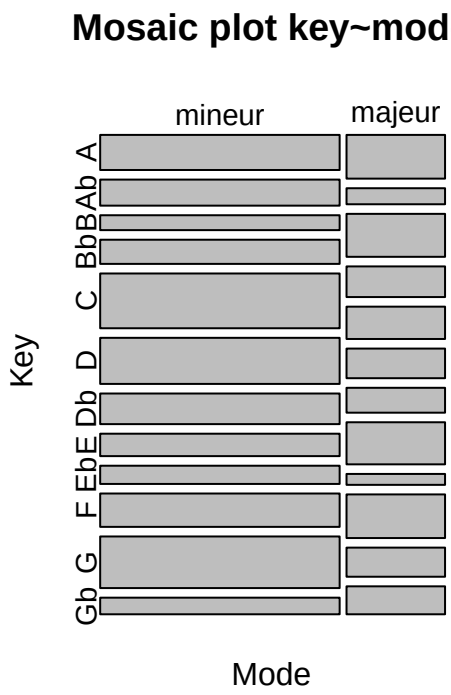





```
### year et popularité
par(mfrow = c(1,2))
Year = spotify$year
table.cont = table(Year, PopClass); prop.cont = prop.table(table.cont)
NewTable = f_regroup_matrice (table.cont,10)
rownames(NewTable) <- f_affiche(table.cont,1921,1930,10)
```

```
colnames(NewTable) <- c("A","B","C","D")
NewProp =  f_regroup_matrice (prop.cont,10)
rownames(NewProp) <- f_affiche(prop.cont,1921,1930,10)
colnames(NewProp) <- c("A","B","C","D")
mosaicplot(NewTable, main="Mosaic plot year~popClass",las = 2, cex.axis = .8, cex.main = .8)
### key et mode
table.cont = table(Mode, Key); prop.cont = prop.table(table.cont)
mosaicplot(table.cont, main="Mosaic plot key~mode", cex.main = .8,cex.axis = 0.9)
```



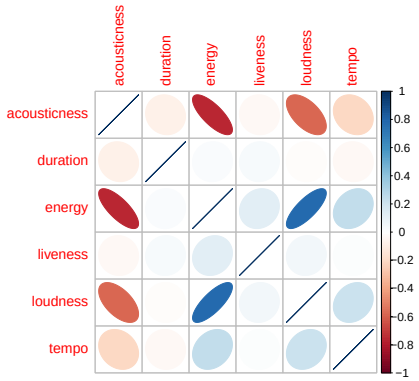


Interprétation des résultats :
- **mode** et **explicit** sont totalement liées car les séparations sont opposées. Ainsi, un morceau commençant par une tonalité mineure aura un contenu non vulgaire contrairement à une musique commençant par une tonalité majeure. Ceci confirme donc l'hypothèse émise lors de l'analyse unidimensionelle.
- **key** et **mode** sont liés. Ceci est étrange car débuter un morceau par exemple en Sol, n'indique par que la tonalité sera en Sol majeur ou Sol mineur.
- **popClass** et **year** sont liés. Ceci est confirmé par l'analyse unidimensionnelle : les morceaux actuels et des années 70 sont les plus populaires.

