Introduction aux traitement de données

Camille Ansel Elsa Catteau Anas EL Farsi

PROBLÉMATIQUE

Quels sont les facteurs qui ont une influencent sur la popularité d’une musique ?

NOTRE DATASET

Identifiants

- instance_id (Entier)
- artist_name (Texte)
- track_name (Texte)
- music_genre (Texte)
- mode (Texte)
- key (Texte)

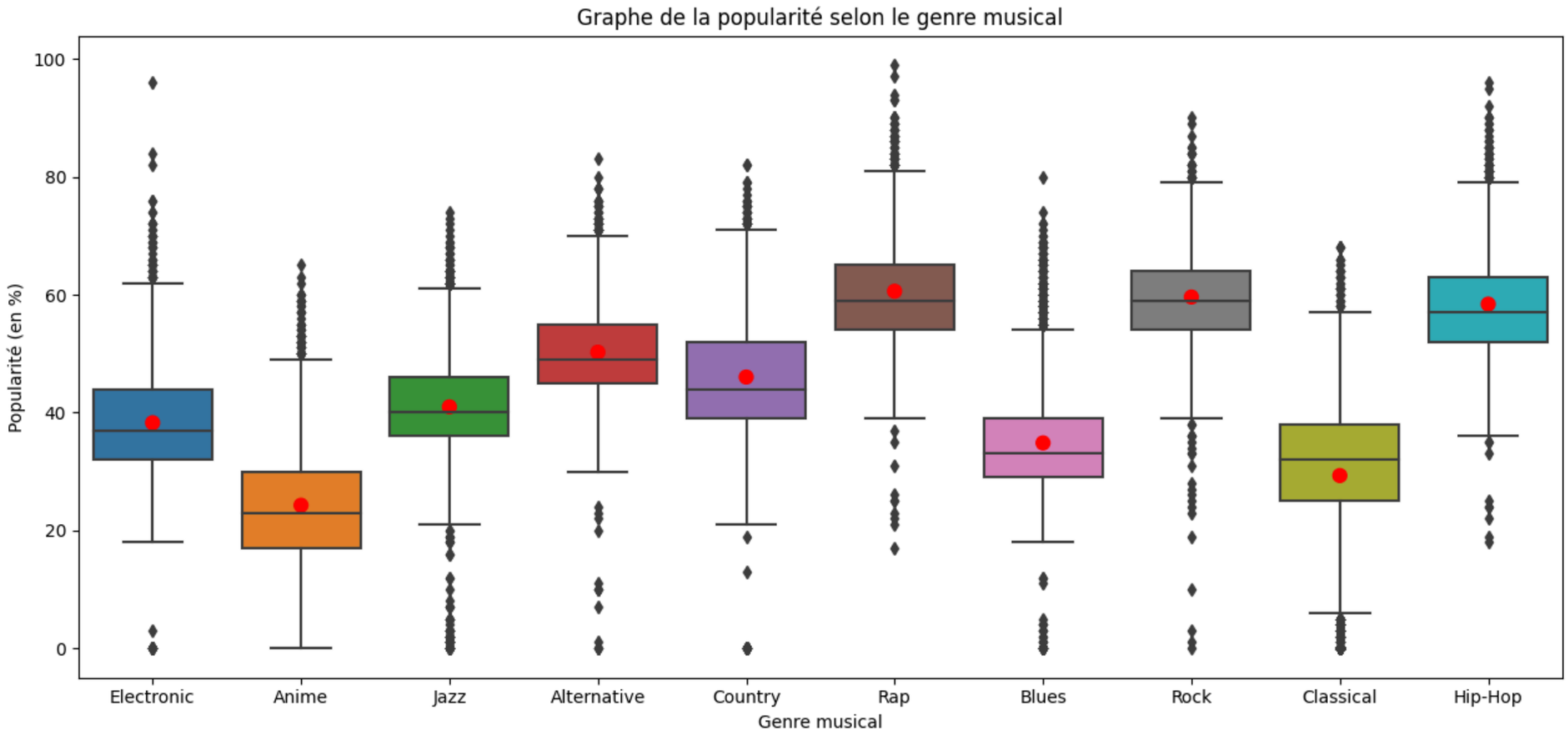
Caractéristiques audio

- popularity (0–100)
- acousticness (0–1)
- danceability (0–1)
- energy (0–1)
- liveness (0–1)
- instrumentalness (0–1)
- duration_ms (En ms)
- tempo (BPM ou “?”)
- loudness (dB)
- valence (0–1)
- speechiness (0–1)

DATACLEANING

- Suppression des variables inutiles
- Suppression des lignes où des informations sont manquantes
- Conversion de la durée en secondes
- Nettoyage des titres
- Encodage des variables catégorielles
- Suppression des doublons

