Analyse descriptive du jeu de données Spotify

Projet en Statistique descriptive

Membres

LOULIDI Younes

PHAM Tuan Kiet

VO Van Nghia

Date

17 Mars, 2021

Table des matières

Table des figures

1 Statistiques descriptives unidimensionnelle et bidimensionnelle

1.1 La nature des jeux de données

1.1.1 Des jeux de données

Ces jeux de données se composent de 10000 chansons extraites de la base de données Spotify. Chaque ligne contient 11 variables statistiques comme suit:

- `year`: année de sortie du morceau,
- `acousticness`: métrique relative interne de l'acoustique morceau,
- `duration`: durée du morceau en millisecondes (ms),
- `energy`: métrique relative interne de l'intensité, des rythmes du morceau,
- `explicit`: vaut 1 si le morceau contient des vulgarités, et 0 sinon,
- `key`: tonalité en début de morceau,
- `liveness`: proportion du morceau où l'on entend un public,
- `loudness`: mesure relative du volume du morceau (en décibels, dB)
- `mode`: mode du morceau (0 si la tonalité est mineure, et 1 si la tonalité est majeure),
- `tempo`: le tempo du morceau, en battement par minute (bpm),
- `pop.class`: la popularité du morceau.

1.1.2 Des variables statistiques

Ici, nous précisons la nature de chaque variable et son format dans R.

Nom de variable statistique	Type de variable	Format dans R
`year`	qualitative ordinale	`integer`
`acousticness`	quantitative continue	`numeric`
`duration`	quantitative discrète	`numeric`
`energy`	quantitative continue	`numeric`
`explicit`	qualitative nominale	`logical`
`key`	qualitative nominale	`factor`
`liveness`	quantitative continue	`numeric`

Nom de variable statistique	Type de variable	Format dans R
loudness	quantitative continue	numeric
mode	qualitative nominale[1]	logical
tempo	quantitative continue	numeric
pop.class	qualitative ordinale	ordered

1.1.3 Charger les jeux de données dans R

```
LoadDataset <- function(fname) {
    colclasses <- c(
        "integer", "numeric", "numeric",
        "numeric", "integer", "factor", "numeric",
        "numeric", "integer", "numeric", "factor"
    )
    dataframe <- read.csv(fname, colClasses = colclasses)
    dataframe$explicit <- as.logical(dataframe$explicit)
    dataframe$mode <- as.logical(dataframe$mode)
    dataframe$pop.class <- ordered(dataframe$pop.class)
    return(dataframe)
}
daf <- LoadDataset("dataset.csv")
str(daf)
```

```
## 'data.frame':    10000 obs. of  11 variables:
##  $ year        : int  1998 1992 1973 1969 2008 2015 1935 1928 2013 1945 ...
##  $ acousticness: num  0.147 0.193 0.388 0.733 0.979 0.0742 0.99 0.995 0.000506 0.98 ...
##  $ duration    : num  148520 189800 289267 170267 438907 ...
##  $ energy      : num  0.74 0.389 0.856 0.454 0.494 0.766 0.42 0.211 0.53 0.106 ...
##  $ explicit    : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ key         : Factor w/ 12 levels "A","Ab","B","Bb",..: 7 5 1 10 11 11 2 6 12 12 ...
##  $ liveness    : num  0.0452 0.154 0.139 0.0889 0.123 0.0827 0.13 0.106 0.0477 0.237 ...
```

[1] On pose FALSE si la tonalité est mineure et TRUE si non.

```
## $ loudness    : num  -8.16 -11.64 -8.4 -8.12 -10.65 ...
## $ mode        : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ tempo       : num  157.1 85.3 101.3 82.4 156.3 ...
## $ pop.class   : Ord.factor w/ 4 levels "A"<"B"<"C"<"D": 3 3 3 3 3 1 4 4 2 4 ...
```

1.2 Analyses unidimensionnelles

1.2.1 Une variable qualitative - `pop.class`

```
summary(daf$pop.class)
```

```
##    A    B    C    D
##  940 2874 3038 3148
```

