Projet Calcul Financier & Statistique : En avant la musique !

MAHI Riad & Jérôme GAMBIEZ

Mise à jour, le 25 Mai 2023

Contents

1. Introduction

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

Dernière mise à jour du jeu de donnée: Mars 2023
Création: Février 2023

Nous avons sélectionné, dans le cadre du projet Calcul Financier & Statistique, un jeu de données provenant des plateformes de streaming musical Spotify et YouTube. Nous avons spécifiquement choisi ce jeu de données en raison de l'impact qu'ont ces applications sur la vie quotidienne de millions d'utilisateurs à travers le monde. Étant donné que ces applications sont utilisées par une population provenant de divers pays, notre échantillon ne représente pas une zone géographique spécifique.

Question à mettre en place: - 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?

Mise en place du jeu de donnée:

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

Les variables qualitatives sont les variables: - Si la musique à été publiée comme un album ou un single - La fréquence d'écoute d'un titre

1. 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.
2. 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.

3. 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.

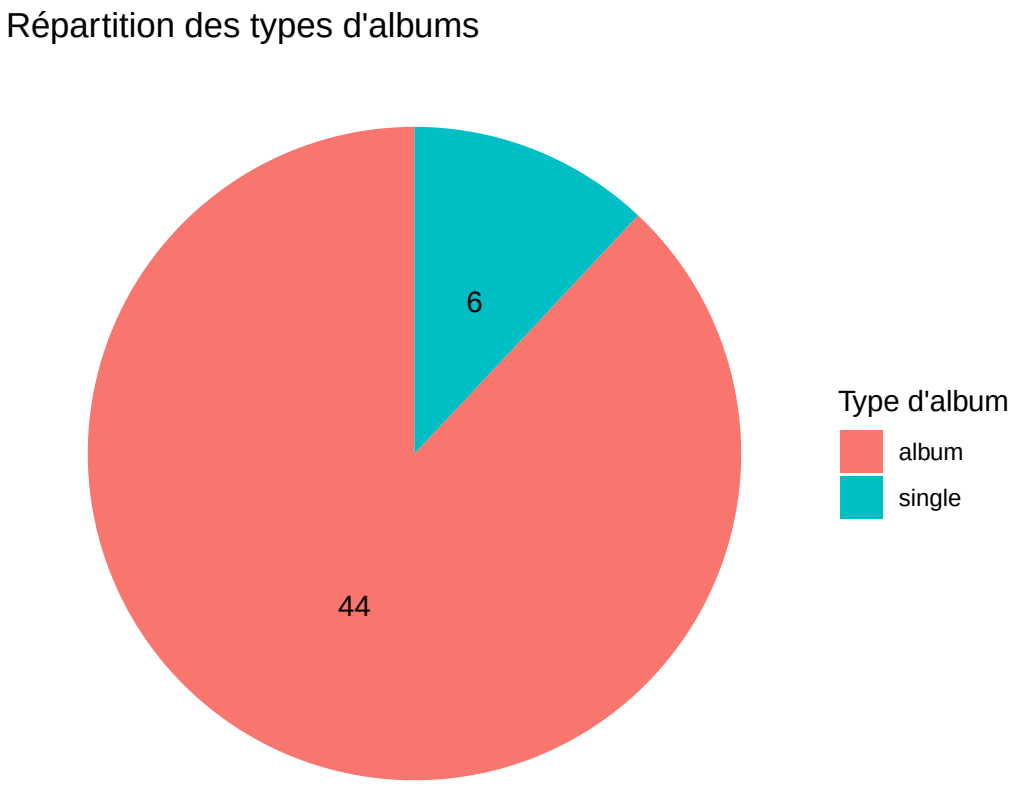
Les variables quantitatives sont: - La dancabilité d'une musique, qui est un indicateur sur la probabilité de dancer sur une musique - l'énergie que la musique à definir!

