Projet Calcul Financier & Statistique : En avant la musique !

MAHI Riad & Jérôme GAMBIEZ

Mise à jour, le 25 Mai 2023

Contents

1. Introduction

Pour ce devoir de stastique nous avons choisi un sujet un peu original qui est d'utiliser des musiques provenant de site de streaming connu comme Spotify ou encore Youtube. Notre travail va être d'analyser un jeu de donnée de 50 titres parmis les 5 artistes les plus connus des plateforme afin d'en déduire une analyse sur des questions qu'on a pu se poser notamment la corrélation entre genre en dancabilité. En effet, plusieurs questions nous sont venu lors de la première étude du jeu de donnée mais également en cours de route notammen, est-ce l'energie et la dansabilité d'une musique ont une relation ? Est-ce l'énergie d'une musique à un impact sur la sensibilité d'un utilisateur à aimer la musique ? Même question pour la dansabilité?

La source de notre jeu de donnée provient du site Kaggle.com:
https://www.kaggle.com/datasets/salvatorerastelli/spotify-and-youtube

Mise en place du jeu de donnée:

```
  library(ggplot2)
  library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
dataPath <- "./Spotify_Youtube.csv"
data <- read.csv(dataPath, header=TRUE, stringsAsFactors=FALSE, nrows=50)
```

Variables qualitative

Nous avons déduit à partir des données extraites des variables qualitatives à minima pertinente.

Notre première variable qualitative est de savoir si la musique est un album ou un single, on la nomera **estAlbum**. Egalement, nous avons choisi de prendre comme variable qualitative la **fréquence d'écoute d'un titre**. Voici une définition des indicateurs de la fréquence d'écoute en accord avec l'échelle des données que nous possédons.

- Peu écouté : Cette catégorie désigne les chansons qui ont reçu un nombre de vues inférieur ou égal à 100 millions. Cela signifie que ces chansons ont été relativement moins populaires ou ont été moins diffusées en comparaison avec d'autres chansons.
- Moyen écouté : Cette catégorie concerne les chansons qui ont reçu un nombre de vues compris entre 100 millions et 500 millions. Ces chansons peuvent être considérées comme ayant une popularité moyenne ou une diffusion modérée par rapport à d'autres chansons.
- Très écouté : Cette catégorie comprend les chansons qui ont accumulé un nombre de vues supérieur à 500 millions. Ces chansons sont considérées comme ayant une grande popularité ou une diffusion élevée, attirant un grand nombre de spectateurs ou d'auditeurs.

```r
#Récupération de la colonne album_type (album ou single)
album_type <- table(data$Album_type)
album_type_df <- as.data.frame(album_type)
album_type_df <- album_type_df[order(album_type_df$Freq, decreasing = TRUE), ]

#Récupération et traitement de la colonne fréquence écoute
data <- data %>%
  mutate(Freq_listen = case_when(
    data$Views <= 100000000 ~ "peu écouté",
    data$Views > 100000000 & data$Views <= 500000000 ~ "moyen écouté",
    data$Views > 500000000 ~ "très écouté",
    TRUE ~ NA_character_
  ))

freq_listen <- table(data$Freq_listen)

freq_listen_df <- as.data.frame(freq_listen)
freq_listen_df <- freq_listen_df[order(freq_listen_df$Freq, decreasing = TRUE), ]
```