Il existe 4 niveaux de popularité (modalités). Commencer par `A` est le plus populaire et décroissant avec `B`, `C`, `D`.

```
pop_class_table <- table(daf$pop.class)
print(label_percent()(c(pop_class_table) / sum(pop_class_table)), quote = F)
```

```
##     A     B     C     D
##  9.4% 28.7% 30.4% 31.5%
```

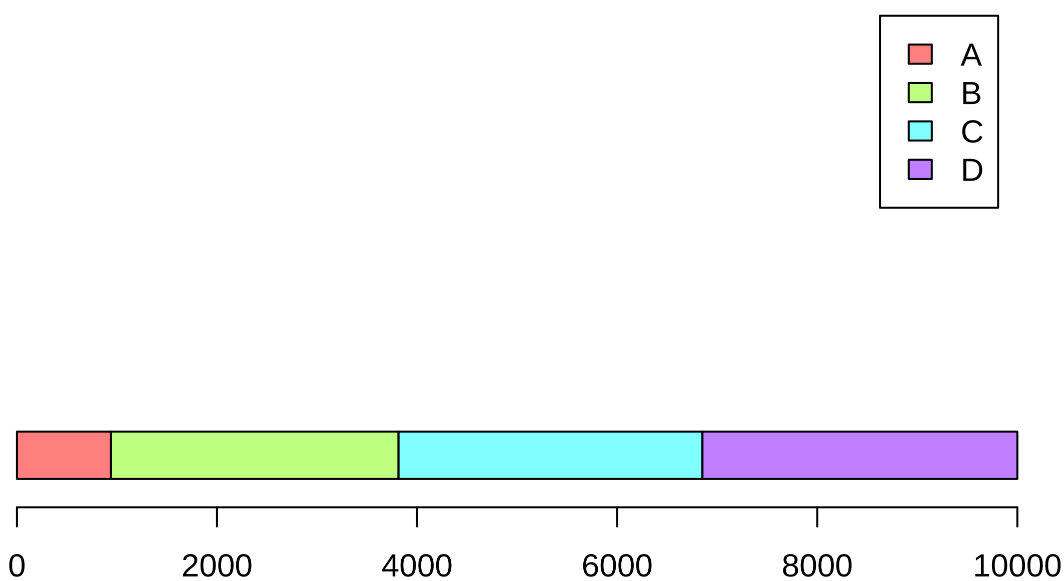


Figure 1: Diagramme en barre de popularité

On peut noter que dans cet ensemble de données, la plupart des chansons ne sont pas populaires $(31, 5)$. Plus le niveau de popularité est élevé, moins les chansons peuvent atteindre ce niveau.

1.2.2 Une variable quantitative - `acousticness`

1.2.2.1 Résumé

```
summary(daf$acousticness)
```

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

D'après le résultat ci-dessus, on a:

- Le premier quartile $q_{0.25}$ est 0.0961
- Le deuxième quartile $q_{0.5}$ est 0.5085
- Le troisième quartile $q_{0.75}$ est 0.8930

1.2.2.2 Distribution

```
skewness(daf$acousticness)
```

```
## [1] -0.01816556
```

Étant donné que son skewness est approximativement 0, nous pouvons conclure que l'ensemble de données est centré autour de sa médiane.

```
kurtosis(daf$acousticness)
```

```
## [1] -1.613205
```

Du fait que son kurtosis est inférieur à $-1, 2$ (le kurtosis de la distribution uniforme[2]), sa distribution aura la forme d'une vallée (car la distribution uniforme est déjà une ligne).