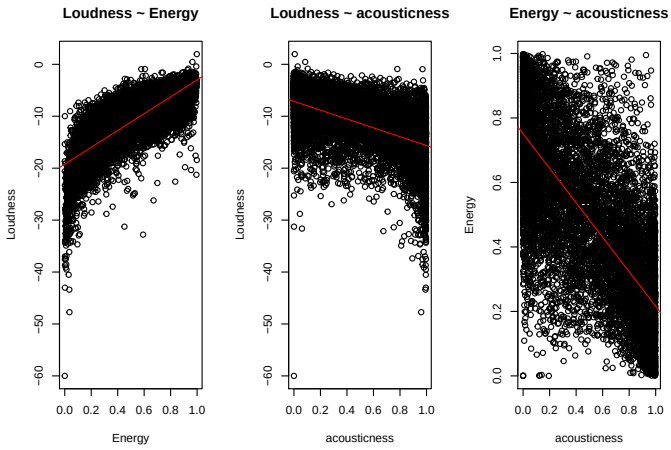
3.2. Entre deux variables quantitatives

Afin de visualiser plus facilement les corrélations entre les variables quantitatives, nous avons affiché la matrice de corrélation à l'aide de la fonction *corrplot*. Puis, si le coefficient de corrélation était suffisamment important (> 0.7 en valeur absolue), nous décidions de continuer l'étude afin de voir comment les variables évoluaient entre elles à l'aide d'une régression linéaire.

```
correlation=cor(spotify[,c(2,3,4,7,8,10)])
corrplot(correlation, method="ellipse")
```




```
par(mfrow = c(1,3))
###energy et loudness
mod = lm(Loudness ~ Energy, data=spotify)
plot(Loudness~Energy, main="Loudness ~ Energy"); abline(mod, col="red")
###acousticness et loudness
 mod = lm(Loudness ~ acousticness, data=spotify)
 plot(Loudness ~ acousticness, main="Loudness ~ acousticness"); abline(mod, col="red")
###acousticness et energy
mod = lm(Energy ~ acousticness, data=spotify)
plot(Energy ~ acousticness, main="Energy ~ acousticness") ;abline(mod, col="red")
```

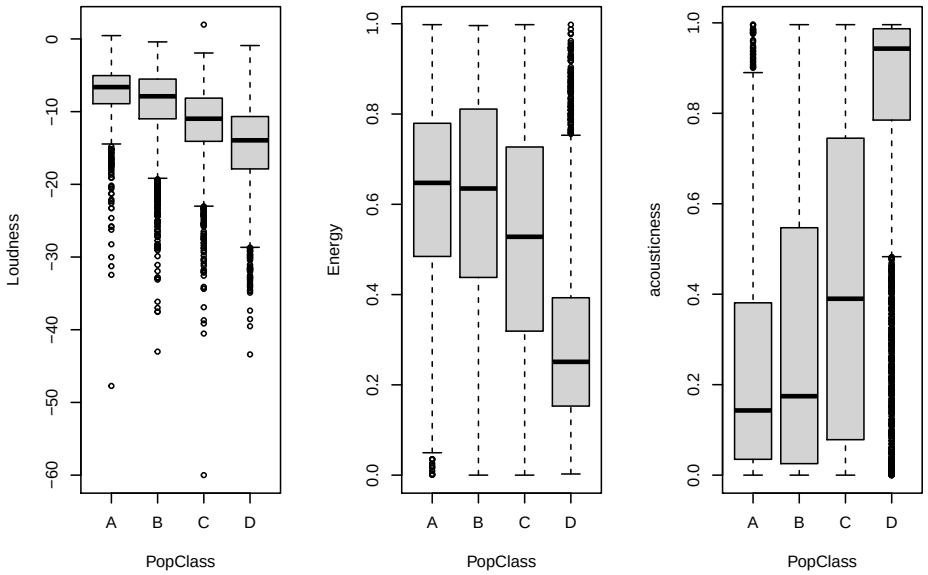


Interprétation des résultats :
Nous pouvons constater que seuls trois couples se démarquent :
- (**energy**, **acousticness**) et (**loudness**, **acousticness**) dont le coefficient de corrélation est négatif. C'est-à-dire que si une variable augmente, l'autre aura tendance à diminuer. C'est d'ailleurs ce que nous constatons sur les tracés des régressions linéaires. Par exemple, lorsque la variable **acousticness** augmente alors la variable **energy** diminue fortement.
- (**energy**, *__loudness__) dont le coefficient de corrélation est positif. Au contraire, ici les variables auront tendance à évoluer dans le même sens comme indiqué sur le tracé de la régression linéaire. Ces résultats sont logiques car si l'on veut rendre une musique plus vive, augmenter le volume sonore semble être une bonne solution et inversement si l'on veut rendre une musique plus douce.

