Analyse d'une Base de Données musicale "Spotify Audio Features"

Rapport final

Projet D'Analyse de Données

Najwa Mourssi, Paolo Conti

Polytech Sorbonne
MAIN 4
2020
Paris, France

Analyse d'une Base de Données musicale "Spotify Audio Features"
Rapport final
Projet D'Analyse de Données
Version n°1

Nous remercions
Le Professeur Fanny Villiers pour son rôle d'encadrante et son aide précieuse.

Table des matières

Table des figures

1. Abstract

ABSTRACT

Aujourd'hui, Spotify est certainement l'une des plateformes musicales les plus utilisées et les plus appréciées. Plus de 40 millions de titres, réalisés par plus de 2 millions d'artistes, sont disponibles, mais seuls quelques milliers sont populaires auprès des utilisateurs. Ces questions viennent donc naturellement : quelles sont les propriétés communes des chansons populaires ? De quoi un artiste a besoin pour que sa chanson grimpe dans les hit-parades ?
Le but de notre projet est d'essayer de comprendre ce qui se cache derrière la popularité d'une chanson et s'il est possible de la prévoir à partir des caractéristiques des chansons. Pour ce faire, il était fondamental d'essayer de diviser les chansons en groupes, selon les genres musicaux, donc un certain regroupement était nécessaire en effectuant des clusters par classification non supervisée.
Durant l'avancement du projet, nous avons réalisé que la seule caractéristique musicale des chansons peut ne pas suffire à les décrire pleinemement.
En effet, il est beaucoup plus facile que votre chanson devienne un succès mondial si vous êtes Rihanna ou Eminem, plutôt qu'un artiste de rue ! Nous avons donc ajouté une nouvelle variable à notre ensemble de données, qui tient compte de la popularité antérieure de l'artiste.
Ensuite, nous procéderons à la construction d'un modèle linéaire afin de prédire la popularité d'une chanson donnée.
Les utilisateurs préfèrent-ils de la musique triste ou joyeuse ? Musique énergique ou musique calme ?
Dans la section Conclusion, nous sommes enfin en mesure de donner une réponse aux questions qui ont guidé notre analyse.

2. Description des données

La base de données est composées de 130 326 chansons (individus) et 17 variables explicatives dont 3 sont descriptives (utilisées pour identifier une chanson), 2 qualitatives et 12 quantitatives . Le jeu de données est disponible sur Kaggle :

"Spotify Audio Features" : https ://www.kaggle.com/tomigelo/spotify-audio-features

Lors de la présentation de notre jeu de données, nous mettrons en évidence les transformations que nous avons opérées sur les variables et, surtout, la procédure que nous avons suivie pour *"nettoyer"* notre base de données afin de ne prendre en compte que la chanson qui nous intéresse. En effet, notre base de données était très hétérogène : d'un point de vue général, nous avons remarqué qu'elle comprenait des bruits, des livres audio, des podcasts, etc. qui ne sont pas utiles pour notre analyse car la plupart des variables ne sont pas significatives pour ce type de données. Par exemple, la popularité d'un podcast serait strictement liée à son sujet, cependant cela n'a pas de sens car aucune valeur correspondante à l'aspect dansant, l'acoustique ou la tonalité n'est à considérer puisque qu'un podcast se résume à un enchainement de mots.

2.1 Variables descriptives

Artist name : Artiste(s) de la chanson.

Track id : Spotify URI (Uniform Ressource Identifier ou identifiant uniforme de ressource) de la chanson.
Cette variable nous permet d'indentifier chaque chanson de manière unique, par conséquent nous l'avons utilisé pour éliminer les doublons dans notre base de données grâce à la commande ci-dessous :

```
data = data[!duplicated(data$track_id),]
```

Track name : Titre de la chanson.

2.2 Variables quantitatives

Acousticness : Une mesure comprise dans l'intervalle de confiance [0,0 ;1,0] indiquant si une musique est plus ou moins à caractère acoustique (*intruments bruts sans arrangement*). La distribution des valeurs pour cette caractéristique ressemble à ceci :

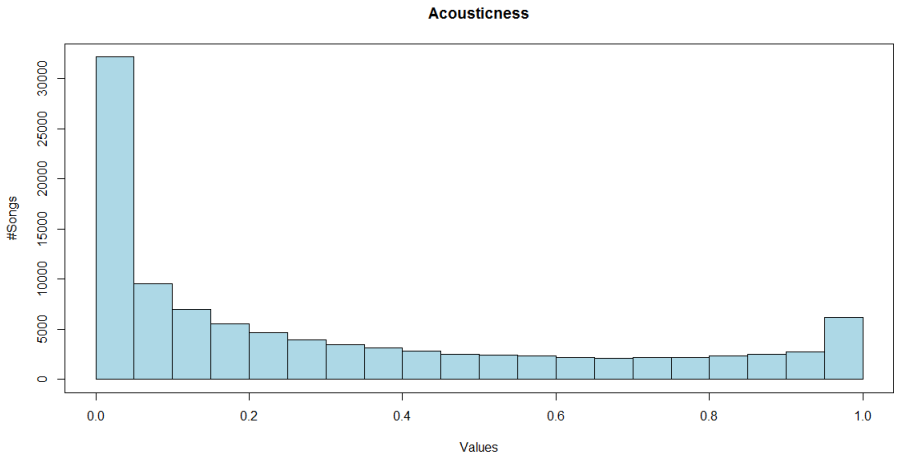

FIGURE 2.1 – Distribution de la variable Acoustique

Energy : L'énergie est une mesure de 0,0 à 1,0 et représente une mesure perceptive de l'intensité d'un titre. En général, les musiques énergétiques sont rapides, fortes et bruyantes. Par exemple, le death metal a une grande énergie, alors qu'un prélude de Bach a une faible note sur cet intervalle. Les caractéristiques perceptives contribuant à cet attribut comprennent la gamme dynamique, l'intensité sonore perçue, le timbre, le taux d'apparition et l'entropie générale. Nous avons remarqué la présence de chansons dont l'énergie était proche de 1 ou de 0 (de 0,998 à 1, et de 0 à 0,001) : en examinant ces données, nous avons découvert qu'elles n'étaient que l'enregistrement et la reproduction de sons d'eau. Comme le but de notre analyse est d'analyser la popularité, nous avons décidé de retirer ces données de notre jeu de données avec la commande :

```
data = subset(data, data$energy<0.998 & data$energy>0.001)
```

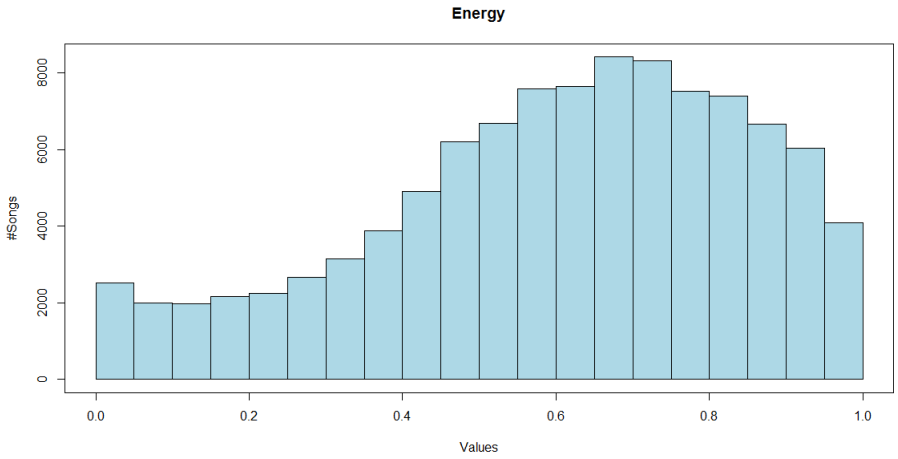

FIGURE 2.2 – Distribution de la variable Energy

Danceability : décrit la pertinence d'un morceau pour la danse en se basant sur une combinaison d'éléments musicaux tels que le tempo, la stabilité du rythme, la force du battement et la régularité générale. Une valeur de 0,0 est la moins dansante et 1,0 est la plus dansante. Nous avons supprimé les bruits en éliminant toutes les données dont les valeurs de cette variable était de 0.

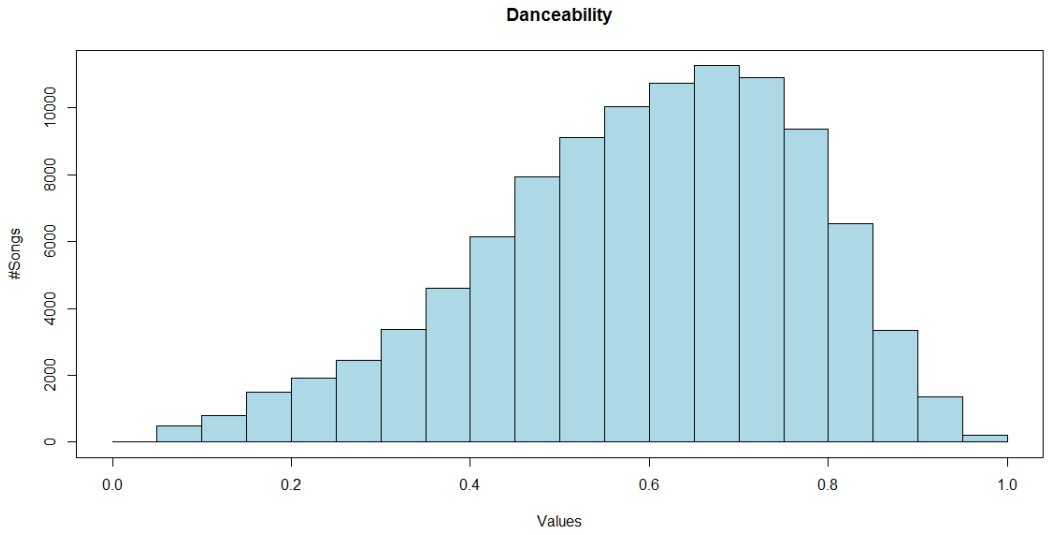

FIGURE 2.3 – Distribution de la variable Danceability

Loudness : correspond à l'intensité sonore globale d'une musique en décibels (dB). Les valeurs d'intensité sonore sont calculées en moyenne sur l'ensemble de la chanson et sont utiles pour comparer l'intensité sonore relative entre les différentes pistes. L'intensité sonore est la qualité d'un son qui est le principal corrélat psychologique de la force physique appelée amplitude. Les valeurs typiques se situent entre -60 et 0 db.

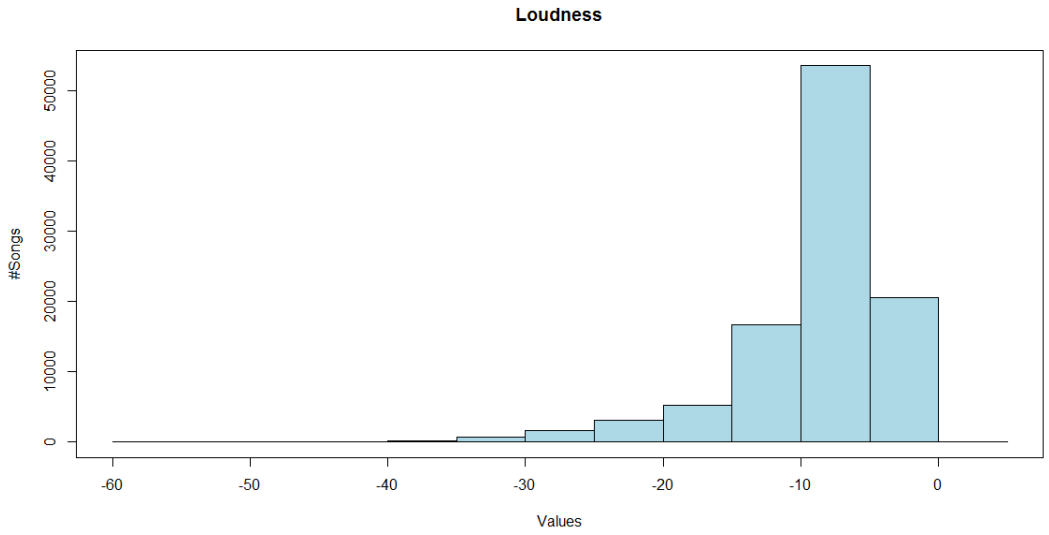

FIGURE 2.4 – Distribution de la variable Loudness

Speechiness : détecte la présence de mots parlés dans un son audio. Plus l'enregistrement ressemble à un discours (par exemple, talk-show, livre audio, poésie), plus la valeur de variable est proche de 1,0. Les valeurs supérieures à 0,66 décrivent des pistes qui sont probablement entièrement constituées de mots parlés. Les valeurs comprises entre 0,33 et 0,66 décrivent des pistes qui peuvent contenir à la fois de la musique et de la parole, soit par sections soit par couches, y compris dans des cas comme la rap. Les valeurs inférieures à 0,33 représentent très probablement de la musique et d'autres pistes contenant uniquement des sons non vocaux.

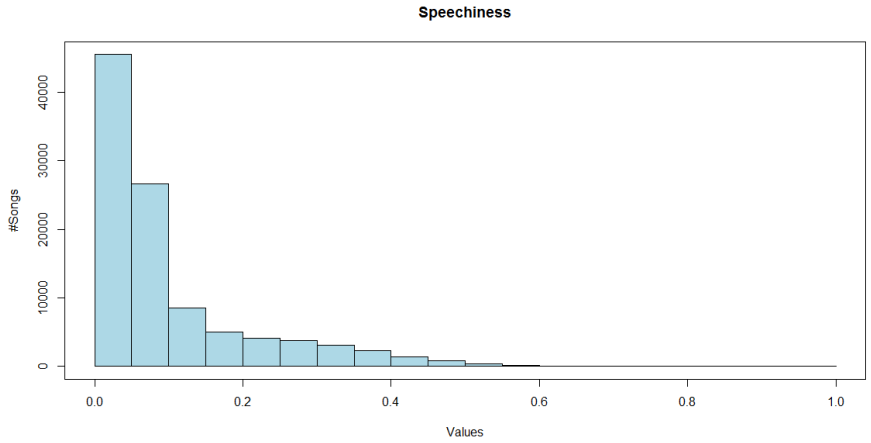

Figure 2.5 – Distribution de la variable Speechiness

Instrumentalness : prédit si un morceau ne contient pas de parties chantées. Les sons "Ooh" et "Aah" sont traités comme des instruments dans ce contexte. Les morceaux de rap ou de slam sont clairement "vocaux". Plus la valeur d'instrumentalité est proche de 1,0, plus il est probable que le morceau ne contienne pas de contenu vocal. Les valeurs supérieures à 0,5 sont censées représenter des pistes instrumentales, mais la confiance est plus élevée lorsque la valeur se rapproche de 1,0.

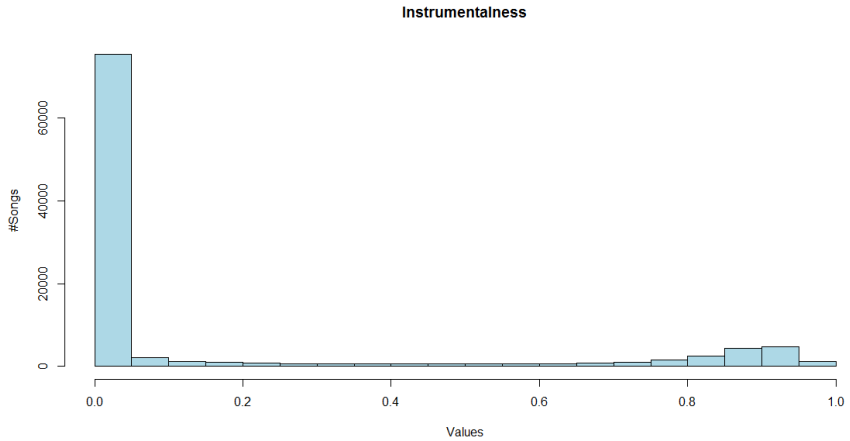

Figure 2.6 – Distribution de la variable Instrumentalness

Environ 67 % des chansons ont une instrumentalité inférieure à 0,1 : ce phénomène s'explique par une instrumentalité nulle dès qu'il y a des paroles dans la chanson.

Liveness : détecte la présence d'un public dans l'enregistrement. Des valeurs de "Liveness" plus élevées représentent une probabilité accrue que le morceau ait été joué en direct. Une valeur supérieure à 0,8 donne une forte probabilité que la piste ait été enregistré en live.