[2] https://en.wikipedia.org/wiki/Kurtosis#Other_well-known_distributions

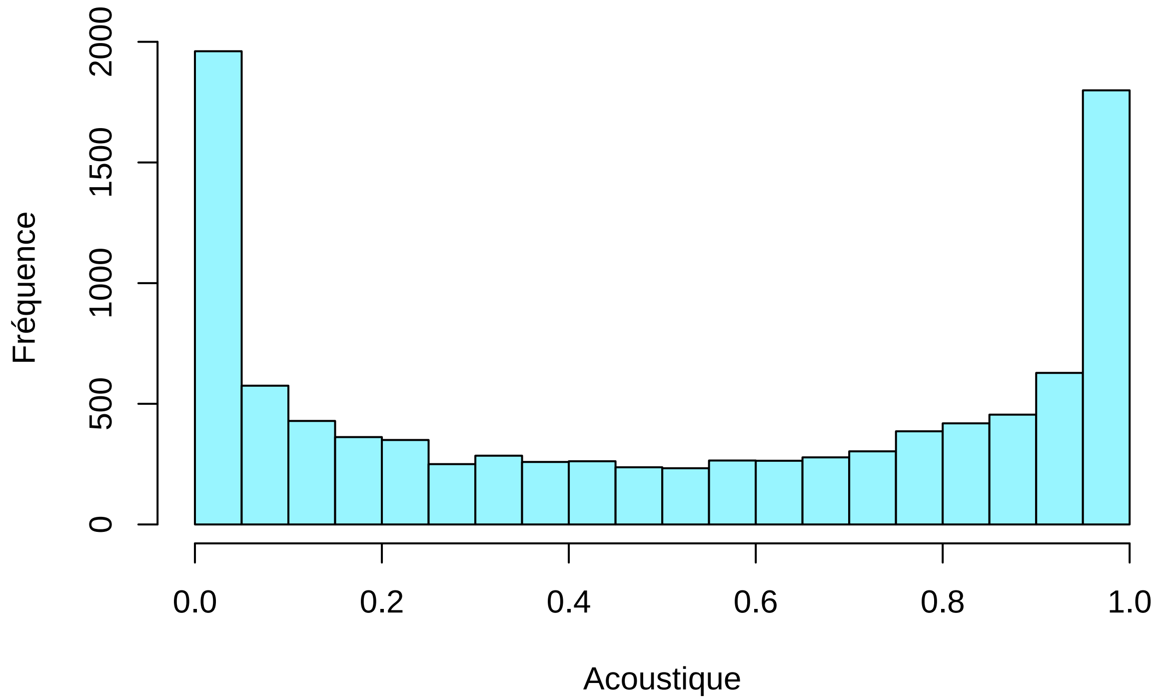


Figure 2: Histogramme d'acoustique des chansons

Vous pouvez voir toutes les caractéristiques mentionnées ci-dessus dans la figure 2.

1.3 Analyses bidimensionnelles

1.3.1 Entre une variable quantitative et une qualitative

Dans cette partie, nous réutiliserons et analyserons les 2 variables précédentes (`pop.class` et `acousticness`).