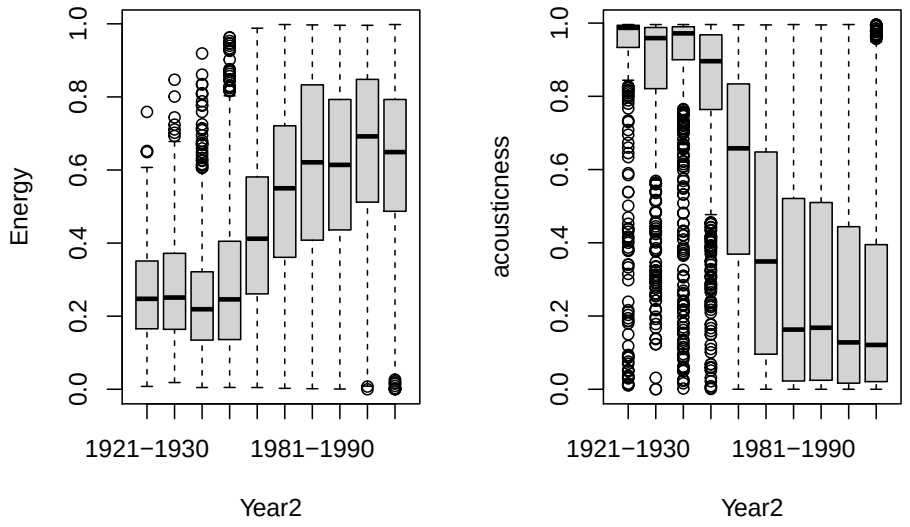
3.3. Entre une variable quantitative et une qualitative

Nous avons tracé ici les boxplots les plus pertinents afin de souligner l'influence d'une variable qualitative sur une variable quantitative.

```
par(mfrow=c(1,3)) ; boxplot(Loudness~PopClass); boxplot(Energy~PopClass); boxplot(acousticness~PopClass)
```




```
par(mfrow=c(1,2)) ; V2 = f_year_biplot (spotify$year,10) ; Year2 = factor(x=V2) ;
boxplot(Energy~Year2) ; boxplot(acousticness~Year2)
```



Interprétation des résultats :
Tout d'abord, nous observons l'influence de la variable **popClass** sur les variables **acousticness**, **energy** et **loudness**. En effet, la différence de position des boxplots révèle une dépendance entre les différents couples. Par exemple, le premier graphe montre que les morceaux ayant un volume sonore faible sont peu appréciés. En particulier parmi les morceaux de la classe D, 75% d'entre eux ne sont pas appréciés et ont un volume sonore inférieur à -10dB. Alors que pour ceux appréciés du public (classe A), 75% ont un volume supérieur à -10dB. Et comme **loudness** et **energy** sont liées, nous pouvons nous attendre au même résultat concernant **popClass** et **loudness**, à savoir : quand l'énergie est importante (et donc le volume aussi) les morceaux ont tendance à être plus appréciés. Enfin, sur le dernier graphe nous remarquons que les musiques très peu appréciées sont plus acoustiques : 75% des morceaux de la classe D ont une acoustique supérieure à 0.8. Au contraire, les musiques très appréciées présentent une faible valeur d'acoustique : 75% des musiques classées A ont une acoustique inférieure à 0.4. En somme, les musiques les plus appréciées des utilisateurs de Spotify sont énergiques et peu acoustiques.