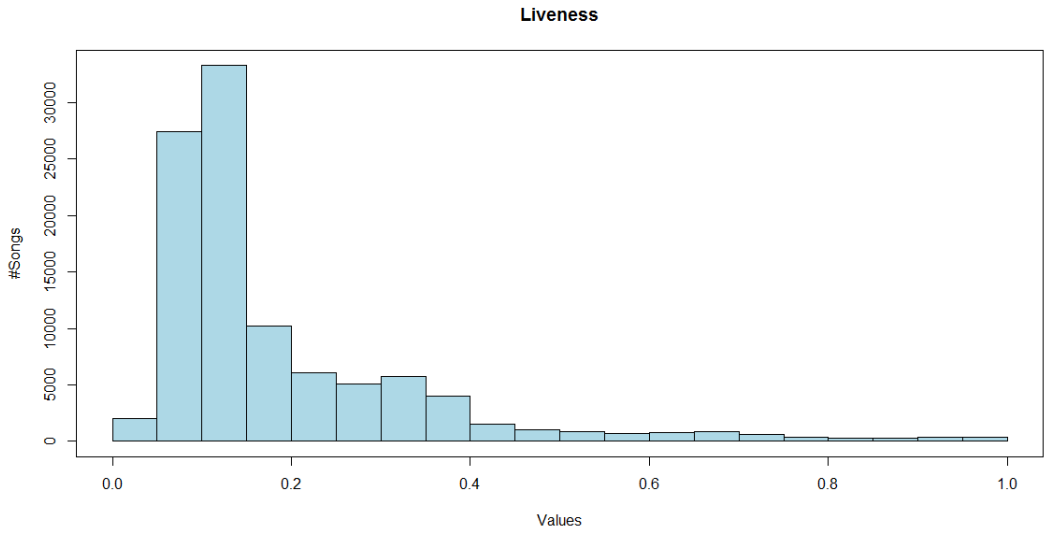

FIGURE 2.7 – Distribution de la variable Liveness

Valence : représente la mesure de 0,0 à 1,0 décrivant la positivité musicale véhiculée par un morceau. Les titres à forte valence ont un son plus positif (par exemple, heureux, joyeux, euphorique), tandis que les titres à faible valence ont un son plus négatif (par exemple, triste, déprimé, fâché).

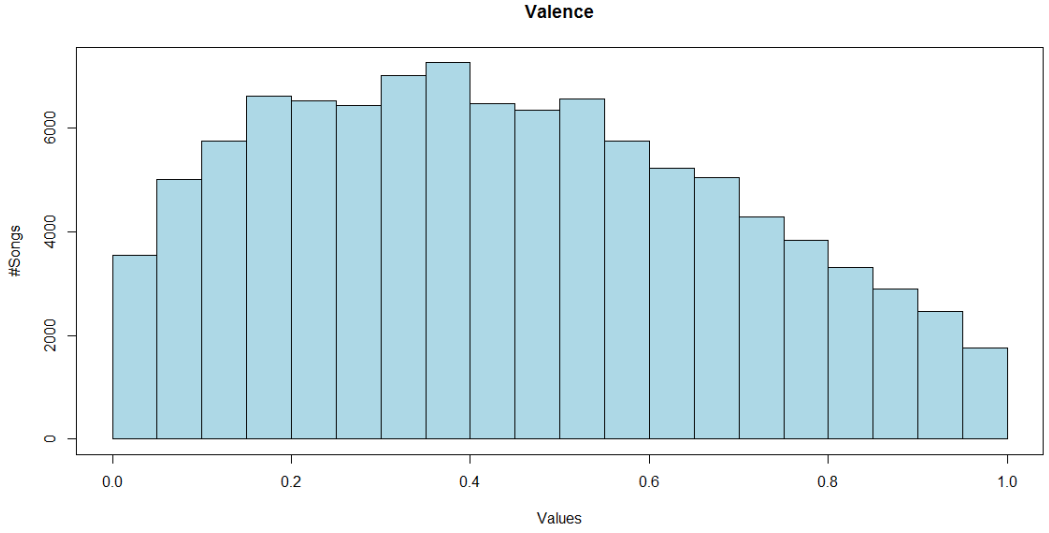

FIGURE 2.8 – Distribution de la variable Valence

Tempo : Le tempo global estimé d'un morceau en battements par minute (BPM). Dans la terminologie musicale, le tempo est la vitesse ou le rythme d'un morceau donné et découle directement de la durée moyenne des battements.

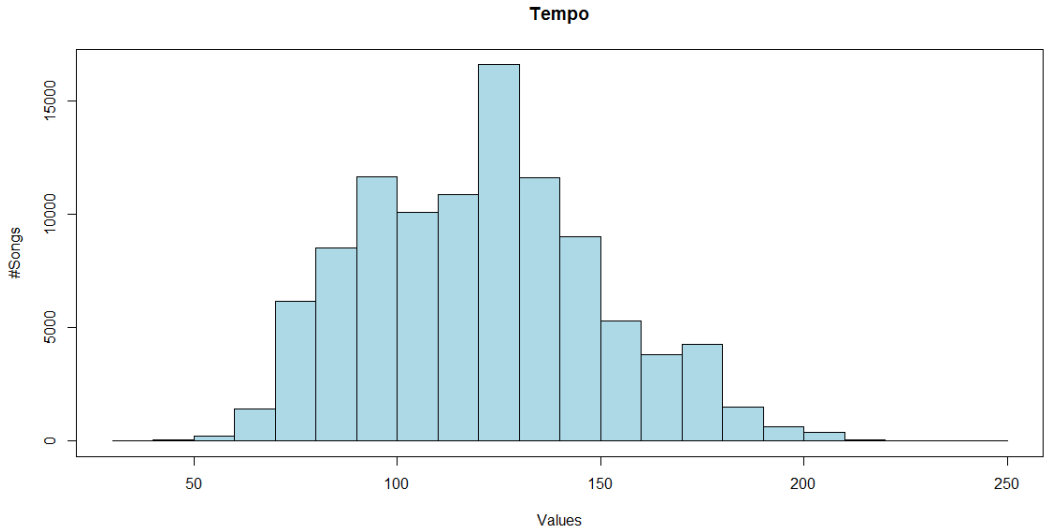

FIGURE 2.9 – Distribution de la variable Tempo

Duration : La durée du morceau en millisecondes.

Nous avons pensé qu'il serait plus judicieux de transformer les variables de millisecondes en secondes, afin d'avoir une interprétation plus claire de ses valeurs. De plus, nous avons remarqué que nous avions une très large gamme : de 3 secondes à 93 minutes ! En gardant à l'esprit que notre objectif est d'étudier la popularité des chansons, nous avons pensé qu'il serait plus approprié de nous concentrer sur les morceaux dont la durée se situe entre 2,5 et 8 minutes, nous avons donc écarté le reste avec la commande :

```
data = subset(data, data$duration>2.5*60 & data$duration<8*60)
```

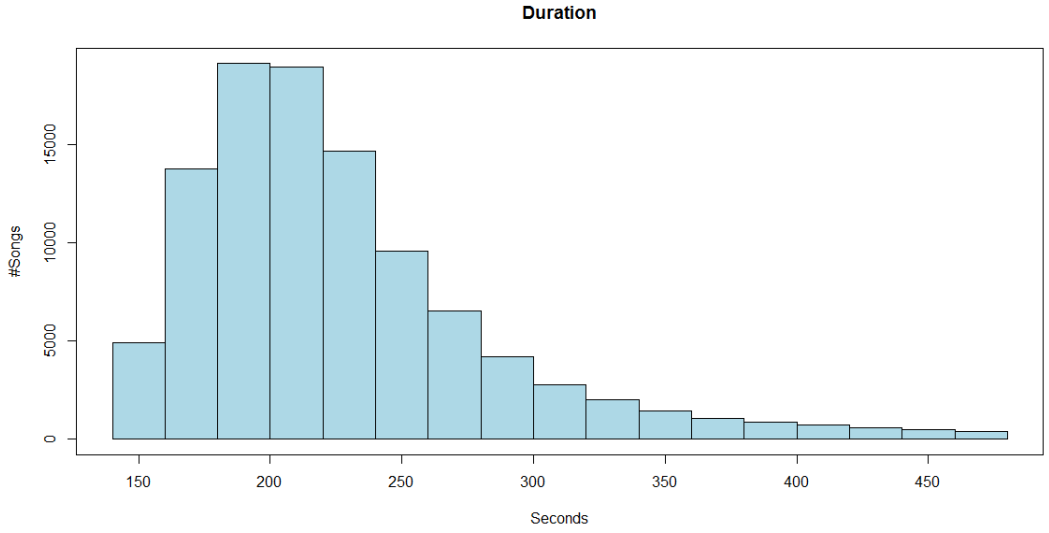

FIGURE 2.10 – Distribution de la variable Duration

Time signature : Une estimation de la signature temp globale d'un morceau. La signature temp est une convention de notation qui permet de spécifier le nombre de battements de chaque barre (ou mesure) soit le nombre et la qualité des notes contenues dans une mesure.

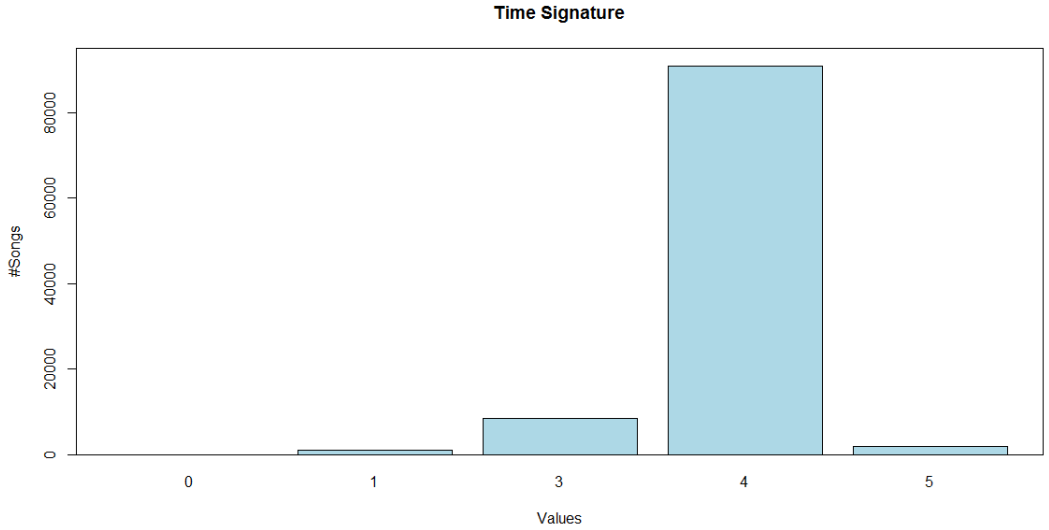

FIGURE 2.11 – Distribution de la variable Time Signature

Popularity : La popularité d'un morceau. La valeur sera comprise entre 0 et 100, 100 étant la valeur la plus haute. La popularité est calculée en fonction du nombre total d'écoutes de la chanson et de leur caractère récent. En général, les chansons qui sont beaucoup jouées aujourd'hui auront une plus grande popularité que celles qui ont été beaucoup jouées auparavant.

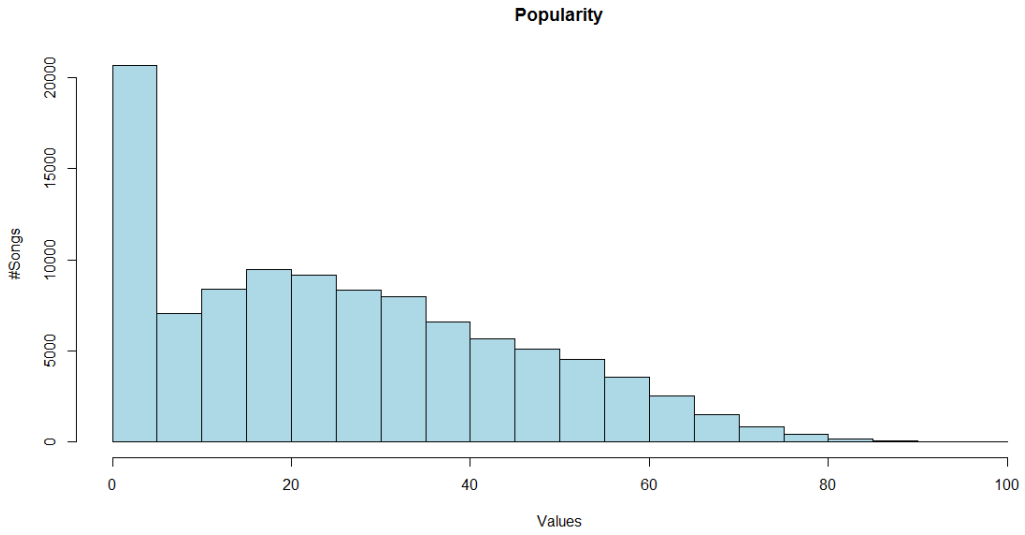

FIGURE 2.12 – Distribution de la variable Popularity

2.3 Variables qualitatives

Key : La clé ou mesure dans laquelle se trouve le morceau. Cela correspond à des nombres entiers fixant le nom et la hauteur de chaque note de musique placée sur les lignes et interlignes d'une portée (six lignes). C'est à dire, 0 = C, 1 = C#, 2 = D, etc.

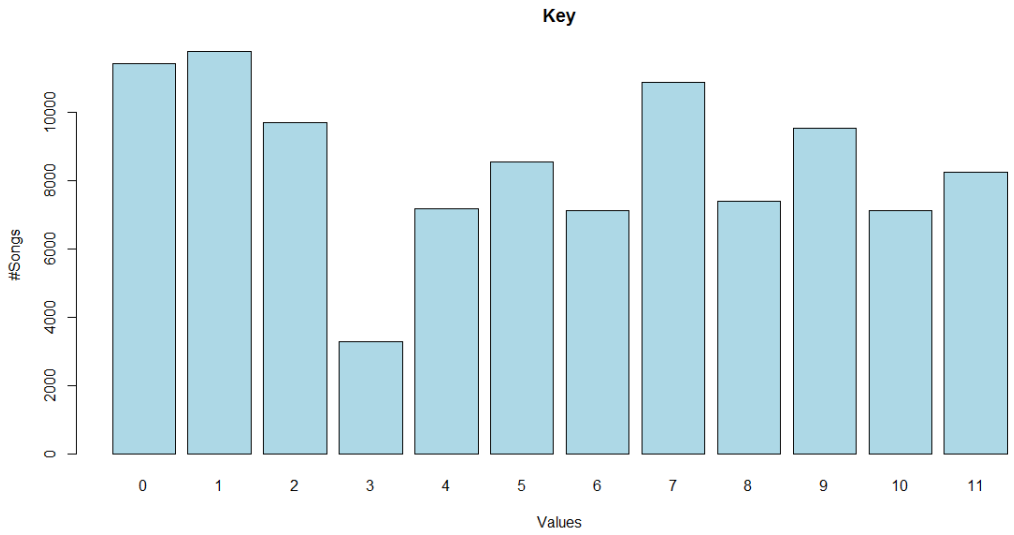

FIGURE 2.13 – Distribution de la variable Key

Mode : indique la modalité (majeure ou mineure) d'un morceau, le type de gamme dont son contenu mélodique est dérivé. Majeur est représenté par 1 et Mineur par 0 .

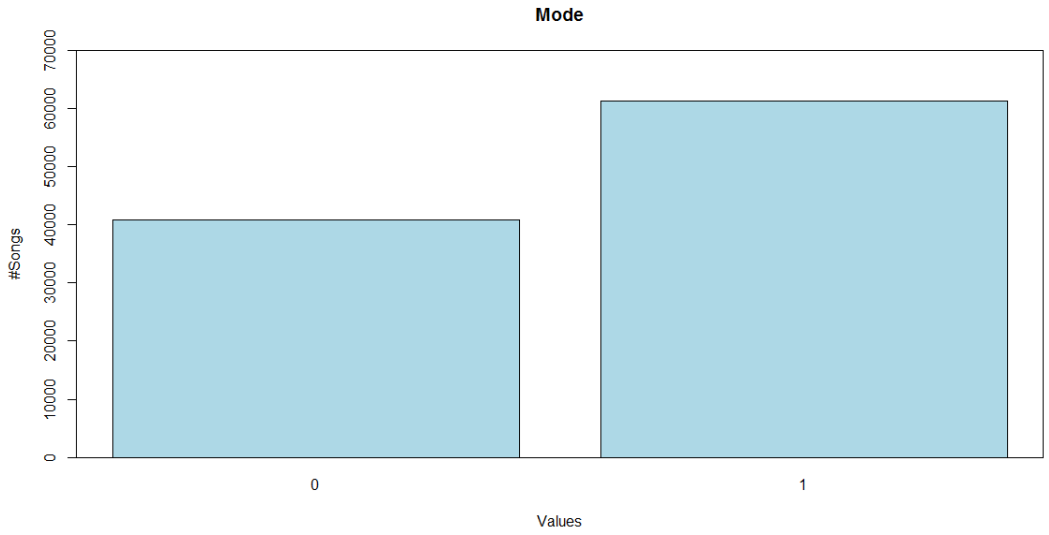

FIGURE 2.14 – Distribution de la variable Mode

3. Résultats et Interprétations de notre Etude

3.1 Résumé statistique

Tout d'abord, visualisons la matrice de corrélation, afin d'avoir un aperçu général de la corrélation entre les variables :

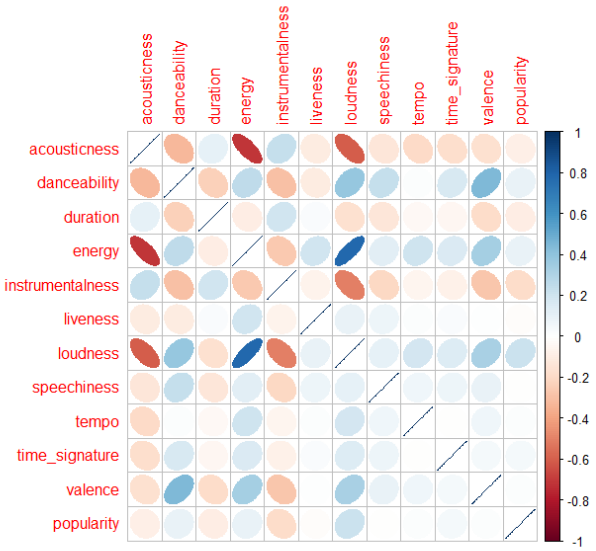

FIGURE 3.1 – Matrice de Corrélation

Nous avons implémenté cette matrice de corrélation à partir de ces commandes :

```
library(corrplot)
matrix_cor=cor(data[, -c(1,2,3,9,12)])      #only quantitative variables
corrplot(matrix_cor, method = "ellipse")
```

Nous constatons une forte corrélation positive entre *energy-loudness* et une légère corrélation positive entre *danceability-loudness* et *danceability-valence*. Nous constatons une forte corrélation négative entre *energy-acousticness*, *acousticness-loudness* et une légère corrélation négative entre *instrumentalness-loudness*.