1.3.1.1 Représentation graphique

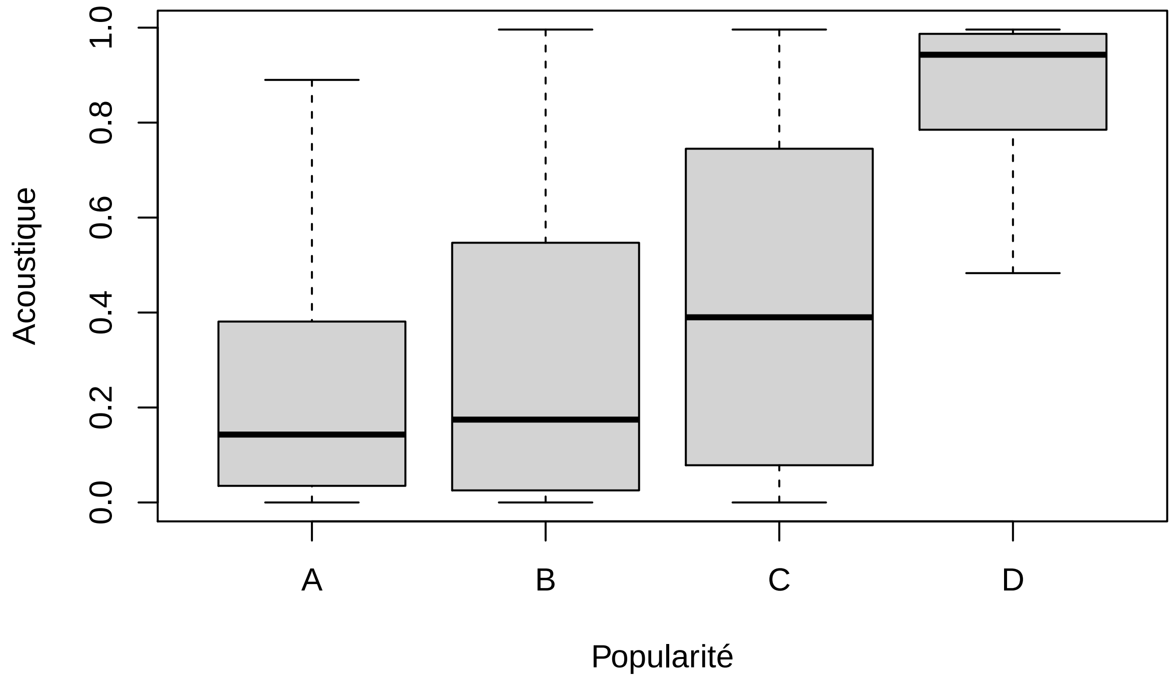


Figure 3: Boxplot parallèles de la relation entre acoustique et popularité

Notez à partir de notre graphique, les cases varient d'un facteur à l'autre, nous concluons que l'acoustique et la popularité sont liées l'une à l'autre.

De plus, 75% des chansons populaires ont une acoustique inférieure à $0,5$ tandis que celle de presque 100% des chansons les moins populaires est supérieure à $0,5$. De l'autre côté, selon la partie précédente, la distribution des chansons avec l'acoustique est symétrique autour de sa médiane (ce qui indique qu'il y a presque le même nombre de chansons de 2 types). Il est démontrable que les gens aiment les chansons électroniques.

1.3.1.2 Indice de liaison

```
eta2(daf$acousticness, daf$pop.class)
```

```
## [1] 0.3673706
```

Avec $c_{y|x} \approx 0,4$, il existe une légère relation entre deux variables.

1.3.2 Entre deux variables quantitatives

Dans cette partie, nous étudions la relation entre le volume et l'énergie.