Interprétation des résultats :
Enfin, d'autres dépendances intéressantes sont celles entre **year** et les variables **acousticness** et **energy**. Pour plus de lisibilité nous avons d'ailleurs regroupé les années en décennies. Ainsi, le premier graphique révèle que parmi les musiques plus modernes (datant des années 80 à nos jours) près de 75% des musiques sont plus énergiques. Ce qui est cohérent car cela correspond à l'explosion de nouveaux genres musicaux plus dansants tels que la *pop music* dont Michael Jackson, Madonna et encore Lady Gaga ont été les précurseurs mais aussi le *RnB* contemporain avec les Black Eyed Peas, Rihanna, Beyoncé, etc. Ces nouveaux genres se développent notamment grâce à l'essor technologique avec de nouveaux instruments : synthétiseur, clavier, guitare électrique, table de mixage... Enfin, le second graphe révèle une évolution remarquable : l'acoustique des musiques n'a pas cessé de diminuer au cours du dernier siècle. En effet, à partir des années 70-80, sur 10 ans les trois-quarts des musiques ont une acoustique relativement faible (<0.5). Ce qui contraste fortement avec les années 1920 où près de 100% des musiques (en négligeant les outliers) avaient un indicateur acoustique quasi-égal à 1. Cela peut s'expliquer notamment par l'émergence de nouveaux instruments de musique contribuant à la diminution de l'acoustique.

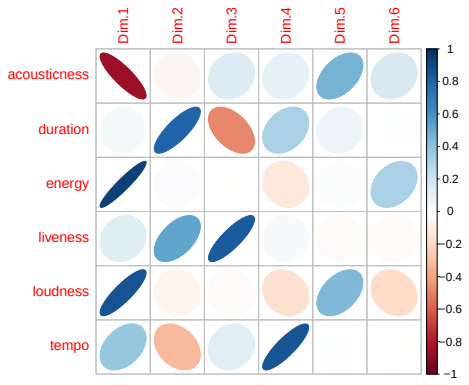
4. Analyse en composantes principales

L'analyse en composantes principales (ACP) permet de résumer et de visualiser les informations dans un ensemble de données contenant des individus décrits par plusieurs variables quantitatives intercorrélées. Chaque variable peut être considérée comme une dimension différente. L'ACP exprime les informations extraites du jeu de données sous la forme d'un ensemble de nouvelles variables appelées composantes principales. Ces nouvelles variables correspondent à une combinaison linéaire des variables d'origine. Dans ce projet, nous avons décidé de réaliser une ACP centrée réduite, car les variables dont nous disposons ont des échelles assez étalées et s'expriment dans des unités différentes (années, bpm...). Les données que nous étudions avec l'ACP sont les variables quantitatives, i.e. **acousticness**, **duration**, **energy**, **liveness**, **loudness**, **tempo**. A la fin de notre analyse, nous chercherons à intégrer les informations des variables quantitatives afin d'en tirer une meilleure interprétation de nos résultats.

```
spotify2=spotify[,-c(1,5,6,9,11)] #pour supprimer les colonnes
write.table(spotify2,file="spotify2.txt",row.names=TRUE,col.names=TRUE)
x <- read.table("spotify2.txt", header=TRUE) #stocke les variables quantitatives
res.acp <- PCA(x,scale.unit=TRUE,ncp=6,graph=FALSE)
```