Afin d'examiner de plus près la corrélation et la signification de certaines variables, nous procéderons à une analyse bivariée exploratoire. Comme nous disposons d'une très grande quantité de données, nous utiliserons des *densités de nuages de points* plutôt que des nuages de points classiques. Dans ce type de graphiques, la couleur rouge indique les régions à forte densité de données et la couleur bleue les régions à faible densité.

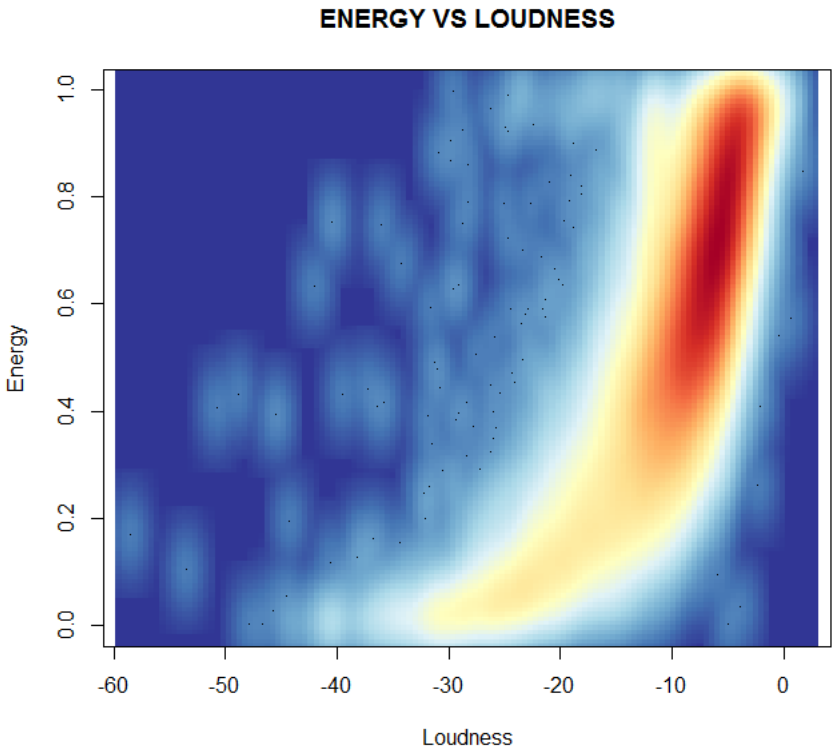

FIGURE 3.2 – Corrélation entre Energy et Loudness

L'énergie est une mesure d'intensité. En général, les morceaux énergétiques sont rapides et bruyants. Par exemple, le death metal a une grande énergie, tandis qu'un prélude de Bach a une faible valeur. Comme nous l'avons remarqué dans le graphe de la matrice de corrélation et maintenant comme nous pouvons le constater sur le nuage de points, ces deux variables sont fortement corrélées positivement. Pour tracer ce graphique, nous avons dû nous appuyer sur ces bibliothèques et ces commandes :

```
library(knitr)
library(ggplot2)
library(colorspace)
library(gridExtra)
library(RColorBrewer)
buylrd = c("#313695", "#4575B4", "#74ADD1", "#ABD9E9", "#E0F3F8", "#FFFFBF",
           "#FEE090", "#FDAE61", "#F46D43", "#D73027", "#A50026")
myColRamp = colorRampPalette(c(buylrd))

smoothScatter(x=data$loudness,y = data$energy, colramp=myColRamp,
              main="ENERGY VS LOUDNESS", xlab="Loudness", ylab="Energy")
```

Dés à présent, considérons deux varaiables ayant une corrélation négative : *acousticness* and *energy*.

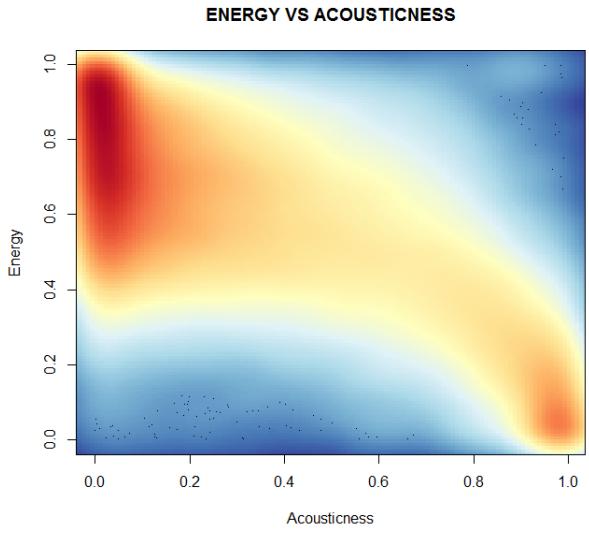

FIGURE 3.3 – Corrélation entre Energy et Acousticness

Nous pouvons clairement voir que l'énergie diminue lorsque l'acoustique augmente et nous repérons deux régions denses opposées : l'une correspondant à des valeurs élevées d'énergie et à des valeurs faibles d'acoustique (coin supérieur gauche) et l'autre correspondant à des valeurs faibles d'énergie et à des valeurs élevées d'acoustique (coin inférieur droit).

Cette analyse exploratoire est également utile pour fournir une interprétation plus précise de la nature de nos variables, même si elles ne sont pas corrélées. Par exemple, on peut étudier le graphique suivant *Instrumentalness* VS *Speechiness*

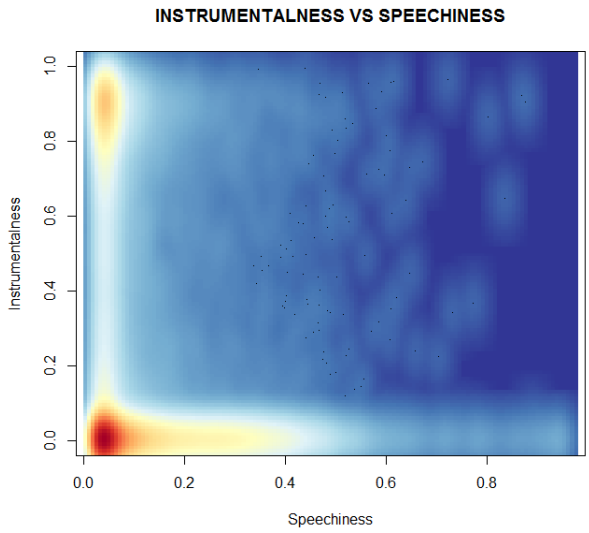

FIGURE 3.4 – Corrélation entre Intrumentalness et Speechiness

Sur le nuage de points, on peut voir un groupe de points avec une forte instrumentalité et une faible élocution (coin supérieur gauche) : on s'attend à trouver des chansons instrumentales qui n'ont pas de paroles et c'est la raison pour laquelle l'élocution est faible. Il y a ensuite un autre groupe de points (coin inférieur gauche) dont l'interprétation est moins intuititive : Pour ceux-là, l'instrumentalité est faible (il y a donc des paroles dans ces morceaux), mais la parole est également peu présente, par conséquent les mots sont probablement chantés et non prononcés, comme dans les chansons Pop typiques. En partant de ces observations et en allant vers la droite (augmentation de l'élocution), la quantité de mots parlés augmente, comme pour des chansons de rap, enfin, avec des valeurs élevées d'élocution, nous faisons l'hypthèse que ces morceaux soient des morceaux de Slam.

Malheureusement, si nous prenons en compte notre variable d'intérêt, *popularité*, nous ne pouvons pas repérer ce genre de comportements. En effet, à partir des nuages de points suivants, nous voyons que ceux-ci sont répartis le long de toutes les valeurs de popularité.

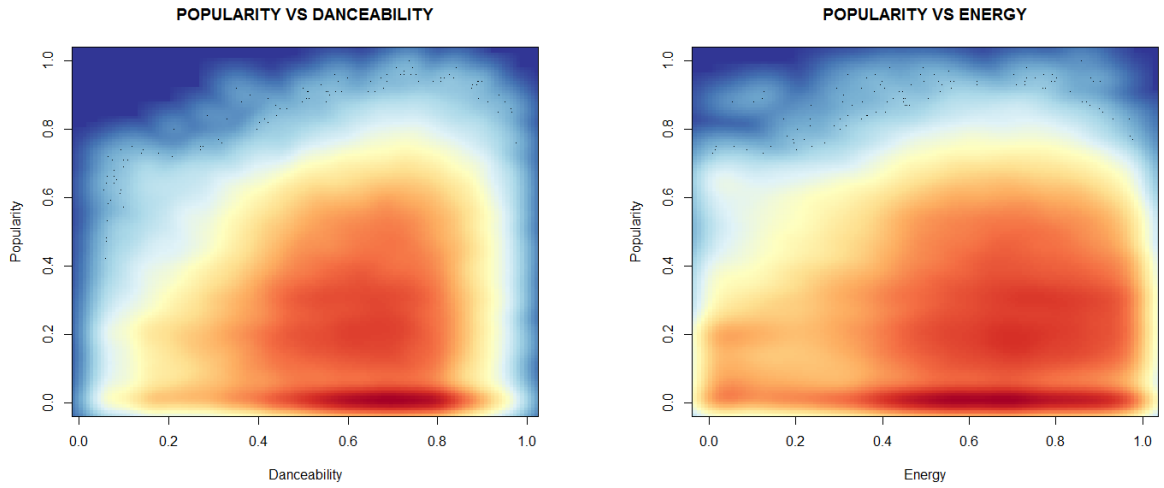

Figure 3.5 – Corrélation de Popularity avec Energy et Danceability

Néanmoins, même si nous pouvions avoir une interprétation plus claire des variables, à partir d'une analyse exploratoire préliminaire[0], nous n'avons toujours pas d'indice clair sur la nature de la variable *popularity*, nous devons donc nous appuyer sur des méthodes statistiques plus spécifiques telles que l'ACP et l'AFC.

3.2 ACP/AFC

Notre base de données étant composée à la fois de variables qualitatives et quantitatives, nous voulions procéder à une AFMD ou Analyse Factorielle des données mixtes (mélange ACP et ACM). Les variables quantitatives et qualitatives ont été normalisé au cours de l'analyse afin d'équilibrer l'influence de chaque ensemble.

On utilise les commandes suivantes :

```
library(FactoMineR)
data$mode = as.factor(data$mode) #qualitative
data$key = as.factor(data$key)#qualitative
str(data) #vérification
data1=data[,4:17]
famd<-FAMD(data1, ncp = 5, sup.var = NULL, ind.sup = NULL, graph = TRUE)
```

Après de nombreuses tentatives nous nous sommes aperçus que l'AFMD prenait beaucoup trop de temps et était infaisable. En effet, cette méthode ne peut être réalisée sur une base de données aussi large que la nôtre. Nous avons donc réfléchi à deux stratégies possibles :

1. *Transformer l'ensemble des données en données quantitatives afin de regrouper les variables quantitatives sous forme de classe.*
   -> **NON** Le nombre de variables quantitatives est nettement supérieur à celui des qualitatives.
2. *Effectue une AFC des variables qualitatives (ou ACM si plus de 2 variables qualitatives) en utilisant les premières composantes principales comme variables quantitatives à la place des variables qualitatives.*
   -> **OUI**, cette approche semble la plus adaptée

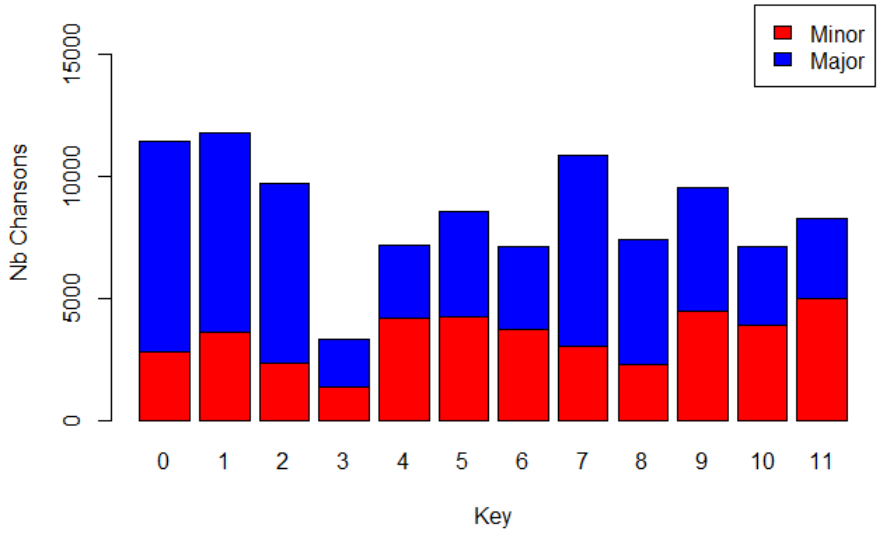

FIGURE 3.6 – Représentation des Variables qualitatives sur l'ensemble du jeu de données

L'AFC a été conduite sur deux variables qualitatives Key et Mode :

```
library(ade4)
res = dudi.coa(qualitative.CA) #AFC
#on selectionne un axe
summary(res)

"Total inertia: 0.07363

Eigenvalues:
  Ax1
0.07363


chisq.test(qualitative.CA)

#Pearson's Chi-squared test
#data:  qualitative.CA
#X-squared = 7521.6, df = 11, p-value < 2.2e-16
```

Le test du Chi-2, nous indique que les variables de ligne et de colonne sont statistiquement significativement associées, ce qui est confirmé par l'AFC, les deux variables sont extrêmement liées.

Afin de procéder à une ACP sur notre jeu de donnée, nous avons créé une nouvelle variable quantitative représentant Key et Mode à l'aide de cette fonction :

```
keymode = data$key #initialisation
for(i in 1:n){
  keymode[i] = res$tab[(data$key[i])+1,(data$mode[i])+1]
}

#Nouvelle base de données avec seulement des variables quantitatives:
new.data = data.frame(data[,-c(9,12)], keymode)
write.table(new.data, file = 'data_quantitative.txt')

data = read.table('data_quantitative.txt')
n = dim(data)[1] # 102156
p = dim(data)[2] #16 et non 17 comme auparavant puisque
#l'on a fusionné deux variables de modalités en une variable quantitative
```

Lors de cette analyse, il est important de choisir le nombre d'axes principaux à prendre en compte pour le reste de notre étude. Ce nombre se détermine par le pourcentage de variances cumulées expliquées par ces axes comme ci-dessous :

```
Cumulative % of variance.
Dim:1       2       3       4       5
    25.633  36.154  44.636  52.565  60.421
```

Il nous a semblé approprié de choisir cinq composantes principales expliquant 60 % de la variance.

Voici le graphique 3.7 obtenu après exécution de l'ACP sur notre nouveau jeu de données :

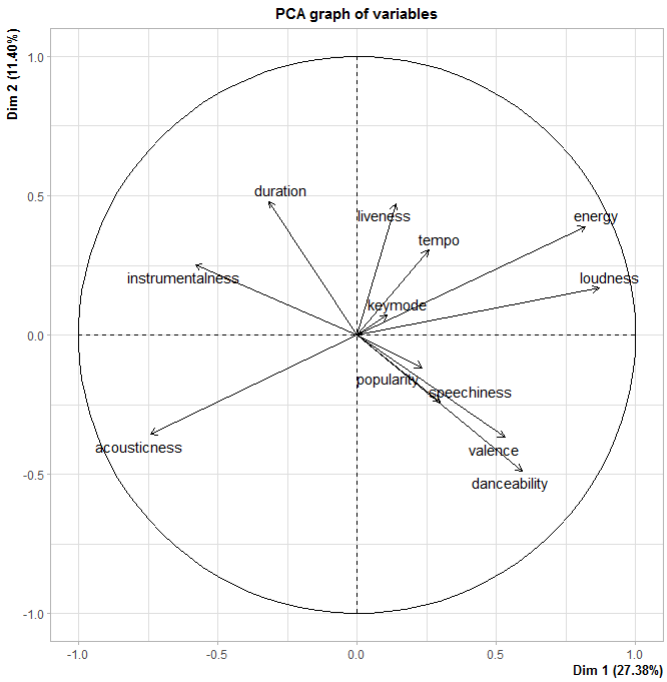

FIGURE 3.7 – ACP axes 1-2

Partie gauche du premier axe : Forte instrumentalness, duration, acousticness et basse energy, loudness et danceability. Cette partie du graphique semble coincider avec le genre *classique*.