La classification en "peu écouté", "moyen écouté" et "très écouté" est basée sur le nombre d'écoute d'un titre :

2. Analyse univariée

Répartitions de la variable qualitative "type album"

```
library(ggplot2)

album_type <- table(data$Album_type)
df <- as.data.frame(album_type)
df <- df[order(df$Freq, decreasing = TRUE), ]
# Générer le pie chart
pie_chart <- ggplot(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)

# Afficher le pie chart
print(pie_chart)
```



Répartitions de la variable qualitative noms des artistes

```
library(dplyr)
```

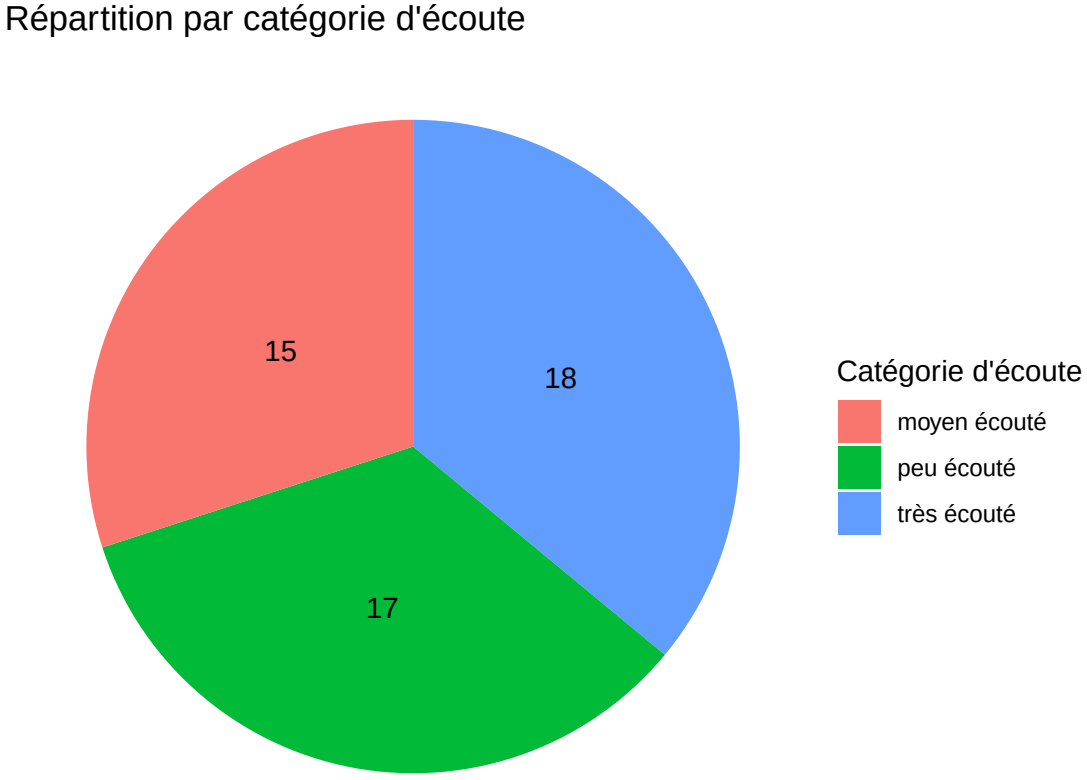
```
##
## Attaching package: 'dplyr'
```