4.1. Choix des composantes principales

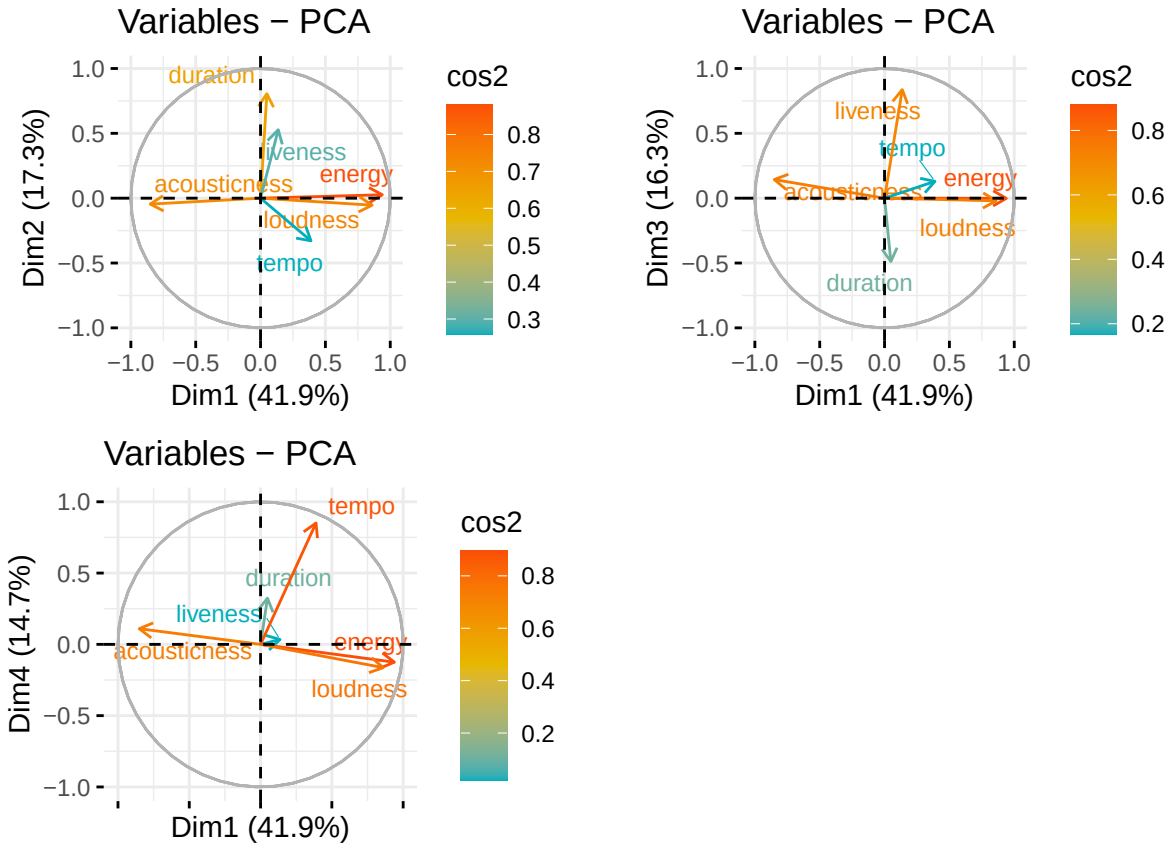
```
eig.val <- get_eigenvalue(res.acp)
p0 = fviz_eig(res.acp, addlabels = TRUE, ylim = c(0, 50))
corrplot(res.acp$var$cor, method="ellipse")
```



A l'affichage des inerties portées par chaque composante, il semblerait que les composantes 1 à 4 soient les plus importantes et que la 5ème et la 6ème puissent être négligées. En effet, les 4 premières composantes principales cumulent 90.2% de l'inertie initiale. Cela se vérifie quand on trace le graphe des corrélations, puisqu'on voit que les fortes corrélations sont toutes distribuées sur les composantes 1 à 4, et que la 5 et 6 ne portent que des corrélations faibles. Cependant, le choix de garder la quatrième composante principale est critiquable dans le sens où l'on pouvait se limiter à près de 80% de l'inertie. Cependant, au vu du graphe des corrélations, nous remarquons que seule la dimension 4 pouvait expliquer la variable **tempo** car cette dernière est uniquement très bien corrélée à cette dimension.