Partie droite du premier axe et basse du deuxième axe : Forte danceability, valence et speechiness. Cette partie semble décrire les caractéristiques de la musique *pop* et du *rap*.

Partie droite du premier axe et haute du deuxième axe : Forte energy, loudness, basse danceability. On suppose que cela correspond au *rock*.

Dans le but d'étayer nos suppositions, nous allons maintenant effectuer une classification non supervisée pour obtenir des classes correspondantes aux genres musicaux que l'on semble se distinguer après cette ACP.

3.3 Classification non supervisé et K-means

3.3.1 Analyse Préliminaire

L'ACP nous suggère que la division des données en différentes classes pourrait nous amener à les identifier comme des genres musicaux. Néanmoins, cette suggestion est basée sur notre intuition, nous voulons donc analyser plus précisément les principales caractéristiques des genres musicaux. *Comment obtenir des informations sur les genres musicaux si nous n'avons pas "d'étiquette" qui attribue chaque chanson à un genre spécifique ?*
Notre idée serait la suivante : nous avons cherché les artistes les plus emblématiques de chaque genre musical et nous avons extrait leurs chansons de la base de données pour analyser les caractéristiques d'un genre spécifique et les comparer à celles des autres genres musicaux et à celles des morceaux de toute la base de données. Cette analyse est aussi utile, car nous espérons qu'elle nous aidera à reconnaître certains genres musicaux dans les groupes que nous allons créer.
Afin de faciliter notre analyse, dans cet environnement de travail, il a été nécessaire d'*échelonner* les variables pour se concentrer sur les différences entre classes de variables et non les différences entre ces variables.

Le Classique : Nous avons recherché les artistes incarnants le mieux ce type de musique pour extraire leurs chansons de la base de données. Voici les compositeurs que nous avons choisi : *Johann Sebastian Bach,Wolfgang Amadeus Mozart,Ludwig van Beethoven, Claude Debussy et Fryderyk Chopin.*
Dans le but d'appliquer la théorie à la pratique, une variable binaire permettant d'identifier le morceau d'un artiste a été créé :

```
classic = ifelse(data$artist_name %in% c("Johann Sebastian Bach",
        "Wolfgang Amadeus Mozart","Ludwig van Beethoven", "Claude Debussy",
        "Fryderyk Chopin") ,1,0)

data_classic = subset(data, classic==1)
```

A l'aide des commandes suivantes, nous avons généré des *boîtes à moustaches stratifiées* pour chaque variable ce qui nous permet de comparer l'ensemble des données avec la classe **Musique Classique** que nous avons créé.

```
#Stratified boxplots:

    par(mfrow=c(3,4))
    for(i in 1:4){
        for(j in 1:3){
            boxplot(data.quantitative[,(i-1)*3+j] ~ classic,
            col = c('green','red'), ylab = names_variables[(i-1)*3+j])
         }
    }
```

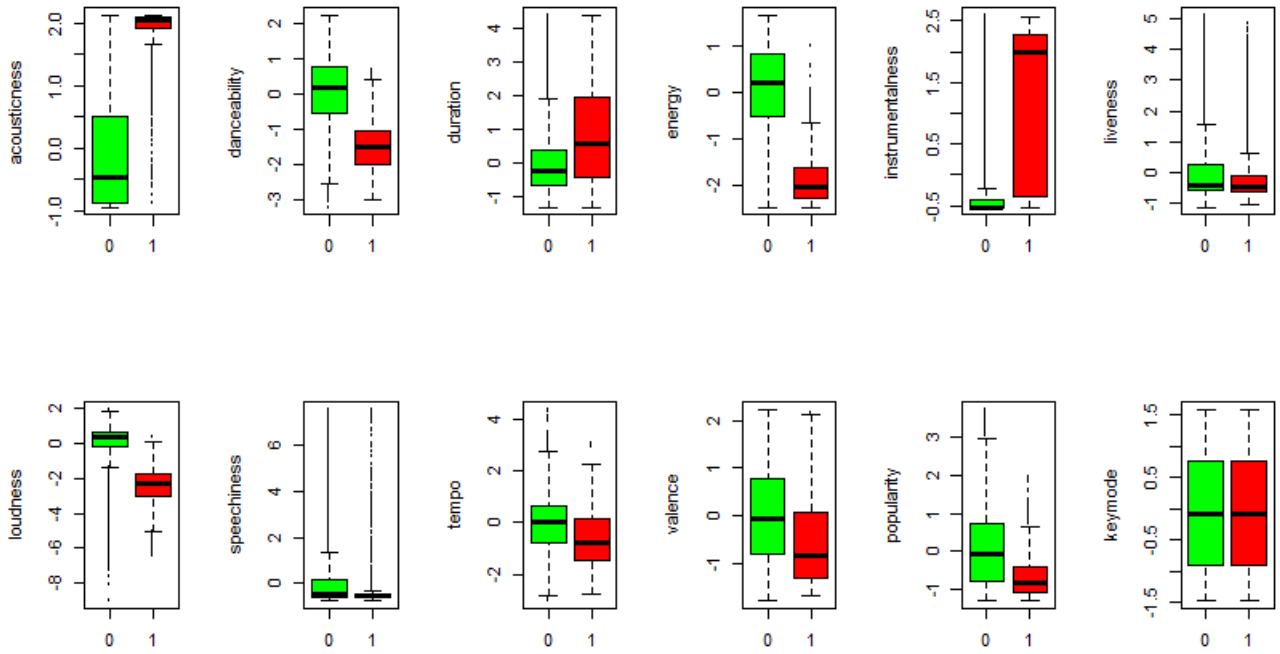

FIGURE 3.8 – Boîtes à moustaches stratifiées pour chaque variable : les boxplots rouges (valeur = 1) représentent les caractéristiques du groupe classique extrait, tandis que les boxplots verts (valeur = 0) représentent les caractéristiques de toutes les autres chansons du jeu de données

Notre hypothèse est confirmée : la musique classique a une très grande acoustique, une grande instrumentalité et une durée plus longue par rapport aux autres genres.3.8

Cette analyse a été réitéré pour les autres genres musicaux. Nous nous contentons d'indiquer les artistes que nous avons pris en compte pour chaque genre musical et représentons un graphique des caractéristiques de chaque groupe.

Le Rap : Liste des artistes considérés pour le Rap : *Eminem, Tupac Shakur, Kanye West, Jay-Z, Drake, Lil Wayne et 50 Cent.*

La Pop : Liste des artistes considérés pour le Rap : *Lady Gaga, Rihanna, Justin Bieber, Michael Jackson, Madonna, Katy Perry, Beyoncé, Ariana Grande, Coldplay, Bruno Mars et Maroon 5.*

La House : Liste des artistes considérés pour le Rap : *Avicii, Martin Garrix, Nicky Romero, David Guetta, deadmau5, Afrojack, Kaskade, Daft Punk et Armin van Buuren.*

Le Rock : Liste des artistes considérés pour le Rap : *Pink Floyd, Guns N' Roses, Nirvana, Led Zeppelin, Bruce Springsteen, The Rolling Stones, Queen, The Beatles, Green Day et Metallica.*

Le Reggaeton : Liste des artistes considérés pour le Rap : *Daddy Yankee, Nikcy Jam, J.Balvin, Don Omar, Bad Bunny, Maluma, Yandel, Wisin, Luis Fonsi et Enrique Iglesias.*

Ensuite, nous avons tracé des diagrammes en bâtons dans lesquels, pour chaque variable, y fi-

gurent les valeurs moyennes de chaque groupe :

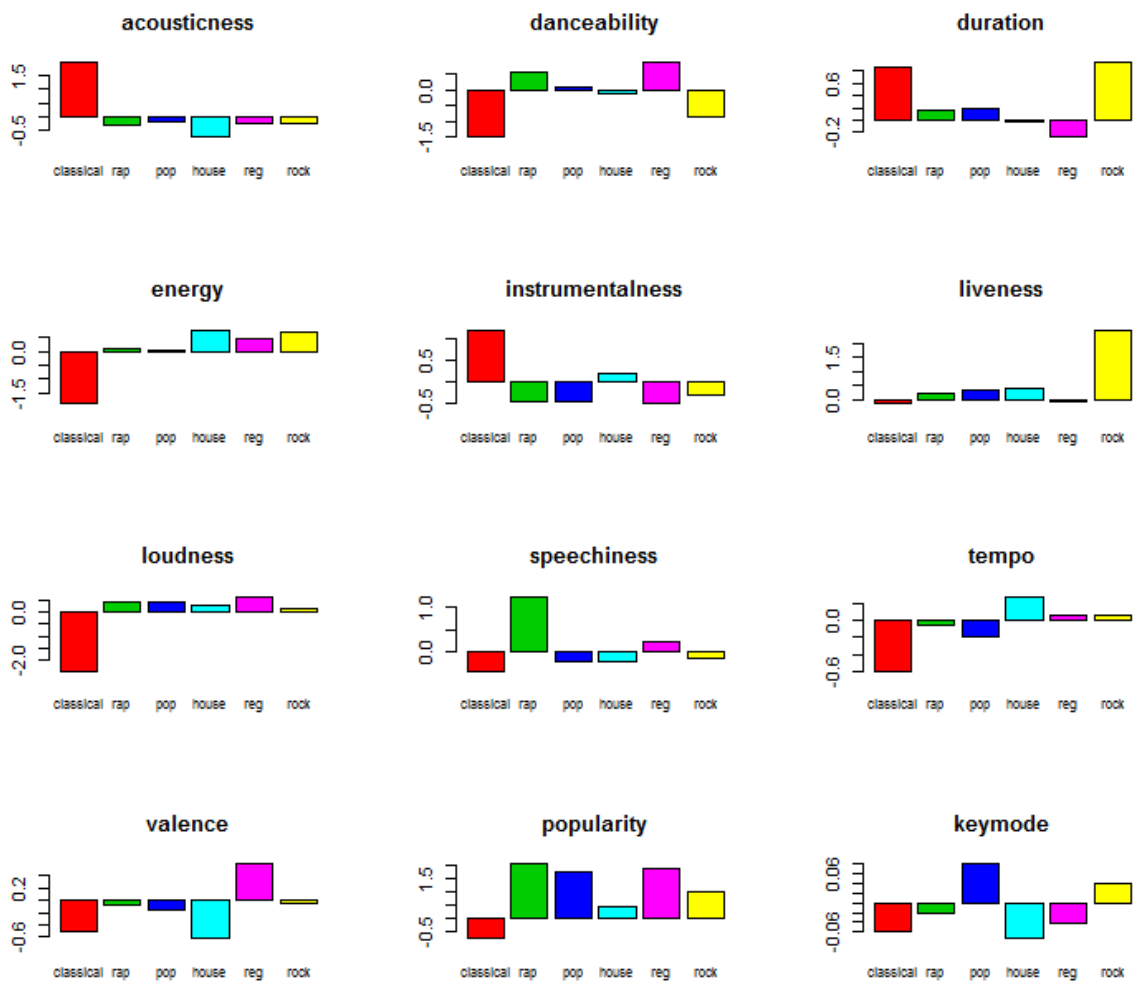

FIGURE 3.9 – Diagrammes en bâtons des différentes variables pour chaque cluster

L'interprétation de *la musique Classique* a été abordé ci-dessus 3.8.
Dans le cas du *Rap*, ce genre se révèle être particulièrement dansant, possède une grande popularité et en particulier une grande élocution (présence de mots parlés, comme attendu dans des chansons de rap). Cependant,ce genre se caractérise également par une très faible acoustique et instrumentalité.
Contrairement, à la *Pop*, qui possède une grande popularité et mode/clé, un caractère dansant lègerement positif mais une valence et instrumentalité faible (conséquence d'un manque d'utilisation de réels instruments dans ce type de chansons).
En interprétant les résultats obtenus pour la *House*, on constate qu'elle a l'énergie la plus élevée parmi tous les genres, une très faible acoustique (en raison de la prédominance des instruments électroniques) ainsi qu'un faible mode/clé et valence.
En ce qui concerne le *Reggaeton*, celui-ci est extrêmement dansant, a une forte valence (chanson typiquement joyeuse et dansante) mais possède les chansons les plus courtes et les moins populaires.
Enfin, pour le *Rock*, les morceaux semblent être les plus long, en effet, de nombreux morceaux sont enregistrés en direct donc l'énergie est élevée et le caractère dansant est vraiment faible.

Il serait également important de comparer les différents groupes musicaux en ce qui concerne les composantes principales que nous avons calculées précédemment. Nous avons donc réécrit les valeurs moyennes des différents groupes dans ces nouvelles coordonnées, en projetant les valeurs précédentes le long des composantes principales (les 5 premières), avec les commandes suivantes :

```
pca = PCA(data.quantitative, graph = F)
moyenne_pca = moyenne %*% pca$svd$V[,1:5]
```

Puis nous avons créé des *diagrammes en bâtons* des cinq premières composantes pour les différents genres musicaux :

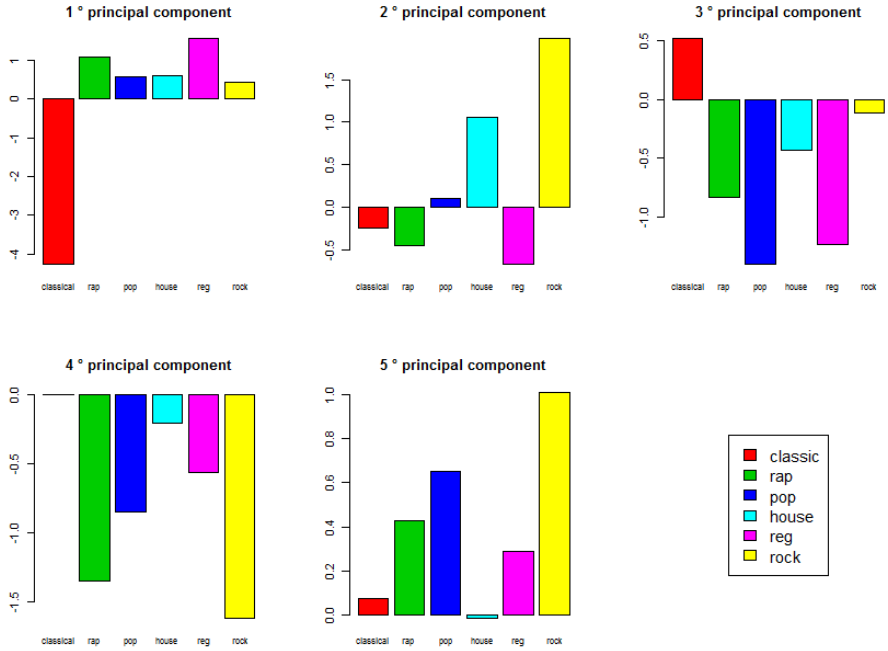

FIGURE 3.10 – Diagramme en bâtons des 5 premières composantes principales pour tous les genres musicaux

Classique : 1°composante considérablement négative, 2° légerement négative et 3° plutôt positive.
Rap : 1° positive et 2°/3° négatives.
Pop : 1° positive, 2° à peine positive et 3° extrèmement négative.
House : 1°/2°positives et 3°négative.
Reggaeton : 1° très positive et 2°/3° très négatives.
Rock : 2° fortement positive et 1°/3° négatives.

Tous les groupes considérés ont une quatrième composante négative et une cinquième composante positive (à l'exception de la *House* qui a une cinquième composante à peine négative)ce qui pourrait indiquer que ces deux composantes décrivent des propriétés attribuables à la musique en elle-même. Enfin, nous sommes prêts à effectuer une *classification non supervisée* en appliquant l'algorithme des *K-means*.2. Nous avons pensé qu'il aurait été plus approprié d'utiliser cette méthode plutôt que la technique CAH (Classification ascendante hiérarchique) non adaptée à un grand jeu de données.