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

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

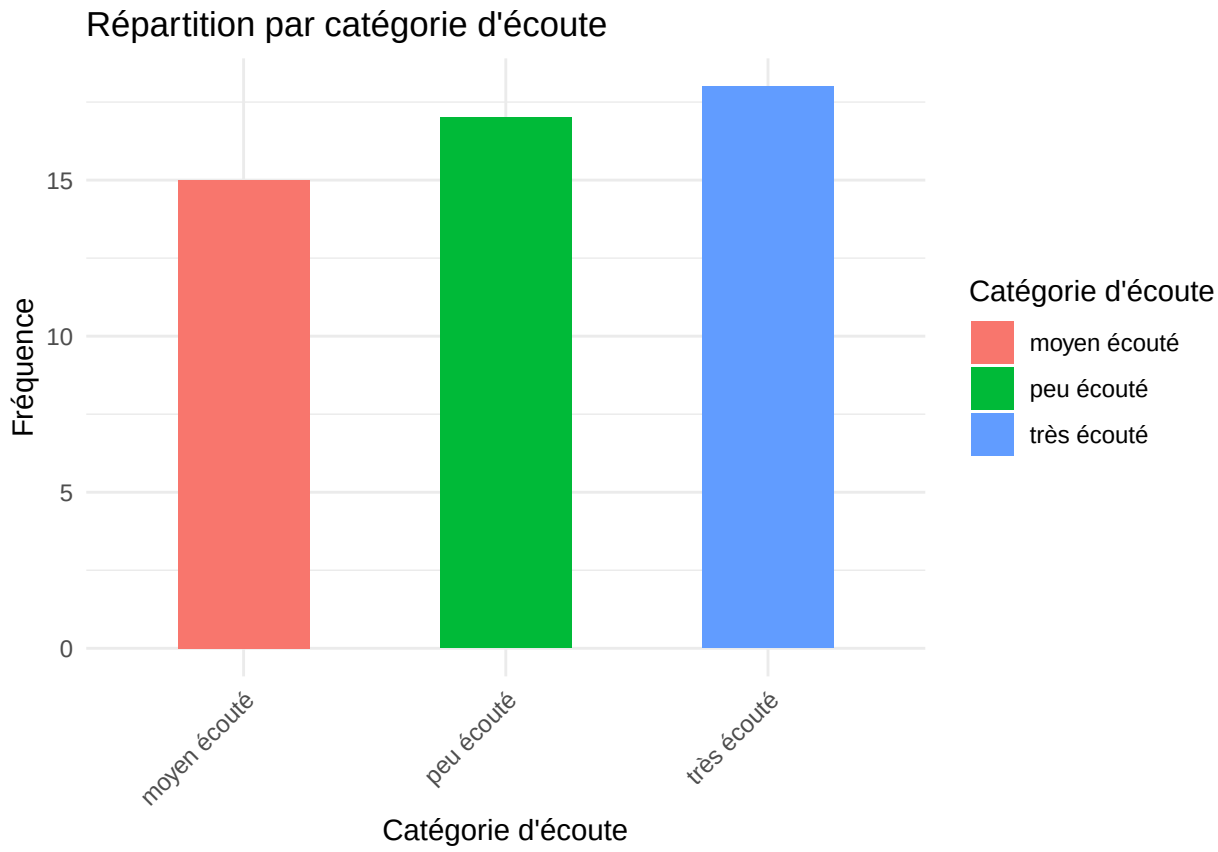
```
# Créer des catégories en fonction du nombre de vues
data <- data %>%
  mutate(Freq_listen = case_when(
    data$Views <= 100000000 ~ "peu écouté",
    data$Views > 100000000 & data$Views <= 500000000 ~ "moyen écouté",
    data$Views > 500000000 ~ "très écouté",
    TRUE ~ NA_character_
  ))

# Créer un tableau de comptage par catégorie d'écoute
```

```r
ecoute_counts <- table(data$Freq_listen)

# Convertir le tableau en data frame
ecoute_df <- as.data.frame(ecoute_counts)
ecoute_df <- ecoute_df[order(ecoute_df$Freq, decreasing = TRUE), ]

# Générer le bar plot
bar_plot <- ggplot(ecoute_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)

# Afficher le bar plot
print(bar_plot)
```




```r
# Générer l'histogramme à barres
bar_histogram <- ggplot(ecoute_df, aes(x = Var1, y = Freq, fill = Var1)) +
  geom_bar(stat = "identity", width = 0.5) +
  labs(title = "Répartition par catégorie d'écoute", x = "Catégorie d'écoute", y = "Fréquence") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_discrete(name = "Catégorie d'écoute")
```

```r
# Afficher l'histogramme à barres
print(bar_histogram)
```