4.2. Graphe des variables

```
var <- get_pca_var(res.acp)
########Composante 1 contre composante 2########
A = fviz_pca_var(res.acp, col.var = "cos2",repel = TRUE, gradient.cols =
                   c("#00AFBB","#E7B800","#FC4E07"), axes=c(1,2),labelsize = 3)
########Composante 1 contre composante 3########
B = fviz_pca_var(res.acp, col.var = "cos2",repel = TRUE, gradient.cols =
                   c("#00AFBB", "#E7B800","#FC4E07"), axes=c(1,3), labelsize = 3)
########Composante 1 contre composante 4########
C = fviz_pca_var(res.acp, col.var = "cos2",repel = TRUE, gradient.cols =
                   c("#00AFBB", "#E7B800","#FC4E07"), axes=c(1,4), labelsize = 3)
ggarrange(A, B, C + rremove("x.text"), ncol = 2, nrow = 2)
```



Composante 1 contre composante 2:
On remarque que les variables **energy** et **loudness** sont fortement corrélées positivement avec la composante 1 grâce, d'une part, au graphe des corrélations (ellipse applatie de couleur bleu foncée) et d'autre part, au graphe des variables (les vecteurs ont un angle faible avec l'axe mais aussi une norme proche de 1). D'ailleurs, cela confirme l'analyse bidimentionnelle car ces deux variables ont tendance à évoluer dans le même sens. Concernant, la variable **acousticness**, elle est corrélée négativement avec la composante 1. Ainsi, **acousticness** aura tendance à varier de façon opposée à **energy** et **loudness** : lorsque ces deux dernières augmentent, elle diminue. Ce qui semble logique car des morceaux joués avec des instruments acoustiques produisent généralement moins de bruit et ont un son plus pur. **Hypothèse :** La dimension 1 se rapporte au son, plus précisément à son profil énergique et acoustique.

Sur ces graphes nous pouvons également constater la forte corrélation positive entre la variable **duration** ainsi que la dimension 2. **Hypothèse** : La dimension 2 se rapporte à la durée d'un morceau.

Remarque : C'est la dimension 1 qui explique les plus de variables (3 exactement). Ceci est logique car c'est la dimension principale et qui explique donc 42% de l'intertie.

Composante 1 contre composante 3
Contrairement à l'analyse précédente, **duration** est corrélée négativement sur la composante 3 tandis que **liveness** est

corrélée positivement, elles auront donc tendance à évoluer de manière opposée si l'on regarde l'information portée par la composante 3. Cependant, la corrélation positive de duration sur la 2ème composante est plus forte que sa corrélation négative sur la 3ème composante, duration est donc majoritairement expliquée par la 2ème composante. **Hypothèse**: La dimension 3 représente si un concert présente des parties en live ou pas.

Composante 1 contre composante 4 On constate que **tempo** est fortement corrélée positivement avec la dimension 4. **Hypothèse**: La dimension 4 représente le rythme musical i.e. s'il donne de l'entrain ou non.

4.3. Graphe des individus

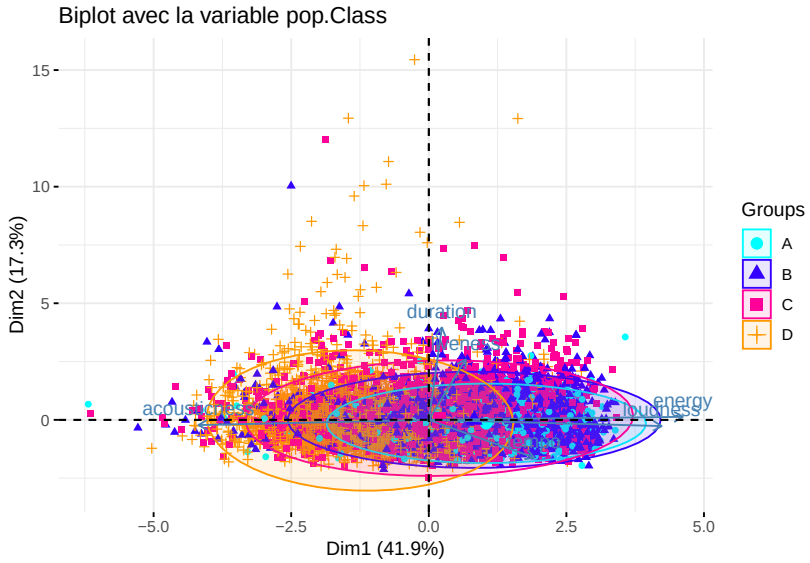
Vérifions nos hypothèses avec les différents graphes des individus. Sur le graphe, les points sont globalement tous regroupés sur la dimension 1. Les individus sont donc fortement corrélés avec la première composante et très peu avec les autres. En particulier, quand on compare les autres dimensions entre elles, on constate que les points forment globalement un amas autour de l'origine du repère, ce qui implique qu'aucune des deux composantes représentées ne porte l'information.