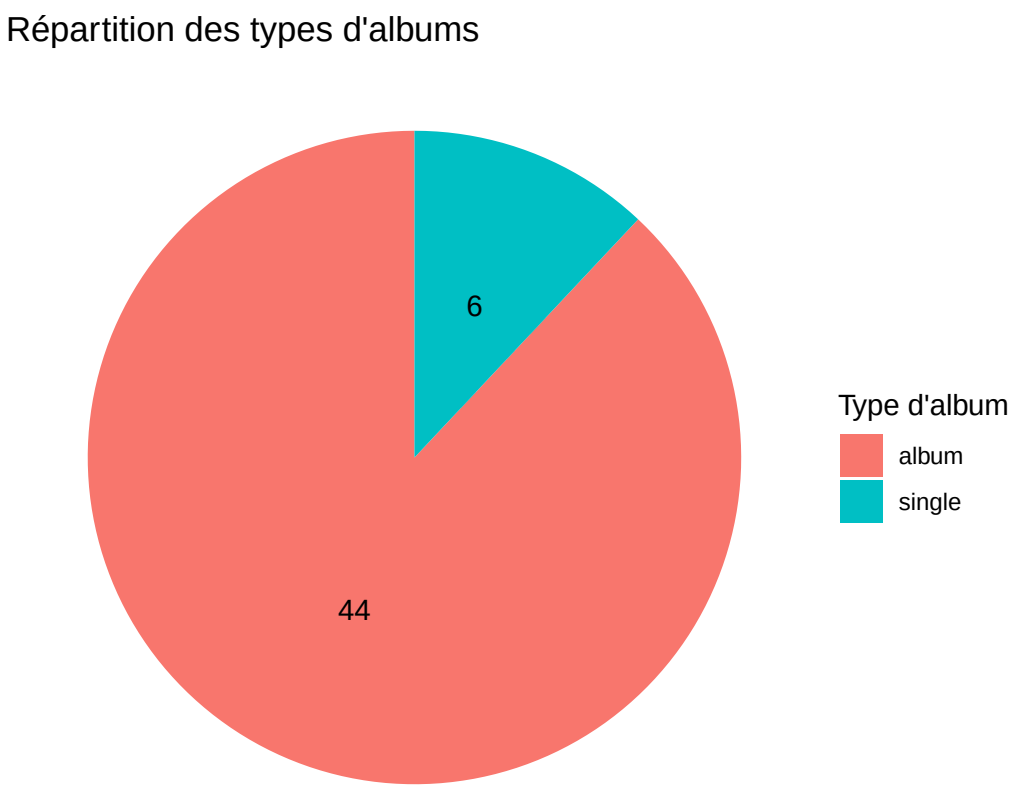
Variables quantitative

Les variables quantitatives sont **la dancabilité** d'une musique, qui est un indicateur sur la probabilité de dancer sur une musique. Puis **l'énergie** qui représente une mesure subjective qui nous aide à évaluer et à caractériser le niveau d'intensité et d'activité ressenties dans un morceau de musique.

2. Analyse univariée

Répartitions des musiques album/single

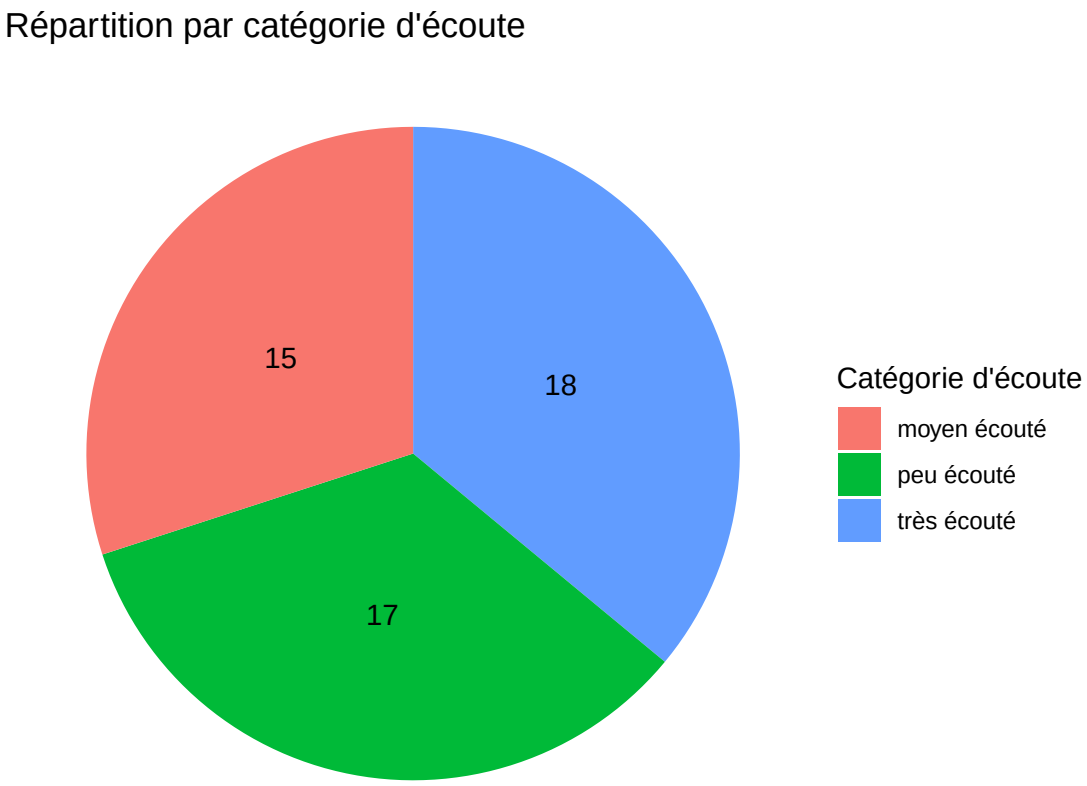
```r
pie_chart <- ggplot(album_type_df, aes(x = "", y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Répartition des types d'albums", fill = "Type d'album") +
  theme_void() +
  geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), size = 4)

pie_chart
```



On observe que parmis notre échantillon, il y a près de 90% musique provenant d'album et 10% de single. Les titres d'albums sont dont prédominant dans nos données.