3.3.2 Choix du nombre de classes K

En utilisant l'algorithme des *K-means*, nous devons indiquer a priori le nombre de clusters que nous cherchons. Cela suit le critère fondé sur les inerties : on recherche un "coude" dans la décroissance de l'inertie intra-classe (on s'arrête quand l'inertie intra-classe ne décroit quasiment plus). On applique ce critère pour 7 classes comme dans le code ci-dessous :

```
inertie.intra <- rep(0,times=7)
for (k in 1:7){
  kmeans.result <- kmeans(scale(data) ,centers=k,nstart=100)
  inertie.intra[k] <- kmeans.result$tot.withinss/kmeans.result$totss
```

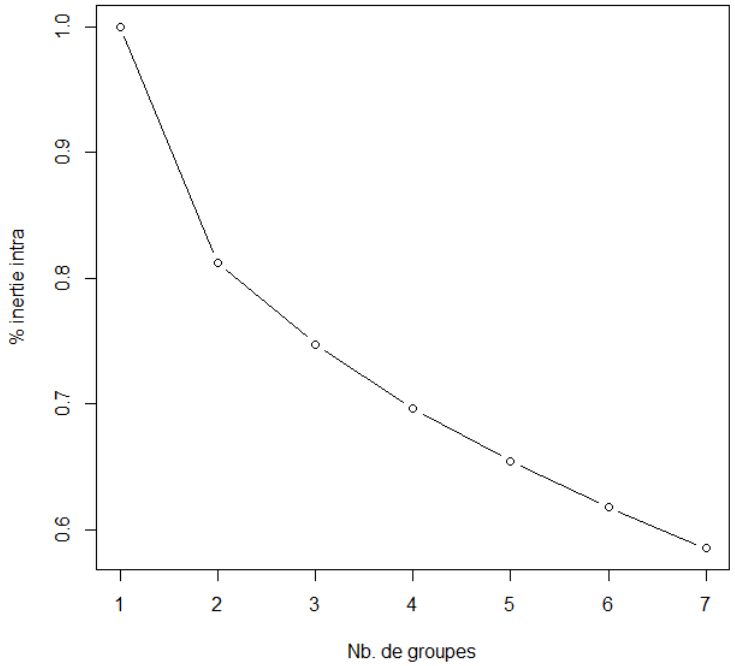

FIGURE 3.11 – Pourcentage d'inertie intra-classe en fonction du Nombre de groupes obtenus par classification non supervisée

Maintenant que nous avons choisi le nombre de classes, il faut se pencher sur le choix des centres. Puisque les K-means.2 fournissent un optimum local et sont sensibles à l'initialisation, notre stratégie consistait à fournir la position des centres initiaux des clusters au lieu de les choisir aléatoirement. Ainsi, nous avons choisi certaines combinaisons possibles des valeurs moyennes des genres musicaux que nous avons considéré dans notre analyse préliminaire pour nos 4 centres initiaux. Malheureusement, notre stratégie a échoué car nous avons reçu un message d'erreur :"*les centres d'entrée ne sont pas assez distincts*". Le problème était possiblement causé par un nombre conséquent de variables (donc de composantes des centres) qui ont des valeurs proches qui se traduit par des centres très proches, bien qu'il y ait des différences évidentes entre les caractéristiques musicales des différents genres musicaux. Notre seule alternative était de travailler avec des centres initialisés aléatoirement.

3.3.3 Algorithme des K-means

Après tous ce travail de préparation, il est temps d'implémenter cet algorithme en utilisant la distance euclidienne entre les observations centrées réduites comme on peut le voir sur ces commandes :

```
result.k = kmeans(scale(data.quantitative), centers = 4, nstart = 100)
```

Après avoir tracé un graphique en 3D, en prenant comme axes les 3 variables les plus pertinentes, *Instrumentalness, Energy et Speechiness*, on obtient la figure suivante :

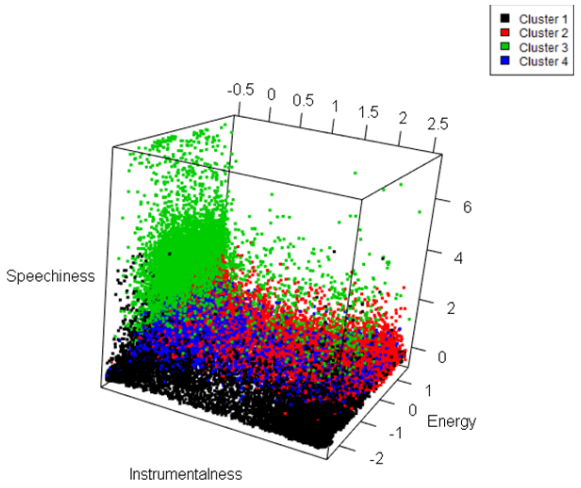

FIGURE 3.12 – Clusters en 3D sur critères Energy, Speechiness et Instrumentalness

On constate que le groupe en *vert* a une forte instrumentalité et le groupe en *noir* a une faible énergie.3.12

Afin d'obtenir une vue d'ensemble, on le fait avec toutes les variables :

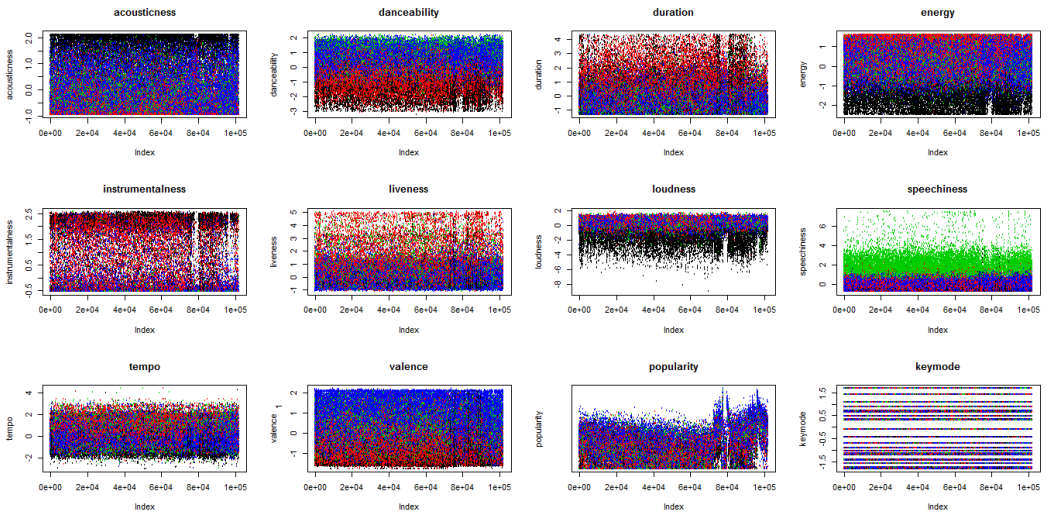

FIGURE 3.13 – Clusters avec l'ensembles des variables

De plus,nous pouvons répéter le procédé cette fois-ci en utilisant les résultats de l'ACP. Le regroupement dans l'espace 3D des trois premières composantes principales est le suivant :

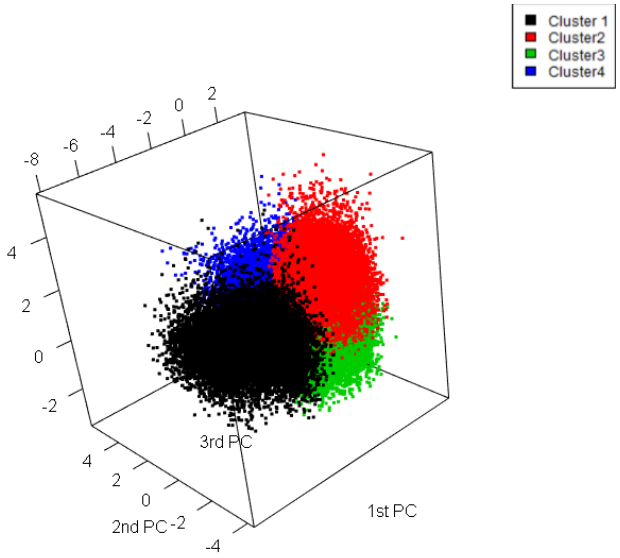

Figure 3.14 – Clusters avec ACP

Nous pouvons très clairement distinguer les quatre classes différentes. Par conséquent, même si les composantes principales ne sont pas aussi faciles à interpréter que les variables initiales, elles fournissent une classification plus précise et claire, de sorte qu'elles s'avèrent également très utiles pour identifier la signification de chacun des groupes.

Comme réalisé auparavant avec l'ensemble des variables, on s'intéresse désormais aux 5 premières composantes principales individuellement :

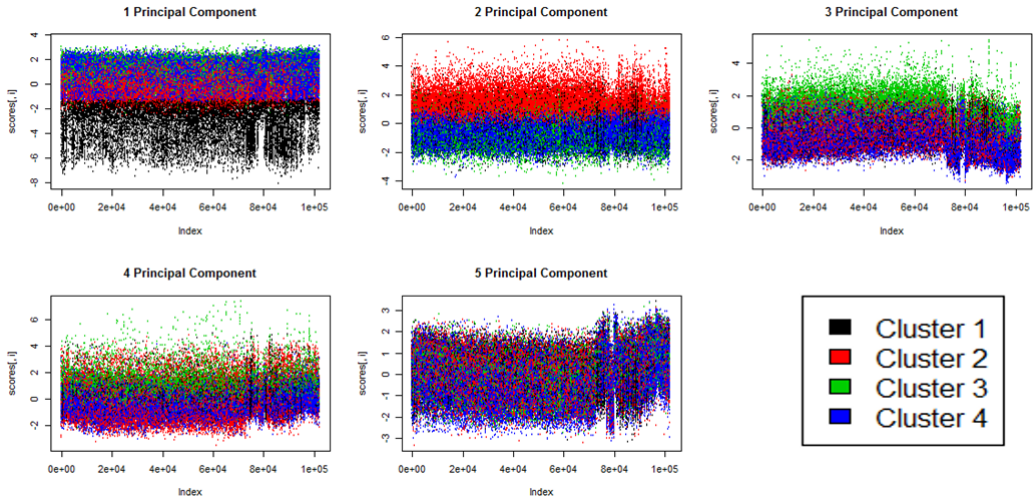

Figure 3.15 – Cluster des 5 premières composantes principales

3.3.4 Analyse et caractérisation des clusters obtenus

Examinons maintenant les composantes du centre (c'est-à-dire les valeurs des variables originales pour chaque centre) de chacune des 4 clusters afin de mettre en évidence leurs différences :

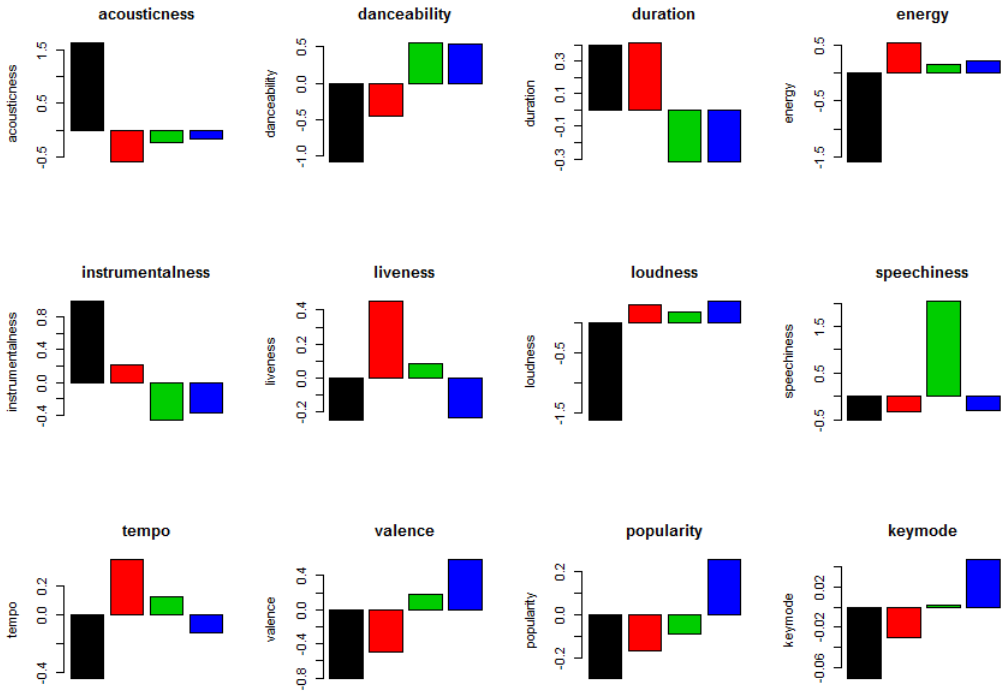

FIGURE 3.16 – Diagramme en bâtons de l'ensemble des clusters en fonction des variables explicatives

Nous avons réitéré la méthode mais cette fois-ci en utilisant les coordonnées des composantes principales comme critères d'analyse de chaque cluster.

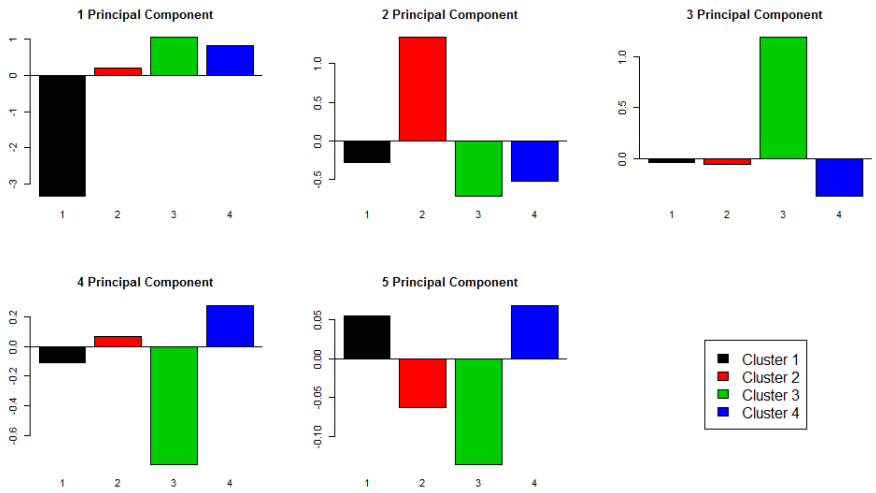

FIGURE 3.17 – Diagrammes en bâtons de l'ensemble des clusters en fonction des composantes principales

Premier cluster : Le CLASSIQUE Le premier groupe est caractérisé par une première composante principale très faible, une deuxième composante faible, une troisième composante nulle. Des valeurs élevées d'acoustique, d'instrumentalité et des valeurs basses du caractère dansant, d'énergie, de volume sonore.
En se basant sur nos premières observations, nous pensons que le premier cluster contient le genre musical : **CLASSIQUE**.
Par les commandes suivantes :

```
result.k$size[1] #16623 tracks
cluster_classic = data[result.k$cluster==1,]
summary(cluster_classical$artist_name)
```

Ce groupe comprend 16 623 morceaux,dont les artistes possédants le plus grand nombre de chansons sont :

```
Johann Sebastian Bach: 1887 chansons
Wolfgang Amadeus Mozart: 992 chansons
Ludwig van Beethoven: 459 chansons
Claude Debussy: 320 chansons
```

Qui se trouvent être parmis les compositeurs classique les plus connus. A partir des résultats de l'analyse préliminaire, nous avons pu comparé les valeurs moyennes des variables pour le groupe test que nous avons sélectionné dans le but de représenter la musique classique avec les coordonnées du centre de ce groupe, représentant probablement les valeurs moyennes de la classe :

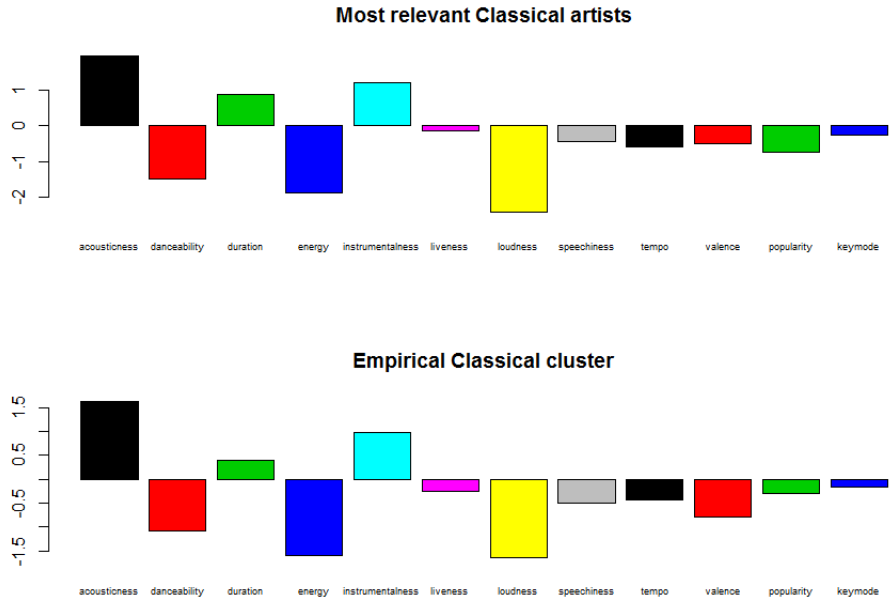

FIGURE 3.18 – Comparaisons des valeurs obtenues pour chaque variables dans le cluster considérant uniquement les morceaux d'artistes emblématiques du genre Classique et le cluster du genre entier

Nous constatons une très grande concordance entre les caractéristiques du cluster classique empirique et celles que nous attendions pour la musique classique à partir de notre analyse préliminaire.