1.3.2.1 Représentation graphique

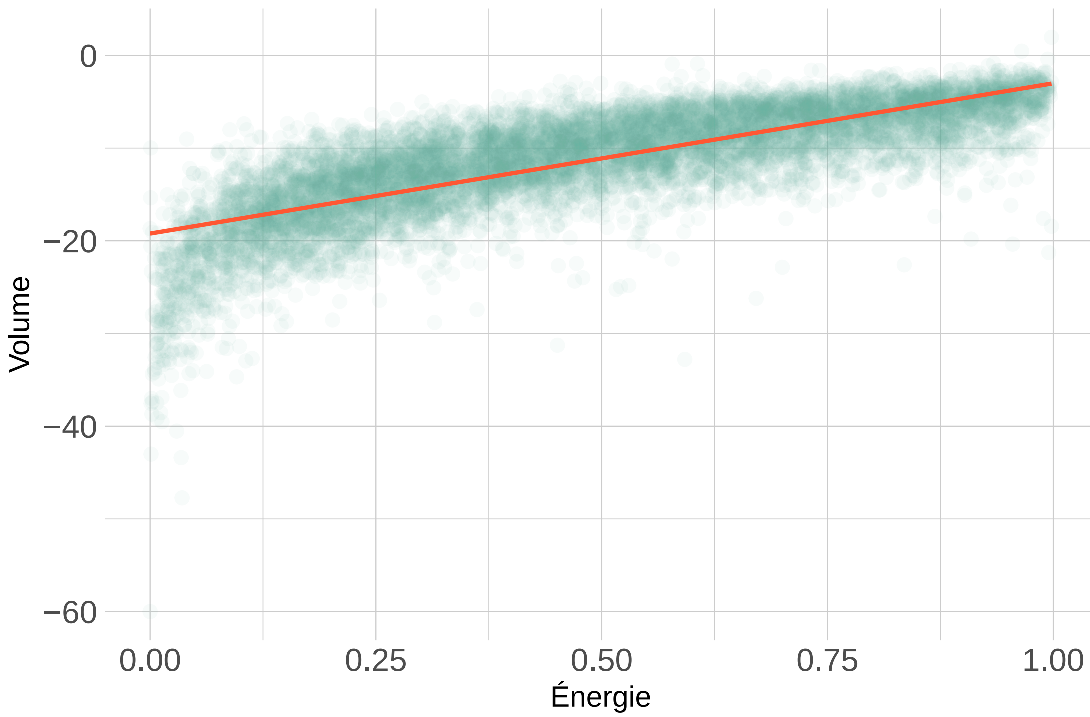


Figure 4: Nuage de points et droite de régression entre l'énergie et le volume

1.3.2.2 Indices de liaison

```
cor(daf$energy, daf$loudness)
```

```
## [1] 0.7744876
```

Avec cette valeur de corrélation et à partir de la figure 4, nous en déduisons qu'il existe un lien fort entre le volume d'une chanson et son énergie: plus le volume d'une chanson est forte, plus elle a d'énergie.

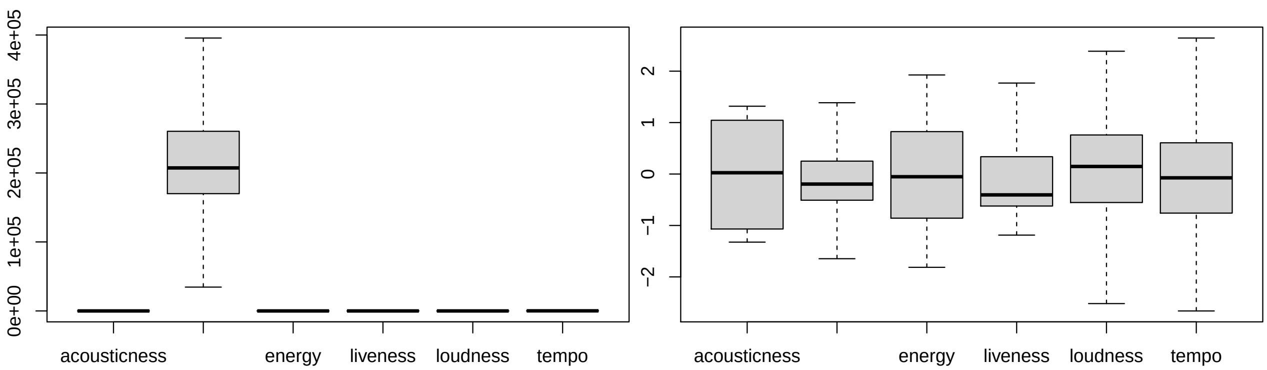
2 Analyse en composantes principales (ACP)

Tout d'abord, nous supprimerons toutes les colonnes qualitatives de l'ensemble de données (qui sont `year`, `explicit`, `key`, `mode` et `pop.class`)

```
dafacp <- select_if(daf, is.numeric)

dafacp <- dafacp[,-1] # remove the first column of `dafacp` (which is `year`)
```

2.1 Choix du type d'ACP réalisé



(a) Des données brutes

(b) Des données centrées réduites

Figure 5: Boxplots des données quantitatives

Ici, le choix de faire une **ACP centrée réduite** s'impose pour deux raisons :

- Les données des différentes variables ne sont pas du tout à la même échelle comme on peut voir dans le figure 5.
- Elles ont des unités différentes.

2.2 FactoMineR

Nous allons, à partir d'ici, utiliser la librairie `FactoMineR` pour effectuer l'ACP en incluant directement l'étape de centrer réduire faite plus haut. L'idée est de se dire que parmi nos 6 variables, nous avons de l'information redondante et donc de passer en dimension plus faible (2 ou 3 pour faciliter la représentation) grâce à des méta-variables. Seulement, nous ne pouvons choisir ces méta-variables au hasard, elles doivent correspondre aux directions selon lesquelles on a le plus de variabilité. La variabilité totale, qui est l'inertie, est répartie entre les 6 dimensions.

```
result_acp <- PCA(daf, scale.unit = TRUE, ncp = 6, quali.sup = c(1, 5, 6, 9, 11), graph = FALSE)
```