Répartition des fréquences d'écoute

```r
pie_chart_freq_list <- ggplot(freq_listen_df, aes(x = "", y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  labs(title = "Répartition par catégorie d'écoute", fill = "Catégorie d'écoute") +
  theme_void() +
  geom_text(aes(label = Freq), position = position_stack(vjust = 0.5), size = 4)

pie_chart_freq_list
```

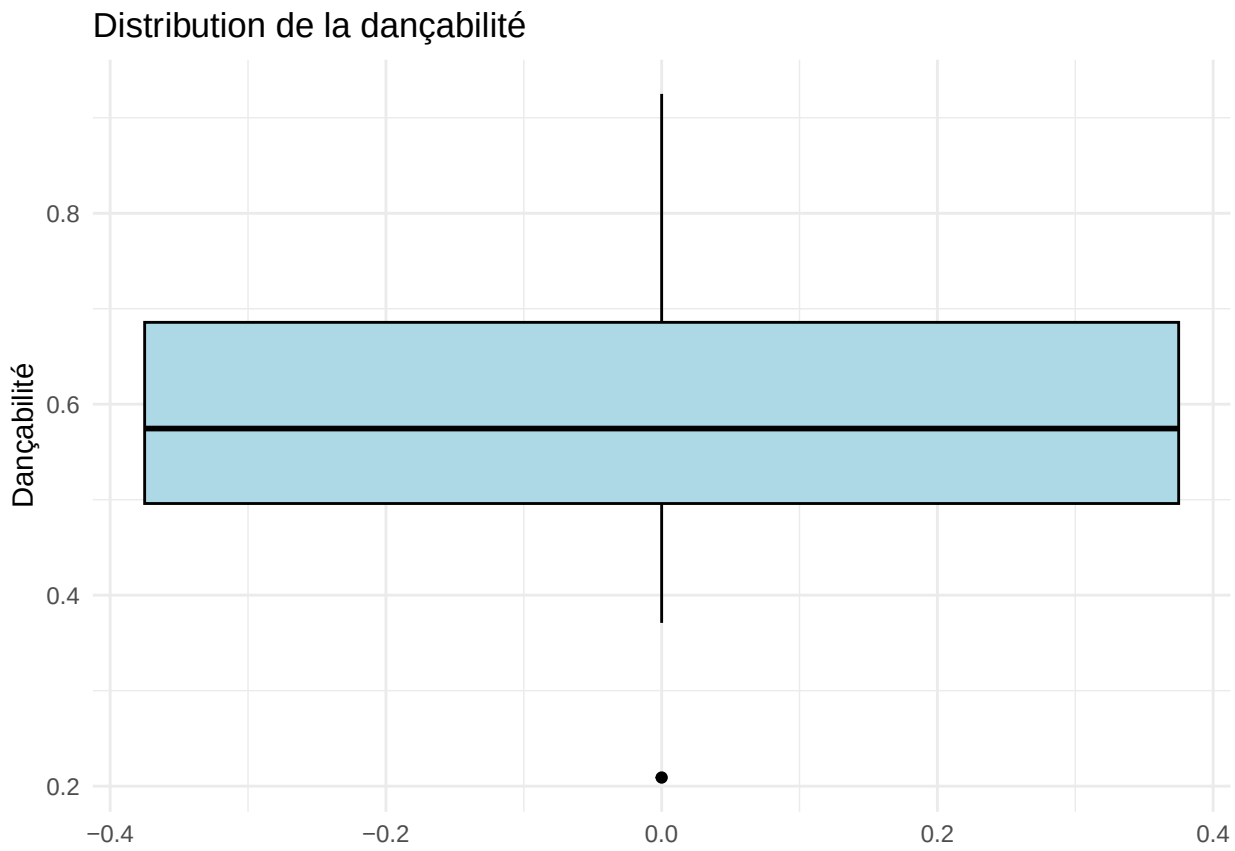


On peut constater que nos fréquences d'écoutes sur les différents titres de nos échantillons on la même proportion à un ou deux titres près.

