

# **Analyse descriptive du jeu de données Spotify**

## **Projet en Statistique descriptive**

### **Membres**

LOULIDI Younes

PHAM Tuan Kiet

VO Van Nghia

### **Date**

17 Mars, 2021

# Table des matières

Table des matières	i
<b>1 Statistiques descriptives unidimensionnelle et bidimensionnelle</b>	<b>1</b>
1.1 La nature des jeux de données . . . . .	1
1.1.1 Des jeux de données . . . . .	1
1.1.2 Des variables statistiques . . . . .	1
1.1.3 Charger les jeux de données dans R . . . . .	2
1.2 Analyses unidimensionnelles . . . . .	3
1.2.1 Une variable qualitative - <code>pop.class</code> . . . . .	3

# 1 Statistiques descriptives unidimensionnelle et bidimensionnelle

## 1.1 La nature des jeux de données

### 1.1.1 Des jeux de données

Ces jeux de données se composent de 10000 chansons extraites de la base de données Spotify.

Chaque ligne contient 11 variables statistiques comme suit:

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

### 1.1.2 Des variables statistiques

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

Nom de variable statistique	Type de variable	Format dans R
<code>year</code>	qualitative ordinale	<code>integer</code>
<code>acousticness</code>	quantitative continue	<code>numeric</code>
<code>duration</code>	quantitative continue <sup>1</sup>	<code>numeric</code>

Nom de variable statistique	Type de variable	Format dans R
energy	quantitative continue	numeric
explicit	qualitative nominale	logical
key	qualitative nominale	factor
liveness	quantitative continue	numeric
loudness	quantitative continue	numeric
mode	qualitative nominale <sup>2</sup>	logical
tempo	quantitative continue	numeric
pop.class	qualitative nominale	factor

### 1.1.3 Charger les jeux de données dans R

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

<sup>1</sup>On choisi son nature est de quantitative continue parce que.

<sup>2</sup>On pose FALSE si la tonalité est mineure et TRUE si non.

## 1.2 Analyses unidimensionnelles

### 1.2.1 Une variable qualitative - pop.class

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

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

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

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

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

```
pie(pop_class_table, col = rainbow(length(pop_class_table)))
```

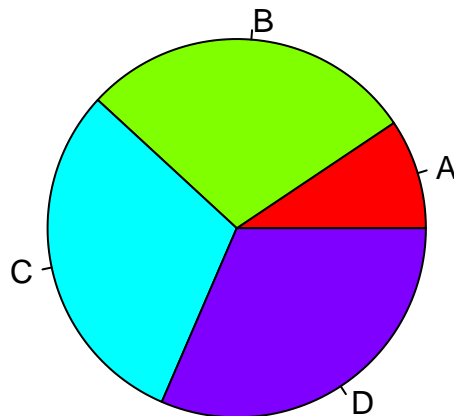


Figure 1: Diagramme circulaire de popularité

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