Avec FactoMineR, nous effectuons ici une ACP à 6 dimensions (`ncp = 6` pour nos 6 variables quantitatives) et nous rajoutons les autres 5 variables comme variables qualitatives supplémentaires (elles nous serviront lors de l'interprétation).

```
result_acp$eig[,"eigenvalue"]
```

```
##    comp 1    comp 2    comp 3    comp 4    comp 5    comp 6
## 2.5169626 1.0407603 0.9768164 0.8844294 0.4183029 0.1627285
```

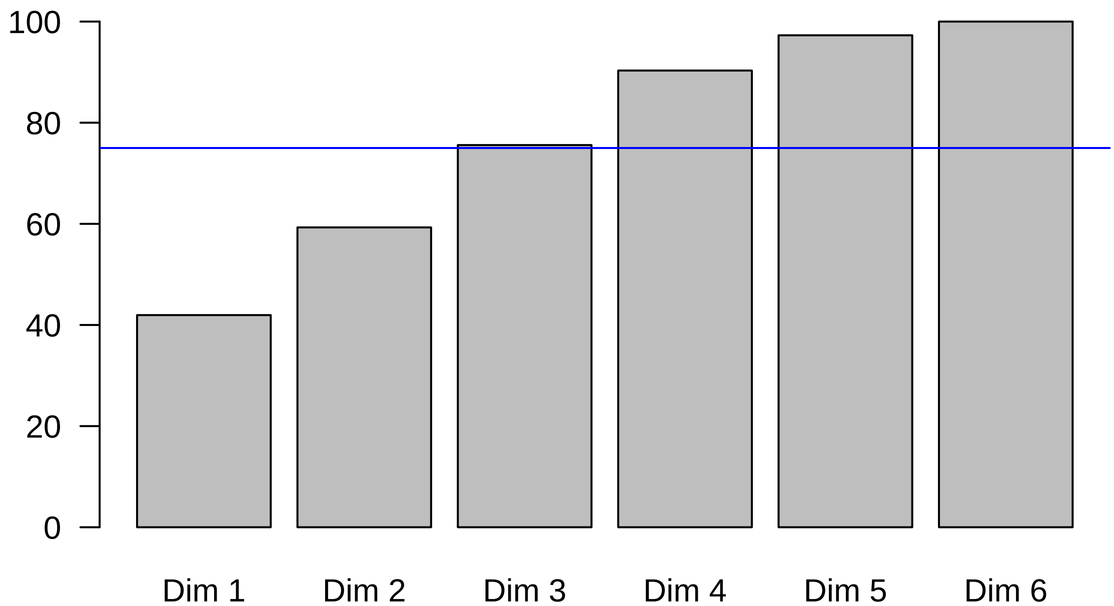


Figure 6: Pourcentages cumulés d'inertie portés par chaque axe

Regardons plus particulièrement les valeurs propres et les pourcentages d'inertie associés à chaque dimension. Comme on s'y attend, l'ACP classe les dimensions de la plus influente à la moins influente. Ici, nous choisissons de garder 75% de variabilité (ce qui est déja un très bon seuil). Ce dernier est atteint par la dimension 3 comme l'indique la droite bleu sur le graphe de pourcentages cumulés d'inertie. Nous prenons donc les 3 premières dimensions.

2.3 Etude des individus

Pour étudier les individus, on se rappelle qu'une ligne du tableau de départ correspond à un individu qu'on veut représenter par un point sur un graphique. Ainsi, à l'issue des 10000 lignes nous auront notre nuage de 10000 individus points représentant les individus. Cela aurait été simple avec deux varibles x et y, le point serait représenter dans un plan à 2 dimensions (de même avec 3 variables en 3 dimensions). Mais nous avons 6 variables qui définissent chaque individu ici, nous sommes donc passées par l'ACP qui a réduit cela à 3 dimensions. Nous allons regarder nos individus dans ces trois dimensions.