Représentation graphique de la répartition

Dancabilité

```
boxplot_danceability <- ggplot(data, aes(y = Danceability)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Distribution de la dançabilité", y = "Dançabilité") +
  theme_minimal()

boxplot_danceability
```



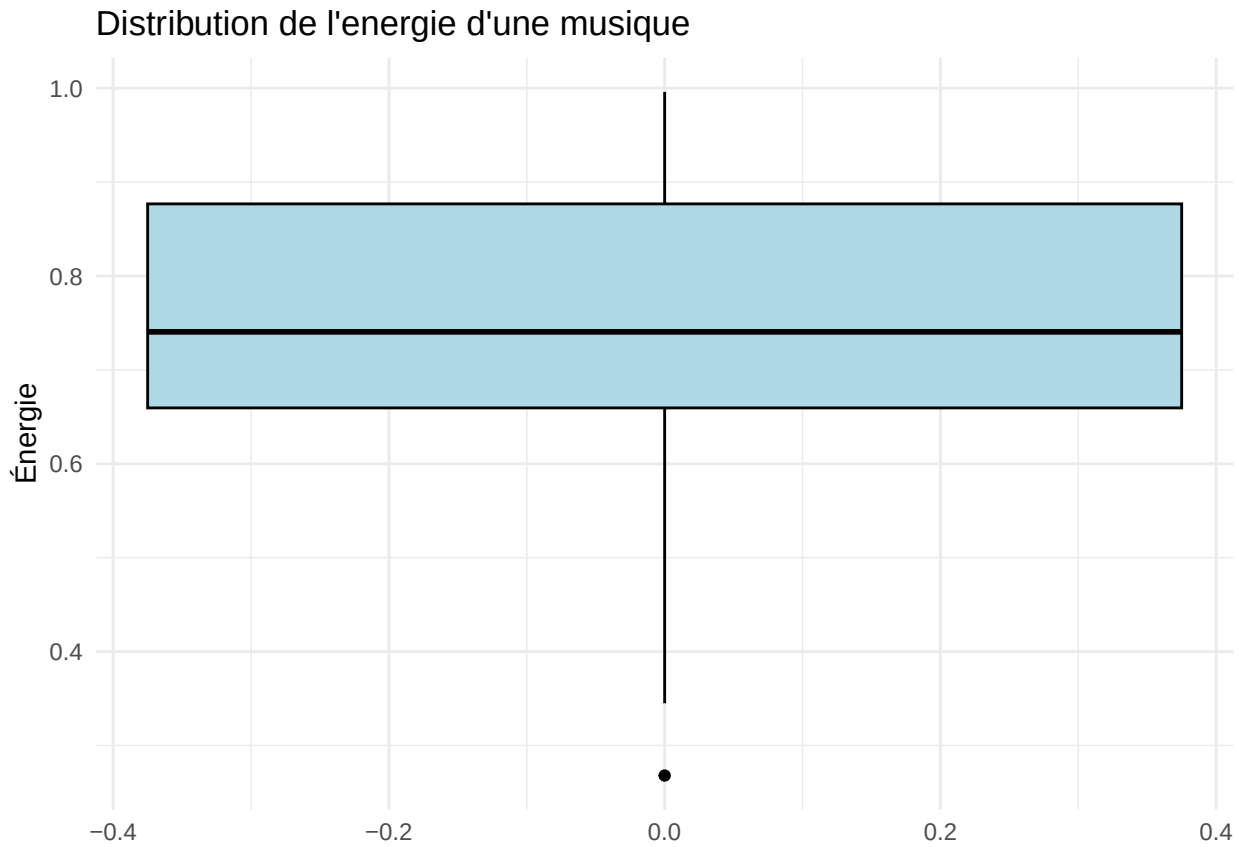
En examinant le diagramme, on peut constater que la majorité des musiques analysées, soit environ 58 %, ont une danceability élevée. Cela indique que ces morceaux sont adaptés à la danse et ont un potentiel élevé pour inciter les gens à se déplacer et à bouger sur la piste de danse.

De plus, le diagramme révèle que 25 % des musiques ont un effet dansant sur environ 70 % de la population. Cela suggère que ces morceaux sont particulièrement accrocheurs et entraînants, capables d'engager un large public et de susciter l'envie de danser chez une majorité de personnes.

Enfin, il est intéressant de noter que 50 % des musiques ont un effet dansant sur une personne sur deux. Cela signifie que la danceability de ces morceaux est plus subjective et peut varier d'une personne à l'autre. Certains individus peuvent ressentir une forte envie de danser en les écoutant, tandis que d'autres peuvent ne pas être aussi réceptifs à leur effet dansant.

Énergie

```
boxplot_energy <- ggplot(data, aes(y = Energy)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Distribution de l'energie d'une musique", y = "Énergie") +
  theme_minimal()
boxplot_energy
```



L'analyse du diagramme en boîte à moustaches révèle plusieurs informations importantes concernant l'énergie des morceaux de musique. Plus l'énergie est proche de 1.0, plus elle est rapides, fortes et bruyantes.

En examinant les différentes parties du diagramme, nous pouvons constater que la médiane à 0.75. On peut donc conclure 75% des morceaux analysés présentent une énergie relativement élevée supérieur à 0.67.