INFLUENCE DES FEATURES INITIALES



POPULARITÉ ET TITRE

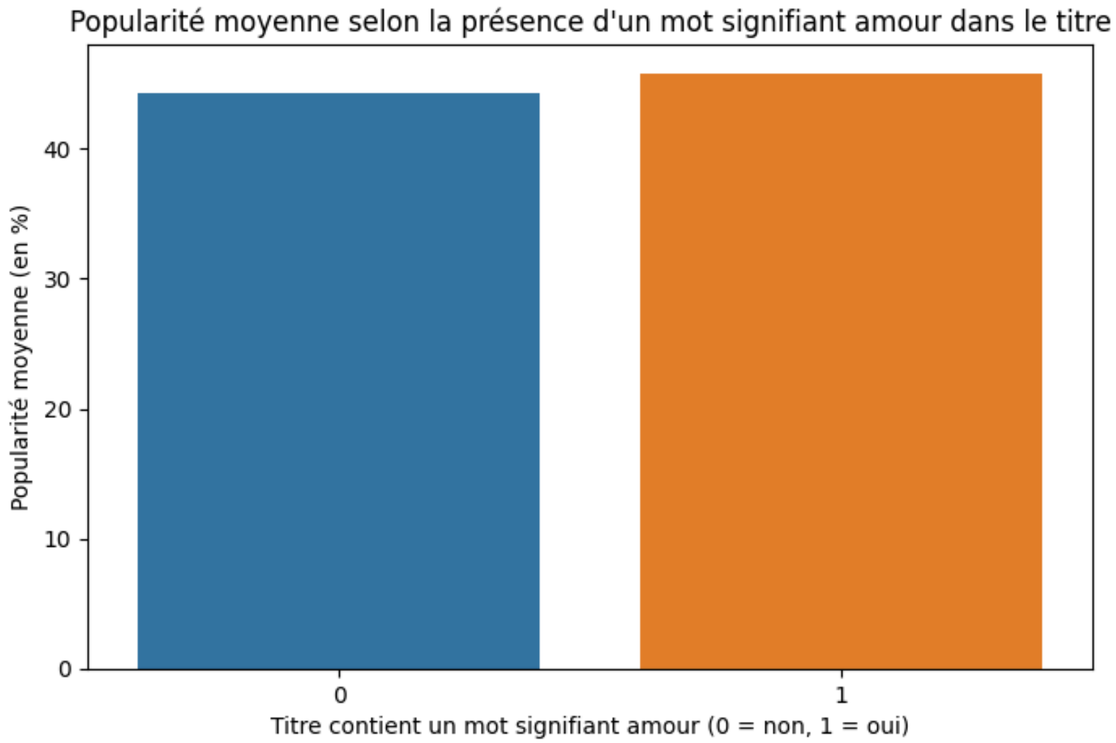
Titre → **Combinaison de mots**

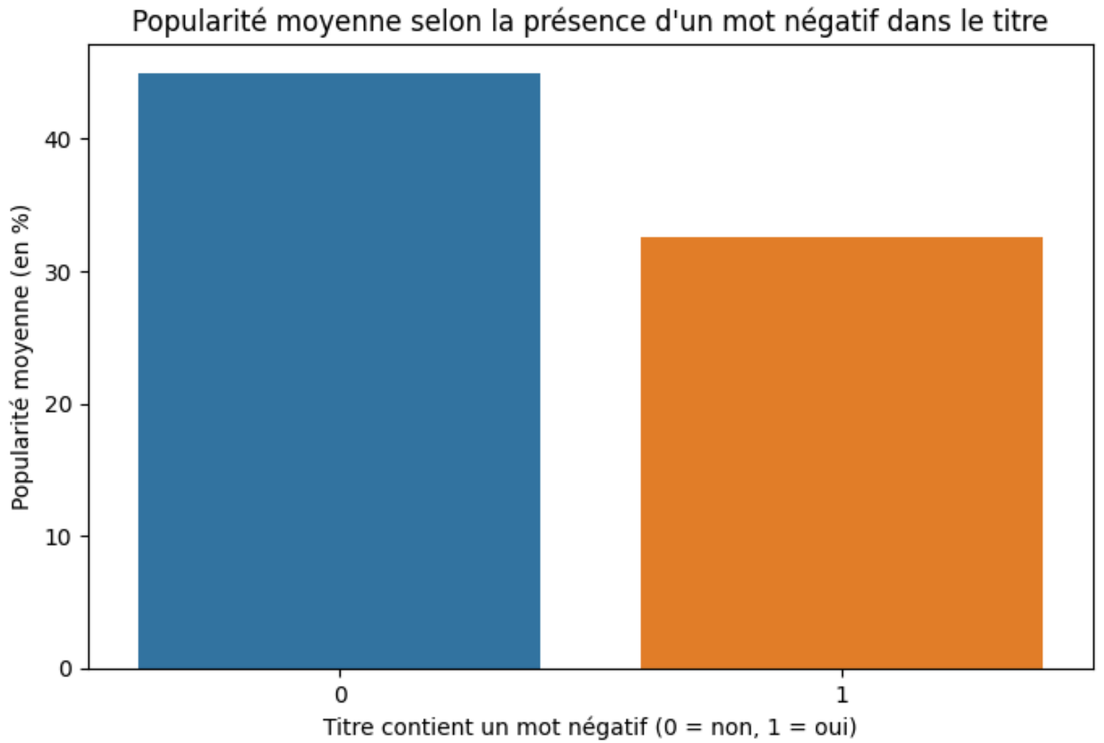
Mots les plus courants :