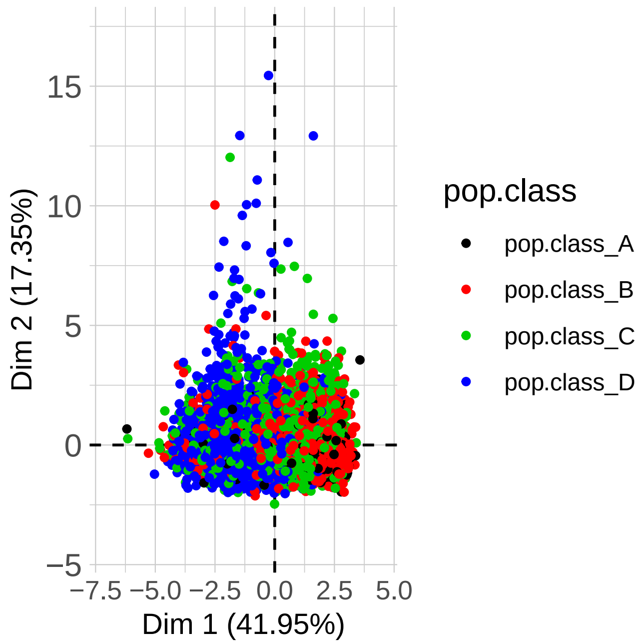


Figure 7: Projection des chansons selon leur popularité dans le plan des axes principaux 1 et 2

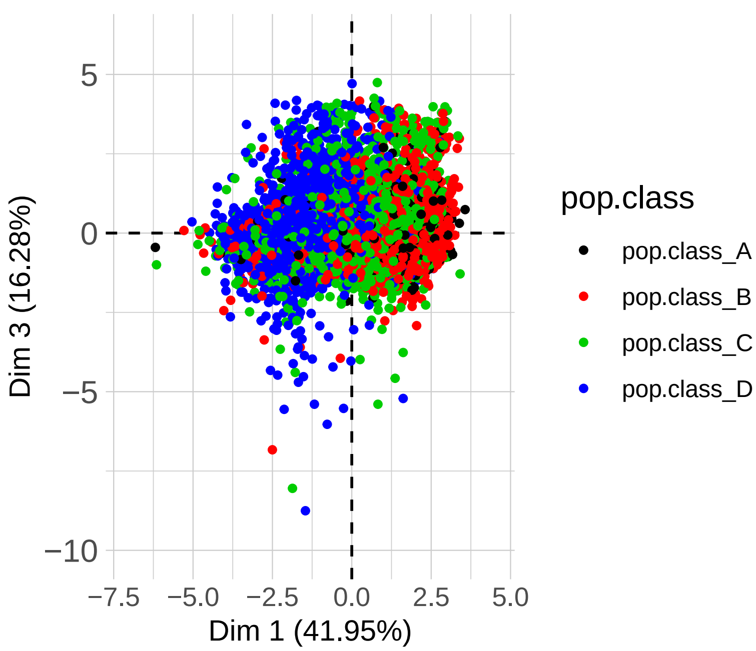
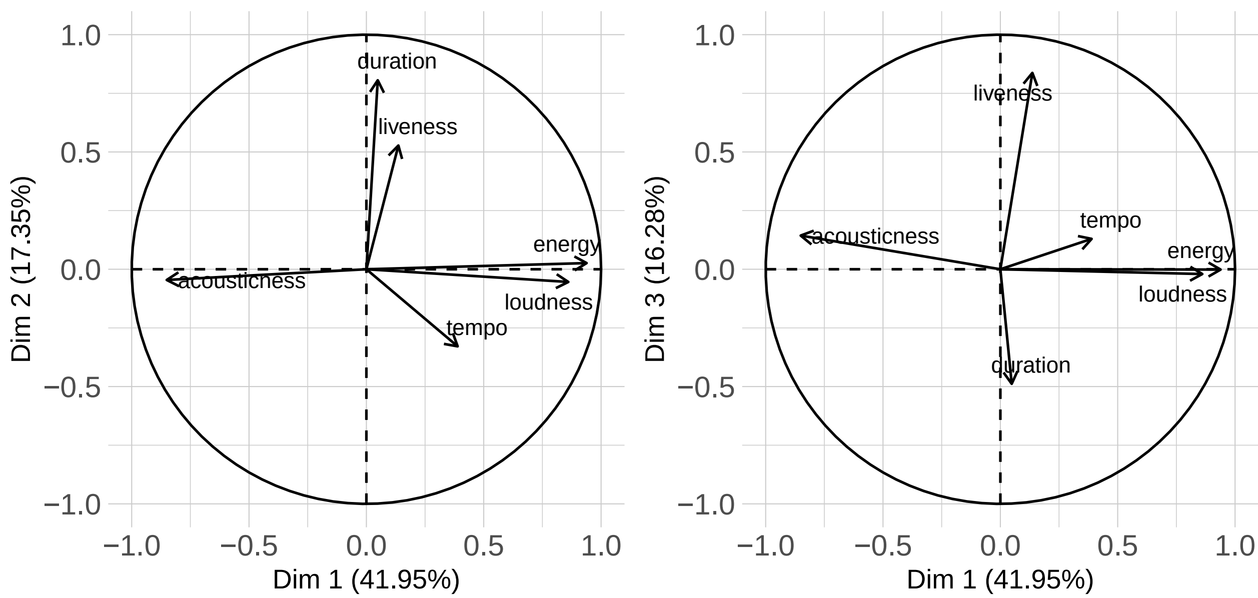