4.4. Les biplots : apport des informations données par les variables qualitatives

Dans l'ACP, nous devons mettre de côté les variables qualitatives, cependant la fonction *fviz_pca_biplot* permet de les prendre en compte dans l'analyse et peuvent apporter des informations aidant à la compréhension de notre jeu de données. Notamment, après avoir tracé les graphes des individus avec chaque variable qualitative, il n'y en a que deux qui ont donné des informations pertinentes : **pop.Class** et **year**.

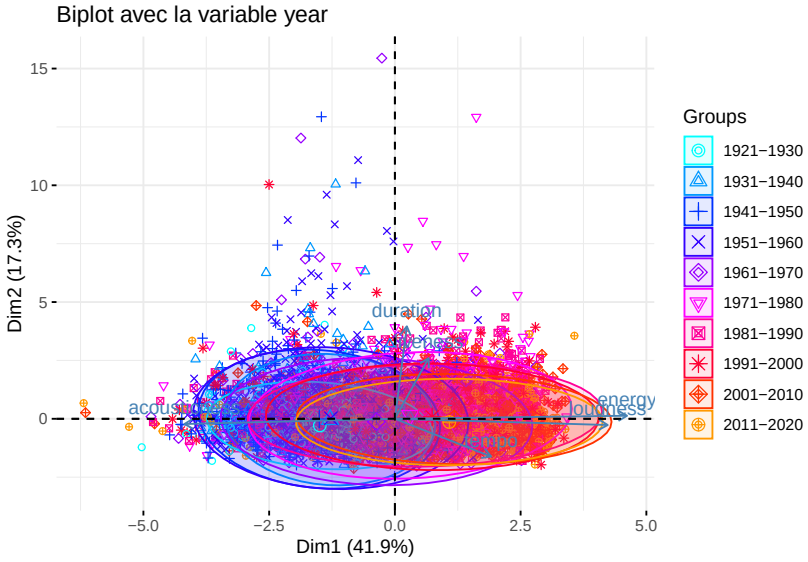
Graphe des individus catégorisé à l'aide de la variable *pop.class*

```
fviz_pca_biplot(res.acp,geom.ind = "point", col.ind = spotify$pop.class, palette = rainbow(4, start=.5,
     end=.1), addEllipses = TRUE, legend.title = "Groups", title="Biplot avec la variable pop.Class")
```



Sur ce graphe les morceaux ont été regroupés selon leur appréciation par les utilisateurs à l'aide des ellipses de concentration. Par exemple, les musiques très peu populaires (classe D) seront davantage concentrées dans l'ellipse orange. Nous observons que les musiques ayant une énergie et un volume plus importants auront tendance à être plus appréciées puisqu'elles sont concentrées dans les ellipses violette et cyan. Au contraire, les musiques acoustiques sont moins appréciées des utilisateurs car elles se retrouvent globalement dans l'ellipse orange. De plus, de la classe D à la classe A on remarque un applatissement des ellipses, ce qui montre que la durée de la musique joue aussi sur son appréciation : plus les musiques sont courtes, plus elles se voient attribuer une bonne note. Pour conclure cette pemière analyse, nous pouvons dire que les utilisateurs de Spotify préfèrent donc les musiques **énergiques**, **rythmées**, **de volume élevé**, **assez courtes** et peu **acoustiques**. Ce résultat est cohérent avec les styles de musique très répandus de nos jours comme la *pop music* et l'*electro* par exemple. Nous pouvons émettre l'hypothèse que c'est en fait l'ancienneté de la musique qui joue sur ces notations (hypothèse vérifiée sur le biplot ci-dessous).

```
V2 = f_year_biplot (spotify$year,10) ; Year2 = factor(x=V2)
fviz_pca_biplot(res.acp, geom.ind = "point", col.ind = Year2, addEllipses = TRUE, legend.title =
"Groups", palette=rainbow(10, start=.5, end=.1), title="Biplot avec la variable year")
```