La boîte à moustaches indique également que 25% des morceaux ont une énergie supérieure à 0,87, ce qui suggère une intensité, une vitesse et un niveau sonore plus élevés. Ces morceaux pourraient correspondre à des genres musicaux tels que le rock, le metal ou d'autres styles de musique agressifs et dynamiques. On retrouve des groupes comme Metallica ou Red Hot Chili Peppers.

D'autre part, 25% des morceaux ont une énergie inférieure à 0,66, ce qui indique une énergie plus basse. Ces morceaux pourraient être plus calmes, plus lents et moins bruyants. On retrouve le musicien Coldlay qui propose des musiques appelé Soft rock.

Conclusion des deux moustaches

En analysant les deux diagrammes ensemble, nous pouvons effectivement observer une corrélation entre la danceability et l'énergie des morceaux de musique. Il semble y avoir une tendance où les morceaux ayant une forte danceability ont également une énergie élevée, supérieure à 0,65.

En se basant sur cette observation, il est intéressant de noter que les musiques les plus dansantes sont celles de 50 Cent, un artiste de hip-hop/rap. Ses morceaux ont une énergie se situant entre 0,65 et 0,75, ce qui correspond à une intensité sonore élevée et une propension à inciter les auditeurs à se déplacer et à danser. ## Estimation et intervalles de confiance

Analyse de la dancabilité

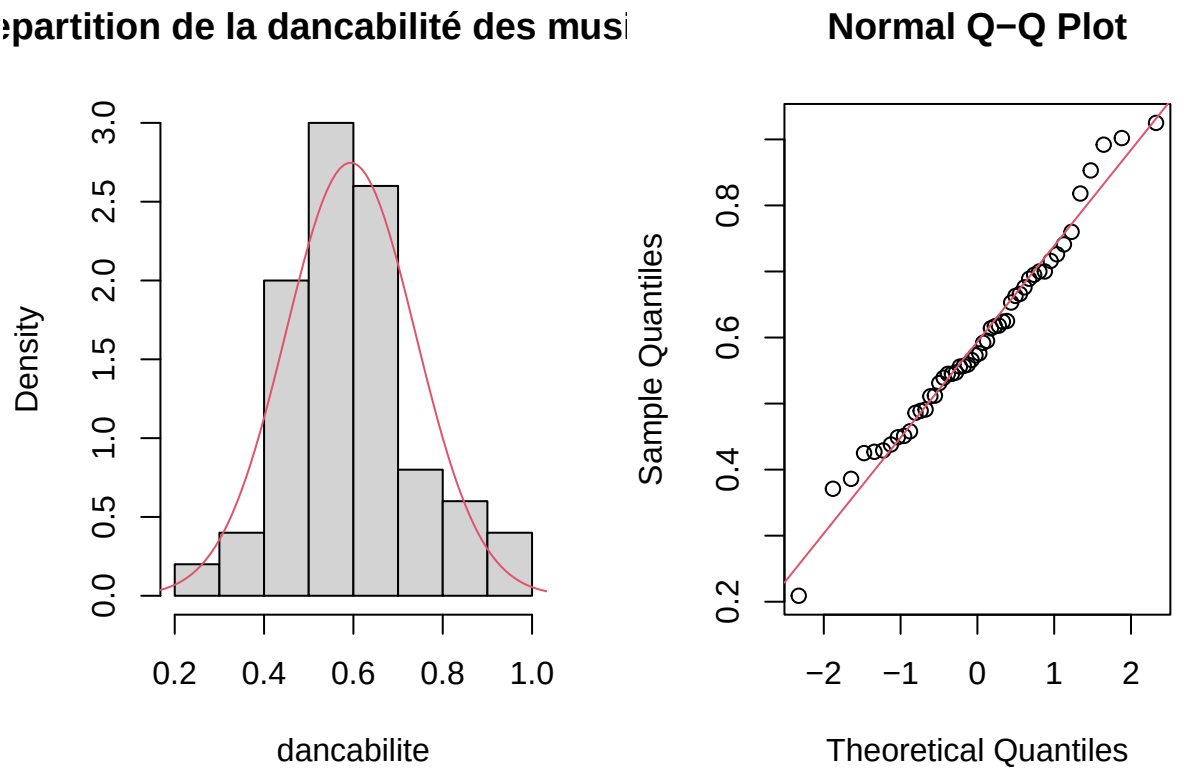
```
dancabilite  <- data$Danceability

var(dancabilite)
```

```
## [1] 0.0210791
```

```
mean(dancabilite)
```

```
## [1] 0.59372
```

```
par(mfrow=c(1,2))
hist(dancabilite , main="repartition de la dancabilité des musiques ",xlab="dancabilite",prob=T)
points(seq(0,30,0.01),dnorm(seq(0,30,0.01),mean(dancabilite),sd(dancabilite)),col=2,type="l")
qqnorm(dancabilite)
abline(mean(dancabilite),sd(dancabilite),col=2)
```




```
interval=t.test(dancabilite)
interval$conf.int
```

```
## [1] 0.5524585 0.6349815
## attr(,"conf.level")
## [1] 0.95
```

L'histogramme représente la répartition de la danceability des morceaux de musique étudiés. La variable dancabilite a une variance (var) de 0.0210791 et une moyenne (mean) de 0.59372. Cela indique que les valeurs de danceability sont relativement concentrées autour de la moyenne et présentent une faible dispersion.