Figure 8: Projection des chansons selon leur popularité dans le plan des axes principaux 1 et 3

On a plusieurs nuages de points et on voit une tendance qui ressortent par lecture graphique. A partir des figures 7 et 8, on observe que la dimension 1 semble être assez représentative de la popularité d'une chanson: bleu quand les coordonnées sont diminués sur la dimension 1 et autre couleurs quand elles augmentent.

2.4 Etude des variables

On représente le graphe des corrélations des variables:



(a) les dimensions 1 et 2

(b) les dimensions 1 et 3

Figure 9: Graphe de corrélation entre les variables

D'après la lecture de la figure 9, nous avons les relations suivantes:

- En suivant la direction positive de l'axe de dim 1, nous avons une augmentation de l'énergie et du volume de la chanson (ce qui est également logique car les 2 variables ont une relation forte selon analyses bidimensionnelles entre le volume et l'énergie)
- Dans le sens opposé, l'acoustique de la chanson monte. Il est également cohérent avec le fait que l'acoustique d'une chanson est généralement inversement proportionnelle à son énergie et à son volume.
- La durée de la chanson pourrait être représentée à la fois par la dimension 2 et la dimension

3.

- La même chose est vraie pour la vivacité (`liveness`).

On peut retrouver ces liaisons en observant la figure 10.

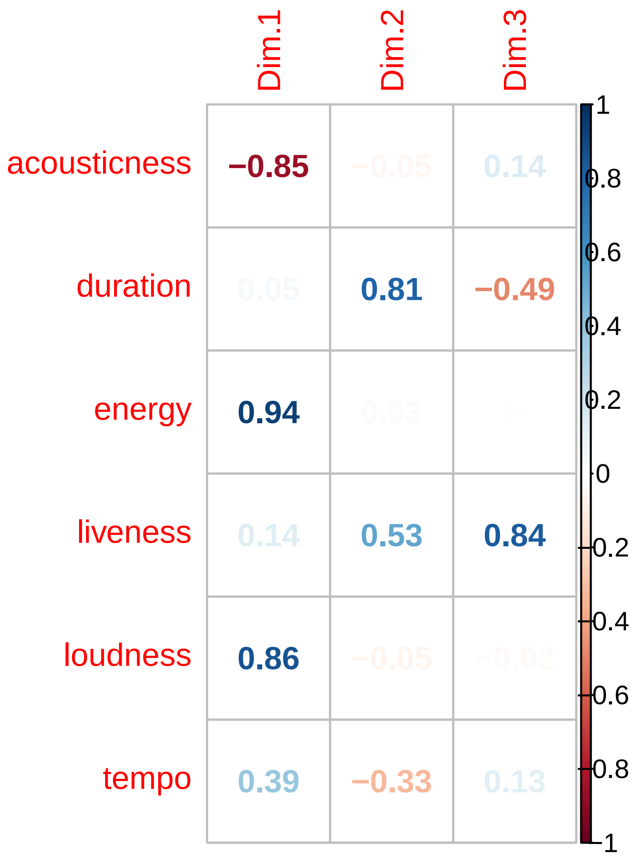


Figure 10: Le graphe de corrélation entre variables et dimensions