Ici les morceaux ont été regroupés selon leur ancienneté. Grâce à cette réprésentation, nous pouvons dire que les musiques récentes (80-90) ont tendance à avoir une énergie et un volume sonore plus importants que les années 30-40. De même, on remarque que les musiques anciennes sont davantage acoustiques que celles d'aujourd'hui, ce qui est cohérent avec la définition d'une musique acoustique que nous avons trouvé sur Wikipédia : "*La musique acoustique n'emploie pas d'instruments électroniques modernes. Si la musique moderne recourt de plus en plus à des moyens automatisés de production sonore, comme les synthétiseurs, les sampleurs, les ordinateurs, etc., la musique acoustique, au contraire, se base sur l'emploi d'instruments de musique « classiques », qui peuvent fonctionner sans électricité.*"

Pour rappel, lors de l'analyse bidimentionnelle des variables **year** et **pop.Class**, nous avions remarqué que les musiques anciennes étaient moins appréciées des utilisateurs. Ainsi, suite à toutes ces analyses nous pouvons avancer que cela s'explique par la différence de style et d'instruments qu'utilisaient les musiciens à l'époque par rapport à aujourd'hui.

Conclusion :

A travers l'analyse du jeu de données spotify, on a donc pu constater que :
- La composante 1 porte les variables **acousticness**, **energy** et **loudness** qui sont corrélées deux à deux : **acousticness** est corrélée négativement avec **energy** et **loudness**, tandis que **loudness** et **energy** sont corrélées positivement.
- La composante 2 porte la variable **duration** et la composante 3 porte la variable **liveness**. Ces deux variables sont liées : le fait d'enregistrer un morceau en live impacte sa durée.
- La composante 4 porte la variable **tempo** qui n'est corrélée avec aucune autre variable. L'analyse unidimensionnelle a montré que le tempo moyen était autour de 117 bpm, caractérisant un rythme dansant.
- D'après l'analyse unidimensionelle, les chansons les plus populaires sont les chansons récentes. La popularité est par ailleurs corrélée négativement à la variable **acousticness** et positivement aux variables **energy** et **loudness**. En effet, les musiques actuelles sont souvent énergétiques, rythmées et de volume élevé tandis que les musiques anciennes sont davantage acoustiques étant donné qu'elles n'utilisent pas d'instruments de musique électriques.
- D'après l'analyse bidimensionelle, les variables **key** et **mode** sont liées, ce qui est surprennant car débuter un morceau par une certaine note n'impose pas au reste du morceau de rester dans la tonalité de cette note.
- Les variables **mode** et **explicit** sont identiques, elles ont exactement la même répartition : un morceau commençant par une tonalité mineure aura un contenu non vulgaire. Ce qui n'est bien sûr pas toujours le cas.

Il aurait été intéressant notamment de prendre en compte l'âge des utilisateurs et leurs notations pour voir si finalement ces résultats sont peut-être biaisés dans le sens où si ce sont des jeunes adultes ou adolescents qui votent, ils auront tendance à voter pour les musiques actuelles, ce qui n'est pas forcément le cas des personnes plus âgées qui préféreront peut-être les musiques de leur époque.