Deuxième Cluster : HOUSE-ROCK La deuxième classe est caractérisée par une deuxième composante très élevée, une première à peine positive, et une troisième composante nulle. Cette classe se distingue par une grande vivacité, énergie, durée des morceaux ainsi qu'un faible caractère dansant et une faible acoustique. Elle comprend 28 064 chansons et les artistes les plus représentés (par leur nombres de chansons) sont les suivants :

```
Armin van Burren: 148 chansons          -> DJ electronique/house
Above & Beyond: 98 chansons             -> electronique/house
Image sounds: 91 chansons               -> electronique/house
R.E.M.: 72 chansons                     -> rock
The Rolling stones: 27 chansons         -> rock
```

Par conséquent notre deuxième Cluster semble correspondre aux genres : **HOUSE ET ROCK**. En comparant, comme précédemment, les valeurs attendues pour la musique house et le rock (nous avons pris la moyenne des deux genres musicaux) avec les valeurs du centre de ce cluster, nous avons obtenu :

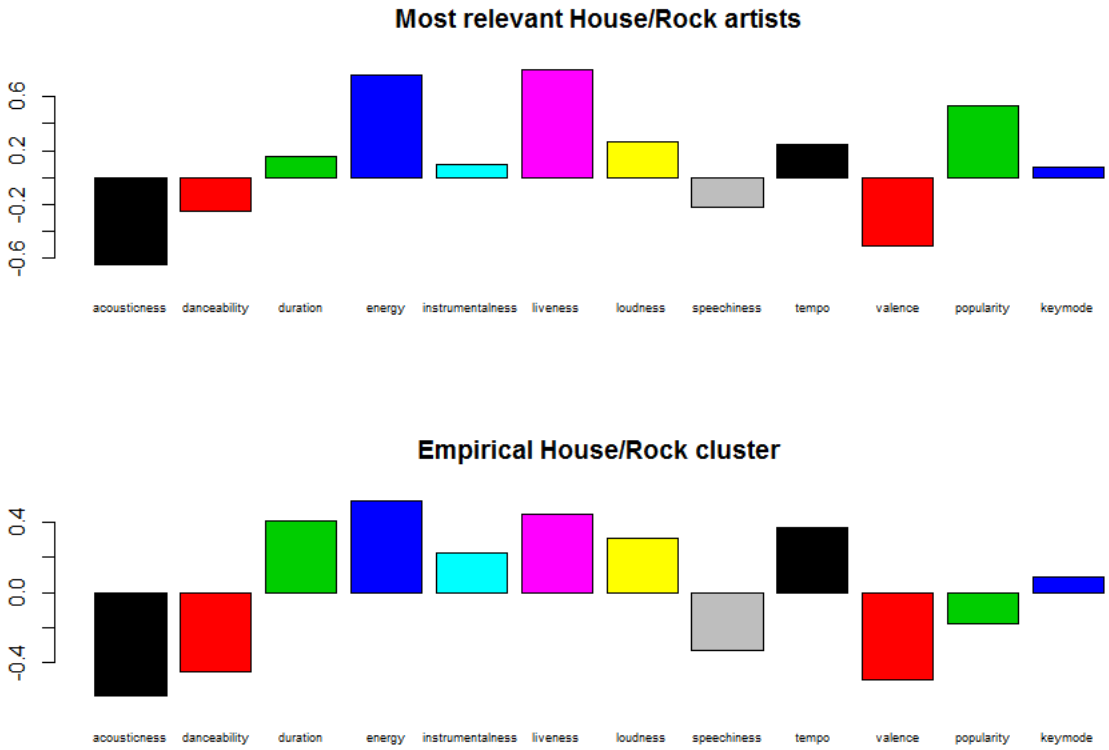

FIGURE 3.19 – Comparaisons des valeurs obtenues pour chaque variables dans le cluster considérant uniquement les morceaux d'artistes emblématiques des genres House/Rock et le cluster des genres entier

Nous retrouvons une très bonne correspondance entre les valeurs attendues pour la house et le rock, avec celles de notre cluster. La seule grande différence réside dans la valeur de la *popularité* qui devrait être beaucoup plus élevée que la valeur empirique. Mais ceci était prévisible puique nous avons choisi *les artistes les plus emblématiques* pour construire nos classes "théoriques",ce sont donc des artistes très célèbres qui auront une popularité bien supérieure aux autres artistes de ces deux genres musicaux.

Troisième Cluster : RAP Cette classe se détermine par une première composante principale élevée, une deuxième composante principale faible et une troisième composante principale très élevée. En outre, la capacité à danser et l'élocution sont très élevées, tandis que l'instrumentalité est, quant à elle, relativement faible. Le groupe contient 15 291 chansons et les principaux artistes sont :

```
Waka Flocka Flame: 110 chansons        -> rap
DJ Fuqua: 81 chansons                  -> rap
Chief Keef: 49 chansons                -> rap
Lud Foe: 49 chansons                   -> rap
```

On peut supposer que ce cluster représente le :RAP. En comparant les mêmes diagrammes en bâtons que ceux réalisés pour les genres précédents :

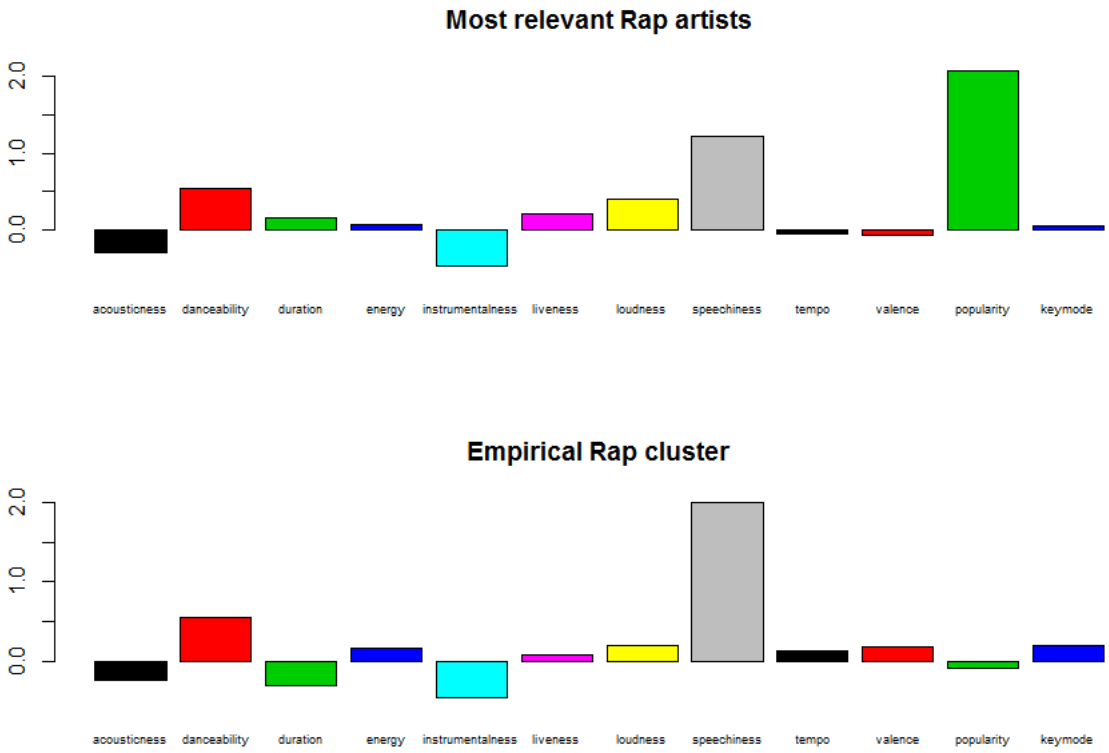

FIGURE 3.20 – Comparaisons des valeurs obtenues pour chaque variables dans le cluster considérant uniquement les morceaux d'artistes emblématiques du genre Rap et le cluster du genre entier

Ce groupe semble être représentatif de la musique rap, à l'exception de la popularité, por laquelle nous observons le même phénomène que pour les autres genres étudiés.

Quatrième Cluster : POP-REGGAETON La dernière classe a une première composante positive, une deuxième négative et une troisième très faible. Elle possède des valeurs très élevées de valence, de popularité et un fort caractère dansant. Ce groupe contient 42 178 titres (il est beaucoup plus grand que les autres) et les artistes les plus représentés sont :

```
Los Cadetes de Linares: 192 chansons   -> sud-américain
Duo Libano: 136 chansons               -> sud-américain
Sia: 31 chansons                       -> pop
Ariana Grande: 23 chansons             -> pop
```

Il semble justifié d'attribué ce cluster aux genres : **POP-REGGAETON**.

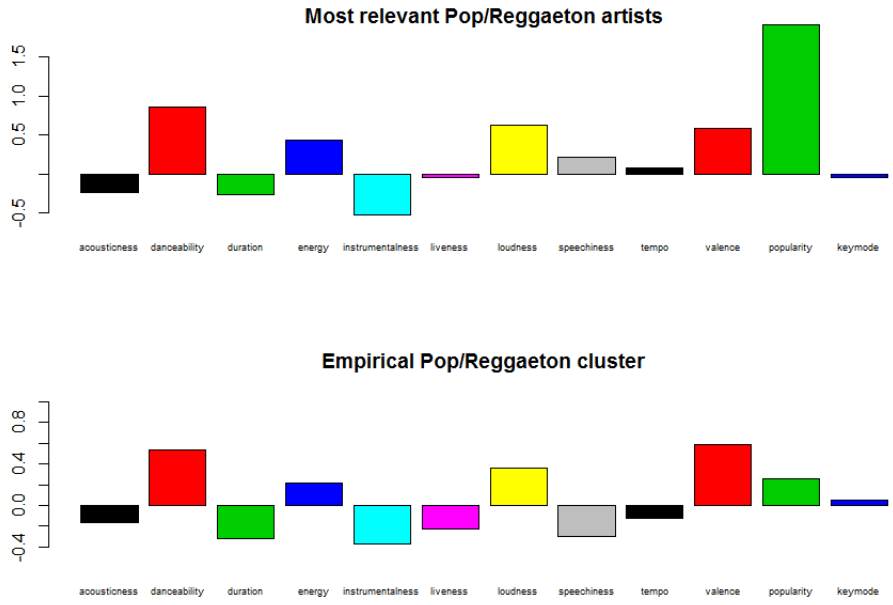

FIGURE 3.21 – Comparaisons des valeurs obtenues pour chaque variables dans le cluster considérant uniquement les morceaux d'artistes emblématiques des genres Pop/Reggaeton et le cluster des genres entier

Dans ce cas également, il y a une bonne similitude entre le groupe empirique et la classe "théorique", mais nous pouvons déceler des différences pour certaines variables. Hormis la différence de popularité que nous avons déjà constatée dans les cas précédents, nous obtenons également des valeurs différentes en ce qui concerne les variables : Liveness, Speechiness et Tempo. Une des raisons principales à cela est l'élimination d'une partie des genres musicaux en se focalisant sur les artistes emblématiques de chaque genre, ce n'est donc pas surprenant d'avoir dû mélanger les genres musicaux pour la construction des 4 clusters. La dernière classe étant beaucoup plus grande que les autres (42 178 chansons), nous pensons qu'elle pourrait contenir plus de genres musicaux que les seules chansons pop et reggaeton (par exemple, la musique latino-américaine se distingue du reggaeton par plus de paroles et plus d'instruments). Or, le but de notre analyse n'est pas de labelliser chaque genre musical parfaitement, nous considérons plutôt qu'il est approprié d'attribuer ce cluster aux genre pop-reggaeton au vu de nos résultats et d'une bonne correspondance pour la plupart des variables.

Ce *camembert* illustre la différence de taille entre les classes obtenues :

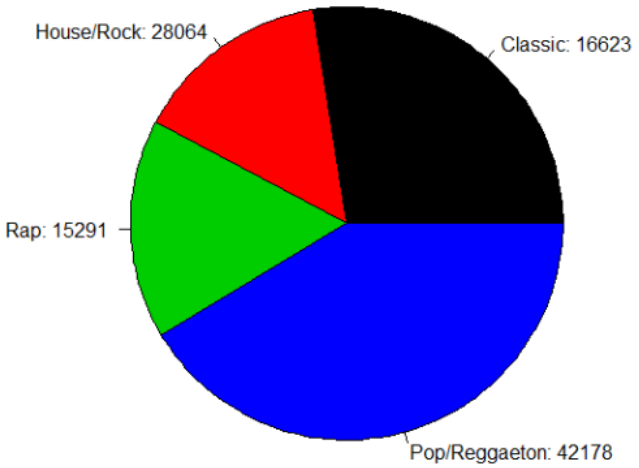

FIGURE 3.22 – Taille des différents Clusters

Pour conclure cette section, voici trois graphiques qui représentent un résumé des résultats que nous avons obtenus jusqu'à présent, en combinant l'ACP et la classification non supervisée. Ceux-ci prennent en considération les 3 premières composantes principales, en traçant deux composantes à la fois qui montrent la contribution de celles-ci pour chacune des variable initiale. Deux clusters voient leur projection se chevaucher dans le plan déterminé par les deux composantes principales considérées. De cette façon, nous pouvons voir graphiquement les caractéristiques de chaque genre musical, non seulement par rapport aux variables initiales mais aussi par rapport composantes principales sur le même graphique.