Analyse de l'energie

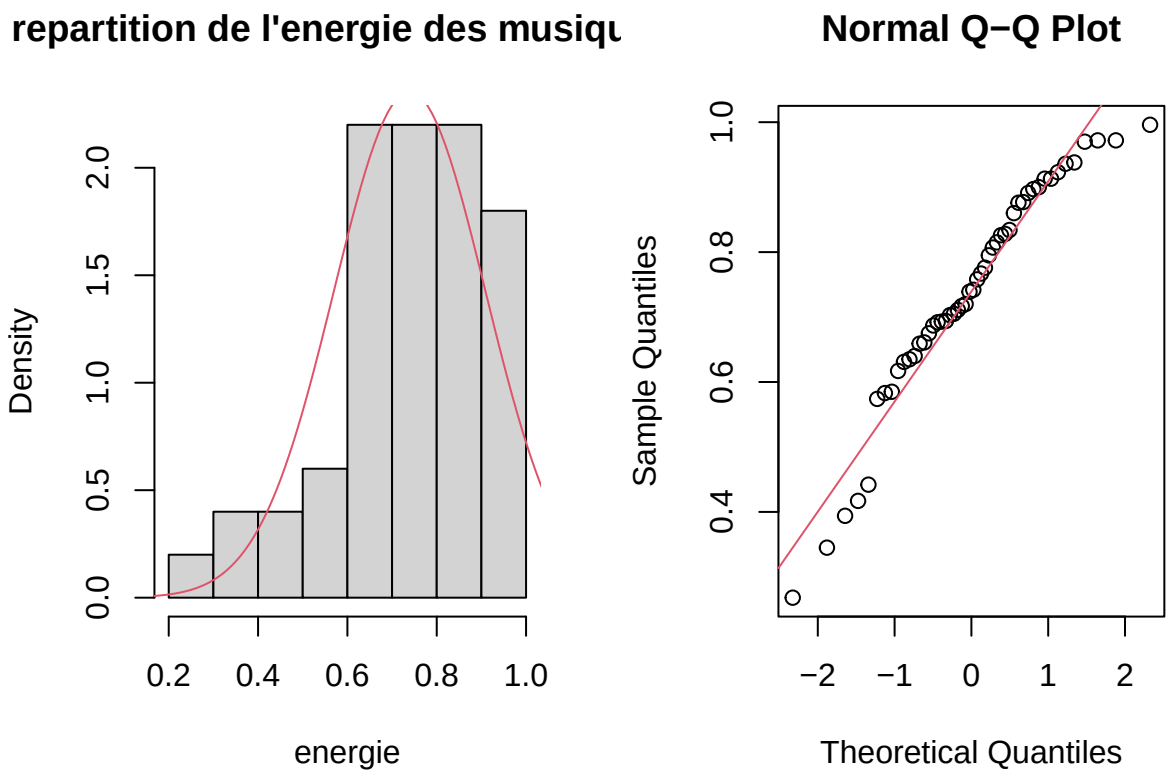
```r
energie <- data$Energy

var(energie)
```

```
## [1] 0.02873404
```

```r
mean(energie)
```

```
## [1] 0.73938
```

```r
par(mfrow=c(1,2))
hist(energie , main="repartition de l'energie des musiques ",xlab="energie",prob=T)
points(seq(0,150,0.01),dnorm(seq(0,150,0.01),mean(energie),sd(energie)),col=2,type="l")
qqnorm(energie)
abline(mean(energie),sd(energie),col=2)
```




```r
interval=t.test(energie)
interval$conf.int
```

```
## [1] 0.6912055 0.7875545
## attr(,"conf.level")
## [1] 0.95
```