Représentation graphique de la dancabilité d'une musique

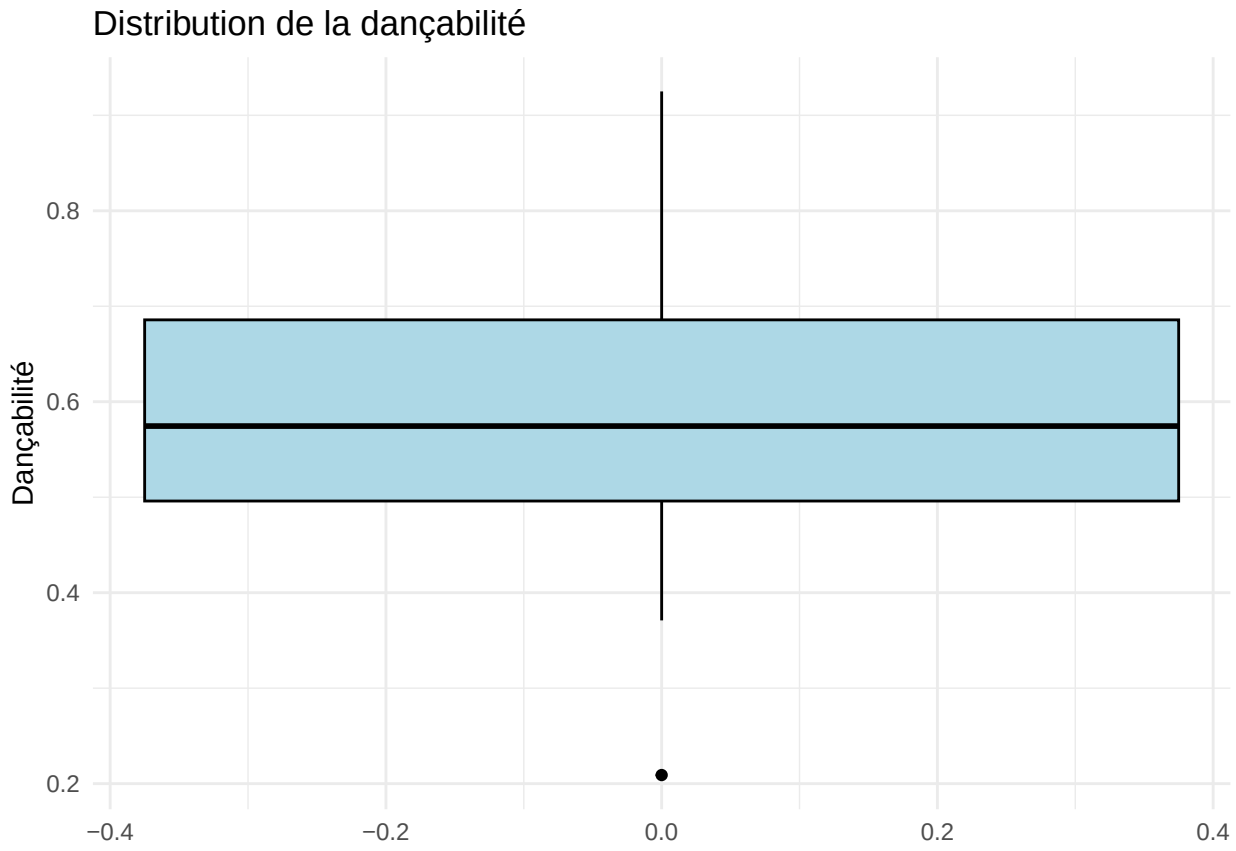
Cela représente la probabilité de dancer sur une musique

```r
data$Danceability
```

```
##  [1] 0.818 0.676 0.695 0.689 0.663 0.760 0.716 0.726 0.741 0.625 0.592 0.559
## [13] 0.618 0.595 0.458 0.427 0.556 0.451 0.666 0.700 0.902 0.614 0.489 0.700
## [25] 0.653 0.925 0.576 0.853 0.892 0.545 0.566 0.547 0.539 0.511 0.531 0.438
## [37] 0.512 0.386 0.624 0.425 0.429 0.486 0.371 0.617 0.557 0.545 0.449 0.491
## [49] 0.573 0.209
```

```r
# Créer un boxplot de la dançabilité
boxplot <- ggplot(data, aes(y = data$Danceability)) +
  geom_boxplot(fill = "lightblue", color = "black") +
  labs(title = "Distribution de la dançabilité", y = "Dançabilité") +
  theme_minimal()

# Afficher le boxplot
print(boxplot)
```

```
## Warning: Use of ‘data$Danceability‘ is discouraged.
## i Use ‘Danceability‘ instead.
```



Les variables quantitatives sont: - La dancabilité d'une musique, qui est un indicateur sur la probabilité de dancer sur une musique - l'énergie que la musique à definir!

```
#var(duration)  # Variance
#mean(v)    # Moyenne
```