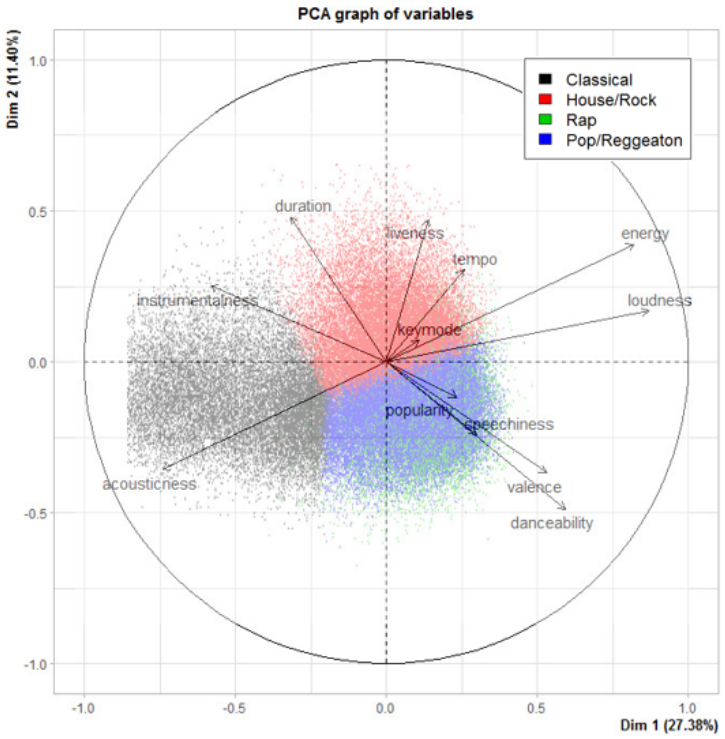

FIGURE 3.23 – Plan des première et deuxième composantes principales

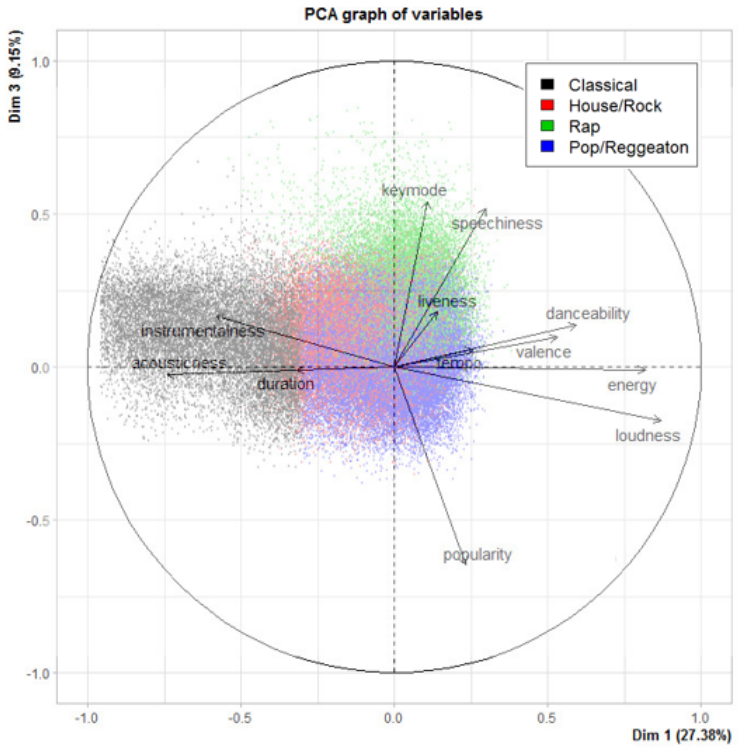

FIGURE 3.24 – Plan des première et troisième composantes principales

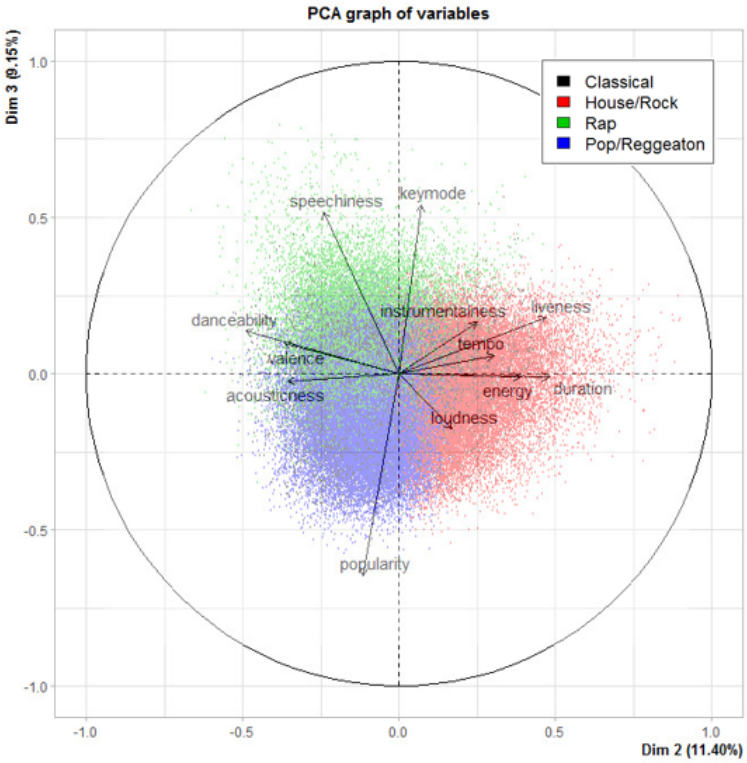

FIGURE 3.25 – Plan des deuxième et troisième composantes principales

3.4 Régression linéaire

3.4.1 Popularité d'un artiste et d'une chanson

Finalement, nous disposons de tous les éléments nécessaires pour tenter de construire un modèle linéaire représentatif. Toutefois, au cours de notre analyse, nous avons formulé les considérations suivantes : Nous craignons que les seules caractéristiques musicales ne suffisent pas à expliquer la popularité. Nous supposons en particulier qu'il y a une popularité liée aux **artistes** eux-mêmes qui influence clairement la popularité de la chanson. L'indice de popularité dans cette base de données n'est qu'une mesure du nombre de fois qu'une chanson a été écouté. Mais si l'on pense au cas où deux artistes (un très célèbre et l'autre beaucoup moins) produisent la *même* chanson : il est naturel d'attendre un indice de popularité plus élevé pour la chanson dont l'artiste est célèbre que l'indice de popularité des chansons dont l'artiste ne l'est pas, même si *toutes les caractéristiques sont identiques*, c'est-à-dire que toutes les variables ont les mêmes valeurs !
Puisque nous souhaitons construire un modèle capable d'effectuer une *prédiction* précise, c'est-à-dire capable de prédire la popularité d'une nouvelle chanson donnée, nous devons tenir compte de la popularité de l'artiste. Construisons donc une variable supplémentaire qui est censée représenter la ***popularité de l'artiste***. Notre stratégie consiste à la créer en prenant la moyenne des indices de popularité parmi toutes les chansons d'un artiste. Les commandes suivantes permettent de calculer l'indice de popularité de l'artiste, de créer une nouvelle variable stockant la popularité de l'artiste pour chaque titre et d'ajouter cette nouvelle variable à la base de données :

```
artist_averages=aggregate(data$popularity,by=list(artist=data$artist_name),FUN=mean)

artist_popularity = rep(0, dim(data)[1])
data = data.frame(data, artist_popularity)
for(i in 1:length(artist_averages$artist)){
  data[which(data$artist_name==artist_averages$artist[i]),18]= artist_averages$x[i]
      #18: column in data, corresponding to the new variable: "artist_popularity"
}
```

3.4.2 Le rôle des genres musicaux

Nous aimerions construire un modèle qui tienne compte du genre de la chanson que nous avons estimé précédemment. Pour ce faire, nous avons construit des variables binaires qui nous ont permis de séparer la contribution de chacune de nos variables dans les 4 genres musicaux différents considérés.

Ensuite, nous construisons 3 variables binaires *factices* d_i , avec i = 1,2,3 afin d'identifier l'appartenance à un groupe

```
d1 d2 d3
1  0  0  : house-rock
0  1  0  : rap
0  0  1  : pop-reggaeton
0  0  0  : classic
```

Ces trois variables ont été créé par les commandes suivantes :

```
dummy_houserock = ifelse(data$cluster == "House_Rock",1,0)
dummy_rap = ifelse(data$cluster == "Rap",1,0)
dummy_popregg = ifelse(data$cluster == "Pop_Reggaeton",1,0)
```

Ici, la variable qualitative *cluster*, que nous avons ajoutée au jeu de données, indique le genre musical auquel une chanson a été assignée pendant la phase de classification.
Ces variables factices sont *quantitatives* et correspondent aux labels des clusters *qualitatifs*. A ce stade de notre étude, notre jeu de données ressemble à ceci :

```
       artist_name                track_id              track_name
1               YG  2RM4jf1Xa9zPgMGRDiht8O               Big Bank
2               YG  1tHDG53xJNGsItRA3vfVgs              BAND DRUM
3          G Herbo  13Mf2ZBpfNkgWJowvM5hXh           Bon appÃ©tit
4  Mr Little Jeans  3Z78Hd9B1OndIo7XJajwYR              Forgetter
5     Orjan Nilsen  16UKw34UY9w40Vc7TOkPpA  Nothing Here But Love
6       Mbo Mentho  3PmPdmHHlDylHFW8yF5f8H               Releduas

 acousticness danceability duration energy instrumentalness liveness loudness
1      0.00582        0.743  238.373  0.339          0.000000   0.0812   -7.678
2      0.02440        0.846  214.800  0.557          0.000000   0.2860   -7.259
3      0.11500        0.885  181.838  0.348          0.000000   0.1070  -12.569
4      0.12500        0.821  254.122  0.512          0.007030   0.0879   -7.138
5      0.04480        0.574  189.467  0.881          0.000000   0.0756   -2.150
6      0.40900        0.574  171.000  0.463          0.000264   0.1180  -23.058

 speechiness   tempo time_signature valence popularity       keymode
1      0.4090 203.927              4   0.118       0.15     0.1576083
2      0.4570 159.009              4   0.371       0.00     0.1525933
3      0.4510 142.111              4   0.180       0.00    -0.1108029
4      0.0596 128.035              4   0.543       0.28     0.2437706
5      0.1660 126.131              4   0.551       0.29     0.2437706
6      0.0499 159.988              4   0.993       0.00     0.5175726

         cluster artist_popularity houserock    rap     popregg
1            Rap         0.4994444         0      1           0
2            Rap         0.4994444         0      1           0
3            Rap         0.2731579         0      1           0
4  Pop_Reggaeton         0.2800000         0      0           1
5  Pop_Reggaeton         0.2235714         0      0           1
6  Pop_Reggaeton         0.0000000         0      0           1
```

3.4.3 Test sur le jeu de données

La dernière astuce avant de construire réellement le modèle consiste à extraire un ensemble de test du jeu de données (nous avons choisi au hasard 300 chansons), que nous allons utiliser pour tester la prédiction du modèle que nous allons construire. Nous avons utilisé les commandes :

```
test = sample(1:dim(data)[1],300)
data_test = data[test,]

data_model = data_quantitative[-test,]
```

3.4.4 Construction du modéle et Sélection

Il est temps de mettre en place un modèle linéaire, prenant en compte toutes les variables quantitatives et les variables factices ainsi que leur interaction avec les variables initiales afin de trouver un modèle qui prend en compte l'appartenance aux genres musicaux. Notre modèle est le suivant :

```
model0 = lm(popularity ~ houserock + rap + popreggaeton +
    acousticness + acousticness:houserock
            + acousticness:rap + acousticness:popreggaeton +
    danceability + danceability:houserock
            + danceability:rap + danceability:popreggaeton +
    duration + duration:houserock  + duration:rap + duration:popreggaeton +
    energy + energy:houserock + energy:rap + energy:popreggaeton +
    instrumentalness + instrumentalness:houserock
            + instrumentalness:rap + instrumentalness:popreggaeton +
    liveness + liveness:houserock + liveness:rap + liveness:popreggaeton +
    loudness + loudness:houserock + loudness:rap + loudness:popreggaeton +
    speechiness + speechiness:houserock
            + speechiness:rap + speechiness:popreggaeton +
    tempo + tempo:houserock + tempo:rap + tempo:popreggaeton +
    time_signature + time_signature:houserock
            + time_signature:rap + time_signature:popreggaeton +
    valence + valence:houserock + valence:rap + valence:popreggaeton +
    keymode + keymode:houserock + keymode:rap + keymode:popreggaeton +
    artist_popularity)
```

Nous disposons désormais d'un modèle comportant de nombreuses variables. C'est pour cela que nous avons commencé par effectuer une *correction de Bonferroni* sur chacune des variables initiales pour comprendre si la contribution globale d'une variable est significative. Nous parlons de test *simultané*, car nous testons 4 coefficients du modèle en même temps, à savoir la contribution d'une variable aux 4 genres musicaux. Par exemple, nous pouvons tester l'importance du caractère dansant en considérant à la fois l'importance des coefficients liés à *danceability, danceability :houserock, danceability :rap et danceability :popreggaeton.* Par ailleurs,nous avons commencé à supprimer du modèle les variables dont la p-value était trop élevée (p-value étant le seuil inférieur 0,01 pour éliminer les variables d'une valeur , c'est-à-dire que nous obtenons un niveau de confiance de 99,9 %), ce qui signifie que la variable correspondante n'est pas significative pour le modèle.

```
       names             p_values
  Intercept          7.42055080530339e-12
  acousticness       0.000221276098752411
  danceability       1.2228968792254e-17
  duration           0.0142877628657674       <- remove
  energy             2.40817873412393e-13
  instrumentalness   0.0688521994275237       <- remove
  liveness           0.0020871101327196
  loudness           3.88135355820952e-41
  speechiness        3.98369427641703e-16
  tempo              0.00720003339164047
  time_signature     0.891077757836686        <- remove
  valence            1.38247730966775e-31
  keymode            0.0528188843523066       <- remove
  artist_popularity  0
```

On commence par supprimer : *duration, instrumentalness, time signature, keymode*. Ensuite, on a pu construire un nouveau modèle avec les variables restantes et avons répété la même procédure pour finalement éliminer également la variable *tempo*.
Nous procédons par la suite à la réduction *stepwise*, nous permettant d'éliminer une ou plusieurs de ses 4 contributions aux genres musicaux. Toutefois nous sommes parvenus à ne négliger que 3 coefficients sur 33, nous avons donc préféré ne pas considérer cette nouvelle réduction pour préserver l'exhaustivité du modèle.
Notre modèle final est donc le suivant :

```
model0 = lm(popularity ~ houserock + rap + popreggaeton +
    acousticness + acousticness:houserock
            + acousticness:rap + acousticness:popreggaeton +
    danceability + danceability:houserock
            + danceability:rap + danceability:popreggaeton +
    energy + energy:houserock + energy:rap + energy:popreggaeton +
            + instrumentalness:rap + instrumentalness:popreggaeton +
    liveness + liveness:houserock + liveness:rap + liveness:popreggaeton +
    loudness + loudness:houserock + loudness:rap + loudness:popreggaeton +
    speechiness + speechiness:houserock
            + speechiness:rap + speechiness:popreggaeton +
    valence + valence:houserock + valence:rap + valence:popreggaeton +
    artist_popularity
)
```

Le R^2 du modéle est très élévé, 0.8321, en regardant les estimations des coefficients :

Figure 3.26 – Estimations des coefficients en fonction de notre modèle

On constate clairement que la popularité d'une chanson est principalement décrite par la popularité de l'artiste et que toutes les autres variables apportent une contribution mineure. Deux raisons peuvent l'expliquer : Premièrement, il est raisonnable de penser que la popularité de l'artiste aura beaucoup plus d'influence sur la popularité d'une chanson que n'importe quelle caractéristique musicale qui lui est attribuable ; Deuxièmement, pour des raisons pratiques, nous avons créé la variable *popularité de l'artiste* à partir de la popularité elle-même, il y a donc forcément une forte dépendance 3.26.
Nous voulons malgré tout nous concentrer sur les valeurs des coefficients des autres variables, parce que, même si elles apportent une contribution mineure à la popularité, elles peuvent nous donner des informations très utiles sur ce qui rend les chansons populaires pour chaque genre musical.

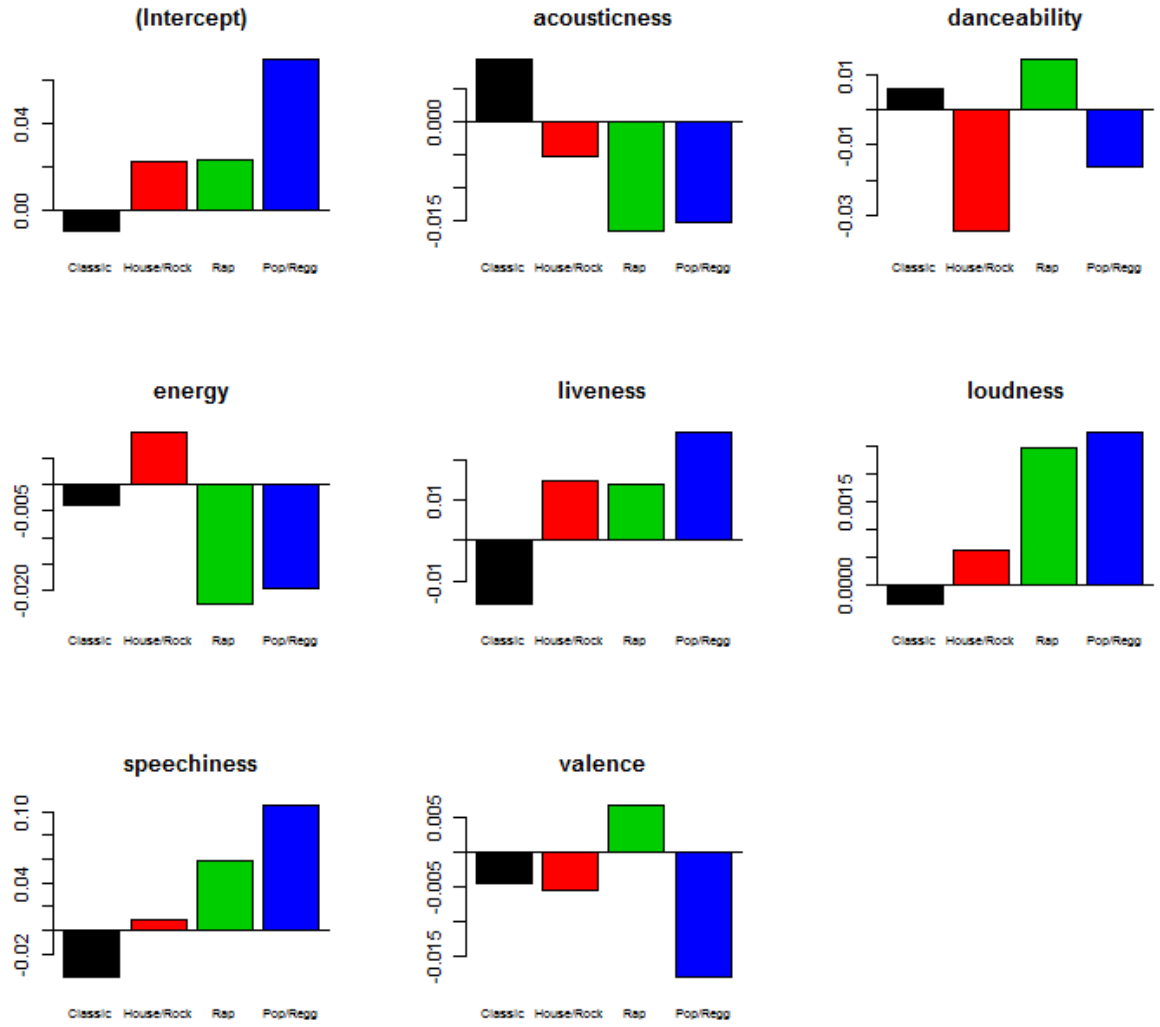

FIGURE 3.27 – Diagrammes en bâtons du poid de chaque variable explicative selon les différents genres musicaux

Intercept : son estimation est très positive pour la *pop-reggaeton*, positive pour le *rap* et la *house-rock*, alors qu'elle est négative pour la musique *classique*. Bien qu'elle représente simplement l'appartenance à un groupe musical, elle a une signification très importante : l'appartenance à un genre musical est déjà un facteur qui influence la popularité ! Le seul fait qu'une chanson soit du genre *pop-reggaeton* la rend déjà plus populaire qu'une chanson du genre *classique*.

Acousticness : Non seulement les chansons *classiques* sont caractérisées par des valeurs élevées d'acoustique, mais plus une chanson classique est acoustique, plus elle est populaire. Alors que l'acoustique a un effet opposé sur les autres genres musicaux, elle influence en effet négativement la popularité des chansons *house-rock, rap* et *pop-reggaeton*.

Danceability : Elle apporte une contribution positive dans la musique *classique* et le *rap*, alors qu'elle apporte une contribution négative à la *pop-reggaeon* et, surtout à la *house-rock*.