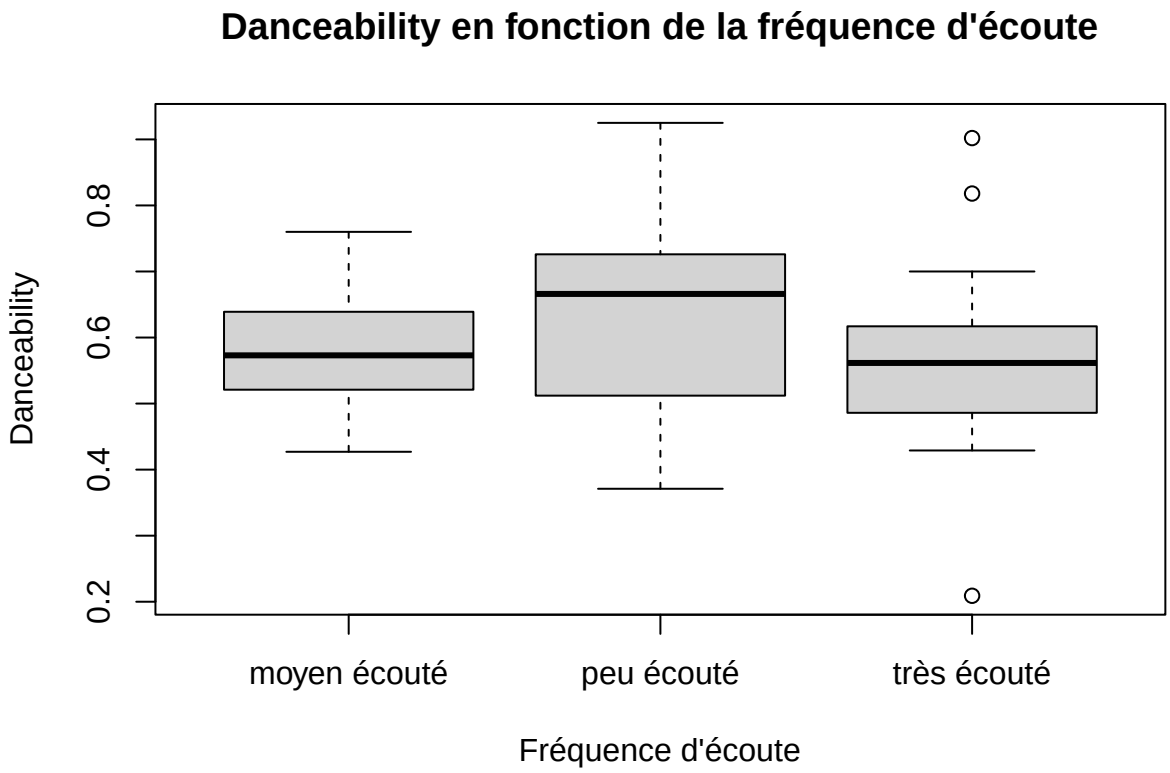
3. Analyse multivariée

Analyse quanti x quali

Nous avons choisi de prendre comme varaible quantitative **la dancabilité** et comme qualitative la **fréquence d'écoute d'un titre**.

- Donner les résumés statistiques, histogrammes et/ou bloxplot par sous-population.

```
boxplot(data$Danceability ~ data$Freq_listen, main="Danceability en fonction de la fréquence d'écoute",
```




```
summary(data$Danceability[data$Freq_listen == "peu écouté"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3710  0.5120  0.6660  0.6312  0.7260  0.9250
```

```
summary(data$Danceability[data$Freq_listen == "moyen écouté"])
```

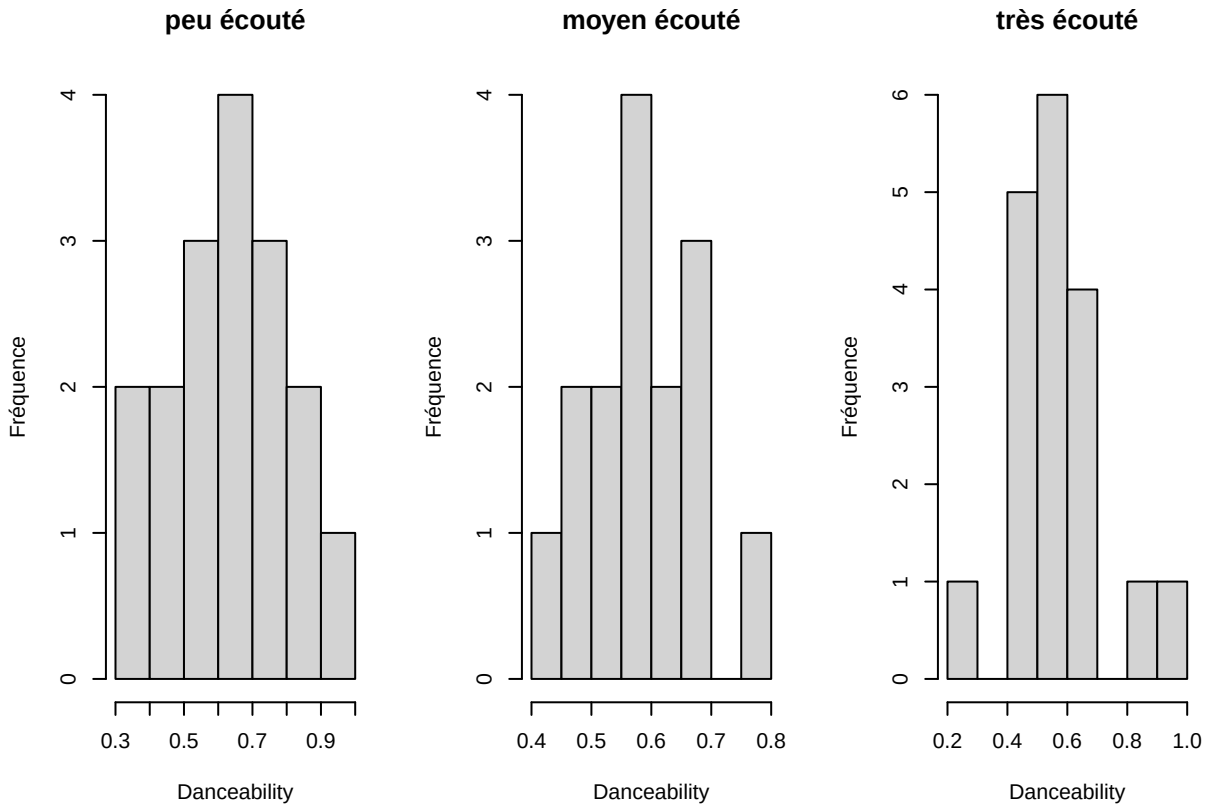
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4270  0.5210  0.5730  0.5825  0.6390  0.7600
```

```
summary(data$Danceability[data$Freq_listen == "très écouté"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2090  0.4873  0.5615  0.5677  0.6162  0.9020
```

Dancabilité par fréquence d'écoute:

```
par(mfrow=c(1,3))
hist(data$Danceability[data$Freq_listen == "peu écouté"], main="peu écouté", xlab="Danceability", ylab="
hist(data$Danceability[data$Freq_listen == "moyen écouté"], main="moyen écouté", xlab="Danceability", yl
hist(data$Danceability[data$Freq_listen == "très écouté"], main="très écouté", xlab="Danceability", ylab
```



Calcule des rapports de corrélations:

```
data$Freq_listen_numeric <- factor(data$Freq_listen)
data$Freq_listen_numeric <- as.numeric(data$Freq_listen_numeric)
correlation <- cor(dancabilite, data$Freq_listen_numeric, use = "pairwise.complete.obs")
correlation
```

```
## [1] -0.0513819
```

Il existe une corrélation très faible et négative (-0.0513819) entre la fréquence d'écoute et la danseabilité des musiques. Cela suggère qu'il n'y a pas de relation significative entre ces deux variables dans l'échantillon étudié. En d'autres termes, la fréquence d'écoute d'une musique ne semble pas être un facteur déterminant de sa danseabilité, du moins dans le contexte de ce jeu de données spécifique.

Cela signifie que, d'après les données analysées, il n'existe qu'une très légère tendance indiquant que plus un morceau de musique est écouté fréquemment, moins il est susceptible d'être dansant. Cependant, il est important de noter que cette corrélation est proche de zéro, ce qui suggère qu'il n'y a pas de relation linéaire significative entre la danseabilité et la fréquence d'écoute des morceaux.

Ce résultat suggère qu'il y a pas de preuve pour dire qu'une fréquence d'écoute élevée soit directement liée à une danseabilité.

- Faire le test d'égalité des moyennes. Spécifier les hypothèses. Conclure ?

Hypothèses :

Hypothèse nulle (H0) : Il n'y a pas de différence significative entre les moyennes des groupes de fréquence d'écoute.

Hypothèse alternative (H1) : Il existe une différence significative entre au moins deux des moyennes des groupes de fréquence d'écoute.

```
# Effectuer le test d'ANOVA
model <- aov(Danceability ~ Freq_listen, data = data)

# Résumé des résultats
summary(model)
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Freq_listen  2 0.0379 0.01896   0.896  0.415
## Residuals   47 0.9950 0.02117
```

Nous avons utilisé un seuil de signification de 0,05 pour évaluer la différence entre les groupes. La valeur p calculée pour notre test est de 0,723, ce qui est supérieur à notre seuil de signification. Cela signifie que nous n'avons pas trouvé suffisamment de preuves statistiques pour rejeter l'hypothèse nulle.

Nos résultats indiquent qu'il n'y a pas de différence significative dans la danseabilité des morceaux de musique en fonction de la fréquence d'écoute. Autrement dit, que les morceaux soient peu écoutés, moyennement écoutés ou très écoutés, cela n'a pas d'impact statistiquement significatif sur leur danseabilité.