Energy : Un peu contre-intuitivement, l'énergie apporte une contribution négative à la popularité, sauf pour la musique *house-rock*.

Liveness, Loudness et Speechiness : Ces variables influencent positivement la popularité, sauf pour la musique *classique*. En particulier, Speechines qui apporte une forte contribution (coefficient d'ordre 10^{-1}, alors que les coefficients de Liveness et Loudness se situent respectivement autour de 10^{-2} et 10^{-3}).

Valence : Il s'avère que les chansons tristes ont une contribution plus positive que les chansons dites joyeuses,à l'exception du *rap* où le phénomène est inversé.

3.4.5 Test des Hypothéses du Modèle

On teste si les hypothéses nécessaires à la construction d'un modéle linéaire sont robustes, nous vérifions alors si nous avons une *normalité*, l'*homoscédascité*, et l'*indépendance* des résidus.

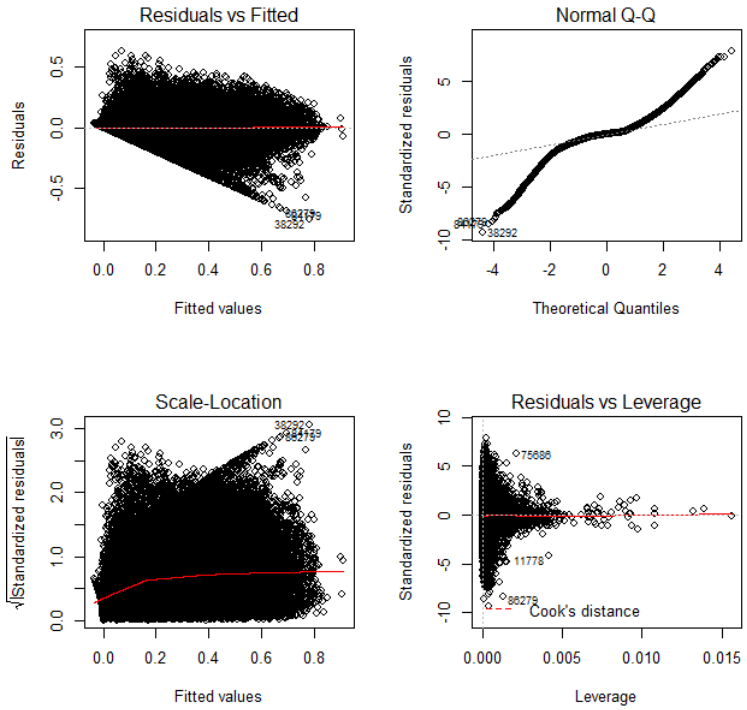

FIGURE 3.28 – Evaluation des hypothéses de validité d'un modéle linéaire

Ces hypothèses ne sont pas vraiment vérifiées pour ce modéle.
En effet,l'*homoscédasicité* est faible (pas de distribution homogène des résidus) et la *normalité* n'est pas forte également (*q-q plot* n'est pas très précis). Pourtant, il semble que nous n'ayons pas de problème de *colinéarité* dans notre modèle réduit final, puisque toutes les valeurs de *Variance Inflation Factors* (VIF) sont inférieures à 2.5 (donc inférieures à 10) :

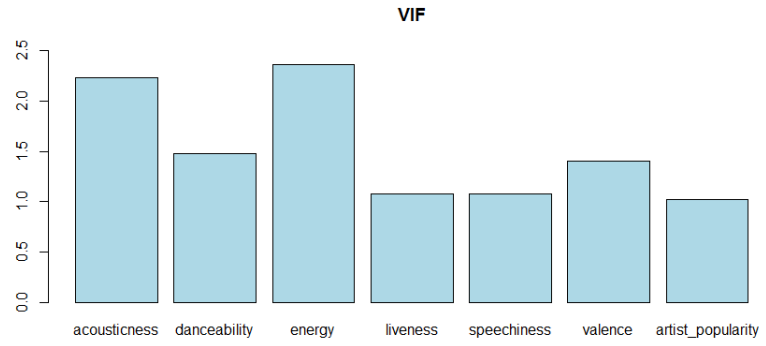

FIGURE 3.29 – Diagramme en bâtons du critère VIF

3.4.6 Prédiction

Nous souhaitons tester la puissance de prédiction de notre modèle au moyen du sous-ensemble *data test* que nous avons extrait de notre ensemble de données. Dans le but d'obtenir une visualisation graphique propre, nous trions les chansons en fonction de leur popularité croissante. Nous avons effectuer une prédiction de la popularité et construit un tableau pour comparer les valeurs prédites avec les valeurs réelles, à savoir les valeurs de *popularity*. Nous avons utilisé les commandes suivantes :

```
#Sort in ascending popularity:
data_test = data_test[order(data_test$popularity),]

#Prediction with our final model:
predicted_pop <- predict(model1, data_test[,])

#Table with the REAL VALUES of POPULARITY and the PREDICTED VALUES of POPULARITY:
true_predicted = data.frame(true = data_test$popularity, predicted = predicted_pop)
```

En traçant les valeurs prédites et les valeurs réelles sur le même graphique, on obtient :

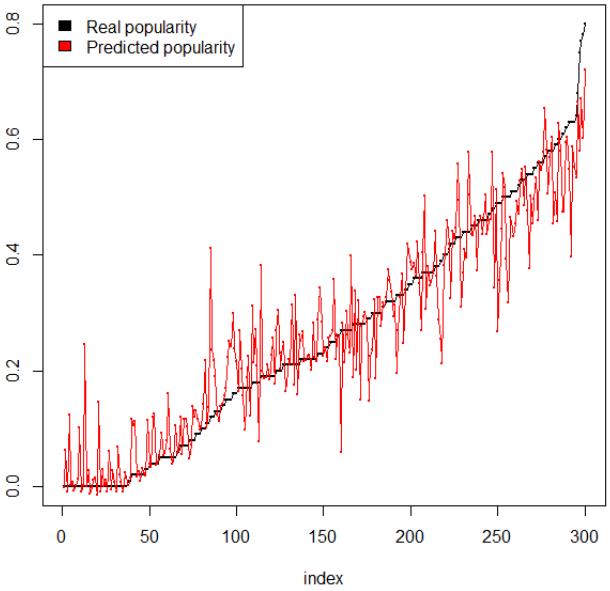

FIGURE 3.30 – Compraison des valeurs prédites de la popularité avec celles de notre base de données

D'après le graphique, nous semblons remarquer que :
- Pour les petites valeurs de la popularité réelle, les valeurs prédites de notre modèle sont généralement plus élevées que celles-ci.
- Pour les valeurs élevées de la popularité réelle, les valeurs prédites de notre modèle sont généralement plus petites.

Nous pouvons souligner cette particularité en traçant un graphique de la différence entre la popularité dite *prédite* et *réelle* des chansons et en classant comme *peu populaires* les 100 chansons ayant la plus faible popularité, comme *moyennement populaires* les 200 chansons suivantes par ordre croissant de popularité et comme *très populaires* les 300 dernières. Nous avons ajouté deux lignes noires en pointillés aux valeurs 0,08 et -0,08. Les chansons entre ces lignes sont celles dont l' *erreur* est inférieure à 8%.

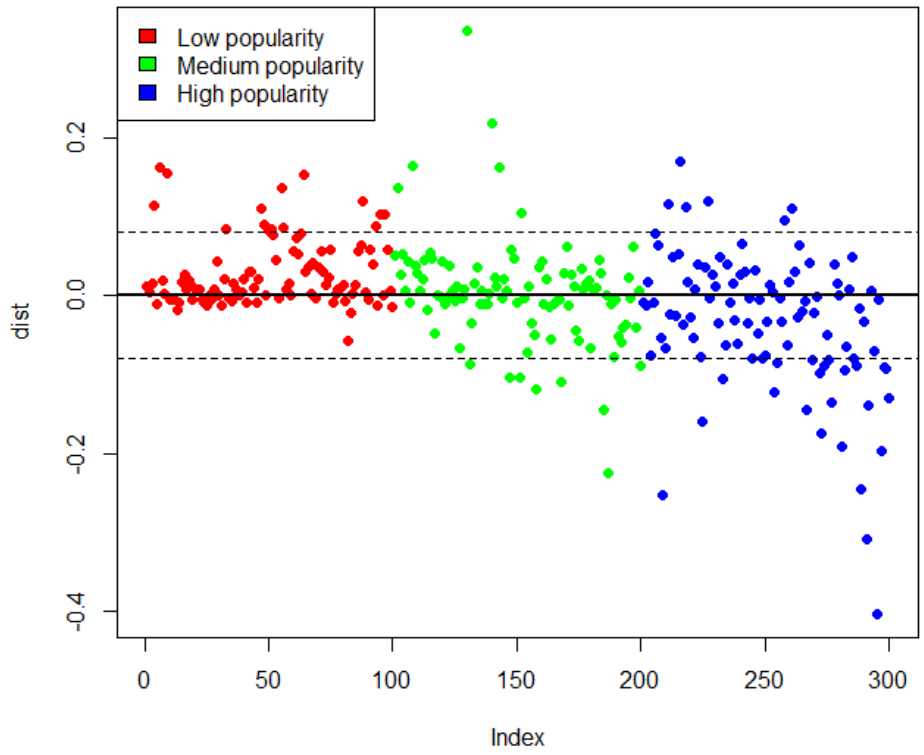

FIGURE 3.31 – Classement des chansons en fonction de leur niveau de popularité

Ce que nous avons repéré auparavant semble plus clair : notre modèle tend à sous-estimer la popularité dans le cas d'une faible popularité des valeurs réelles et à la surestimer pour les chansons très populaire. Notre modèle repose principalement sur la popularité de l'artiste, qui est calculée en prenant la *moyenne* de toutes les chansons d'un artiste, c'est pourquoi notre prédiction est biaisée positivement ou négativement en cas de très faible popularité réelle ou de très forte popularité réelle respectivement.

De plus, nous pouvons voir qu'une grande quantité de données (précisément les 80 %) se trouve entre les deux lignes pointillées, ce qui indique que la différence entre la popularité prédite et les valeurs réelles de la popularité est inférieure à 0,08 (8 %). L'erreur moyenne est de 0,0481 (environ 5%), tandis que sa variance est de 0,0032, ce qui indique un très bonne prédiction du modèle.

4. Conclusion

À la fin de cette analyse, nous ne sommes pas en mesure de donner au lecteur une recette parfaite pour être dans le top 10 sur Spotify. Ce qu'on pu relevé comme point important est que la popularité d'un titre dépend fortement de la popularité antérieure de son artiste, comme nous l'avons repéré dans notre modèle.
Quoi qu'il en soit, en fonction du genre musical des chansons, nous avons pu fournir quelques suggestions, en nous appuyant sur les résultats présentés dans la figure 3.27.
Nous avons mis en évidence que les utilisateurs de Spotify qui écoutent de la *musique classique* préfèrent les chansons à forte acoustique et qui se prêtent à la danse. Si vous êtes plutôt un artiste qui souhaite faire de la *musique house-rock*, vous devez tenir compte du fait que les caractéristiques les plus influentes sont l'énergie et la vivacité de votre morceau. Si vous vous intéressez par la musique *pop-reggaeton*, il est préférable de vous concentrer sur les paroles, le volume sonore et préféré une thématique de chanson assez triste (faible valence). Enfin, si vous êtes un *Rappeur*, vous devriez vous occuper du caractère dansant de votre morceau et du texte, tout en réduisant l'acoustique et l'énergie.

Les limites de notre modèle proviennent des faits suivants. Tout d'abord, le modèle est basé sur les subdivisions en genres musicaux, ce qui est déjà une estimation que nous avons dû calculer par l'algorithme des K-means. Deuxièmement, les hypothèses requises pour la régression et l'ajustement d'un modèle linéaire ne sont pas fortes, comme nous l'avons souligné dans la section **3. 4. 5**. Par conséquent, même si nous n'avons pas pu atteindre notre objectif (ambitieux) de décrire précisément les caractéristiques qui déterminent la popularité d'un artiste ou d'une chanson, nous sommes satisfaits des résultats obtenus dans la recherche d'un moyen de classer les chansons dans leurs genres musicaux et dans la capacité de prédiction du modèle que nous avons construit.

5. Appendice

.1 Definitions

Statistique Exploratoire ou Descriptive : Les objectifs de la statistique descriptive sont :

1. Résumer, synthétiser l'information contenue dans la série statistique, mettre en évidence ses propriétés.
2. Suggérer des hypothèses relatives àla population dont est issu l'échantillon.

De nombreux outils sont utilisés tels que les Tableaux (table des fréquences, de contingence, ...), Graphiques (box-plots, histogrammes,..) et des Indicateurs (moyenne, corrélation,..). L'utilisation de ces outils dépendent de la nature de la série (uni oumultidimensionnelle) et des variables (quantitatives discrètes, continues ou qualitatives).

Dans notre étude, nous allons procédé à une statistique multi-variés. Le principe est le même que pour une seule variable, sauf que toutes les caractéristiques (moyenne, mode, écart type, etc) sont bi variées (des vecteurs). Il y a d'autre part une caractéristique supplémentaire : la corrélation. Elle est une mesure linéaire de la dépendance entre les différentes composantes de la variable multi variée.

Statistique Inférentielle : L'inférence statistique est l'ensemble de techniques permettant d'induire les caractéristiques d'un groupe général (la population) à partir de celles d'un groupe particulier (l'échantillon), en fournissant une mesure de la certitude de la prédiction : la probabilité d'erreur.

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

ACM ou Analyse des Correspondances Multiples : Cette méthode revient à effectuer une AFC du tableau disjonctif complet. En ACM, on traitece tableau disjonctif comme une table de contingence.

AFC ou Analyse Factorielle des Correspondances : Cette méthode revient à réaliser l'ACP du tableau des profils-lignes afin d'obtenir la représentation graphique des individus. Ensuite, on procéde de la même manière sur le tableau profils-colonne. On peut finalement montrer que ces deux graphiques se superposent et en déduire la liaison entre deux variables

Clustering ou Classification non supervisée : Un processus qui permet de rassembler des données similaires. Le fait qu'il ne soit pas supervisé signifie que des techniques d'apprentissage machine vont permettre de trouver certaines similarités pour pouvoir classer les données et ce de manière plus ou moins autonome. Ce type d'analyse permet d'avoir un profil des différents groupes. Cela permet donc de simplifier l'analyse des données en faisant ressortir les points commun et les différences et en réduisant ainsi le nombre de variable des données.

K-means ou Partitionnement en K-moyennes : Une méthode de partitionnement de données et un problème d'optimisation combinatoire. Étant donnés des points et un entier k, le problème est de diviser les points en k groupes, souvent appelés clusters, de façon à minimiser une certaine fonction. On considère la distance d'un point à la moyenne des points de son cluster ; la fonction à minimiser est la somme des carrés de ces distances

Régression linéaire : modèle de régression linéaire est un modèle de régression qui cherche à établir une relation linéaire entre une variable, dite expliquée, et une ou plusieurs variables, dites explicatives.En général, le modèle de régression linéaire désigne un modèle dans lequel l'espérance conditionnelle de y sachant x est une transformation affine Modèle :Lourd paramètres. Cependant, on peut aussi considérer des modèles dans lesquels c'est la médiane conditionnelle de y sachant x ou n'importe quel quantile de la distribution de y sachant x qui est une transformation affine en les paramètres

Inertie : C'est la mesure de la dispersion du nuage de point en se basant sur la distance euclidienne et le centre de gravité du nuage (moyennes des individus). $I_T = \frac{1}{n} \sum_{i=1}^{n} d(G;x_i)^2$.

AFMD ou Analyse Factorielle des données mixtes : une méthode destinée à analyser un jeu de données contenant à la fois des variables quantitatives et qualitatives. Elle permet d'analyser la similitude entre les individus en prenant en compte des variables mixtes. De plus, on peut explorer l'association entre toutes les variables, tant quantitatives que qualitatives. Pour faire simple, l'algorithme AFDM peut être considéré comme mixte entre l'analyse en composantes principales (ACP) et l'analyse des correspondances multiples (ACM). En d'autres termes, il agit comme l'ACP concernant les variables quantitatives et comme l'ACM concernant les variables qualitatives.

.2 Algorithme K-means

Entrées :
- K : le nombre de cluster à former
- Training Set (matrice de données)

DEBUT
1. Choisir aléatoirement K points (une ligne de la matrice de données). Ces points sont les centres des clusters (nommé centroïd).

REPETER
2. Affecter chaque point (élément de la matrice de donnée) au groupe dont il est le plus proche au son centre
3. Recalculer le centre de chaque cluster et modifier le centroide

JUSQU'A CONVERGENCE
OU (stabilisation de l'inertie totale de la population)
FIN

Bibliographie

[1] Université Aix-Marseille, *Cours 2 : Rappels de Statistique descriptive*, 2019. http://iml.univ-mrs.fr/~reboul/cours2.pdf

[2] Ph.D Villiers Fannyn, *Cours d'Analyse De Données*, 2020.

[3] Richard A. Johnson, Dean W. Wichern *Applied Multivariate Statistical Analysis* Pearson, sixth edition.