('no', 2035), ('love', 1054), ('all', 461), ("don't", 436), ...

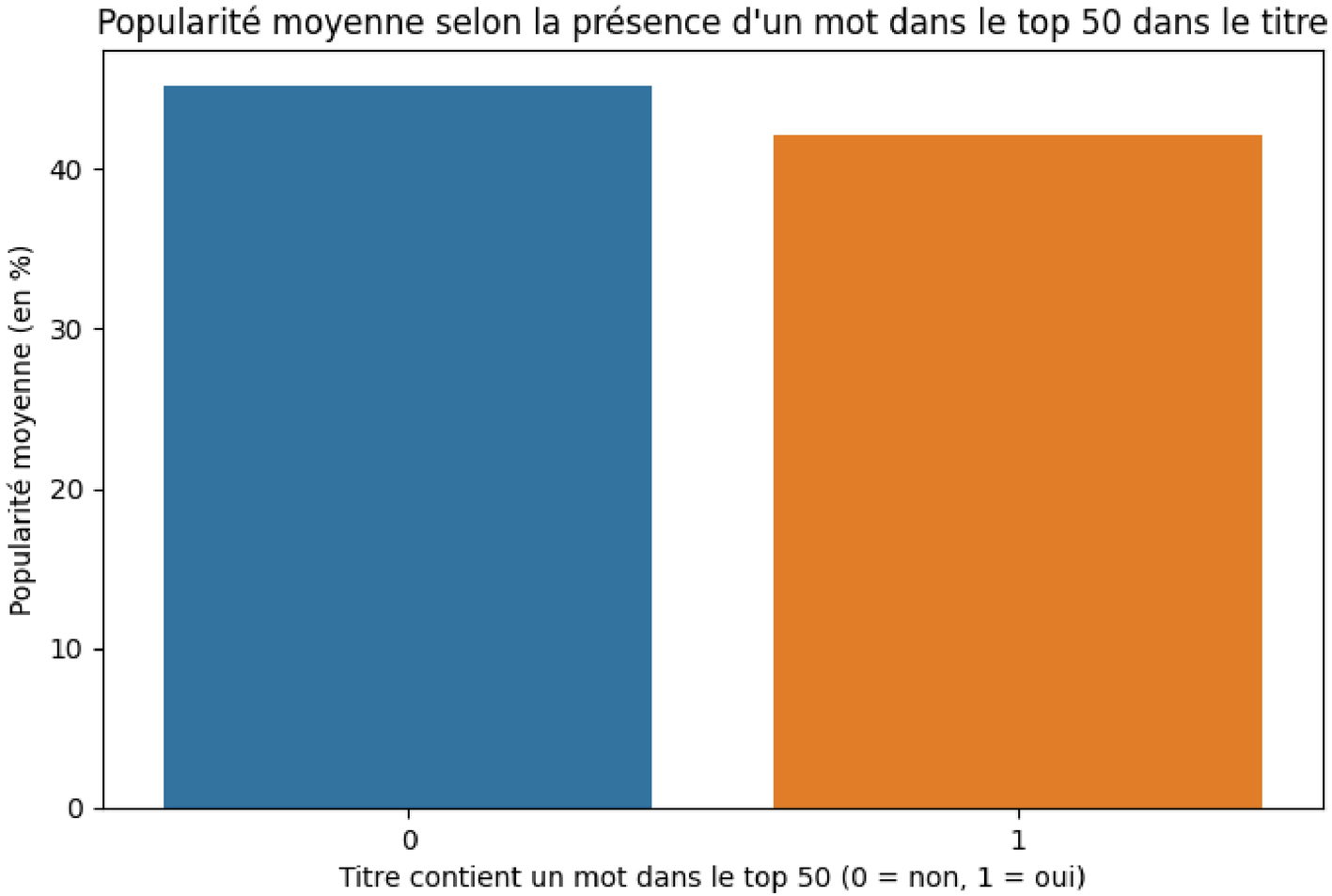
```
Pourcentage de musiques avec has_no = 1 : 5.17%
Pourcentage de musiques avec has_love = 1 : 2.69%
Pourcentage de musiques avec has_top = 1 : 28.70%
```

POPULARITÉ ET TITRE

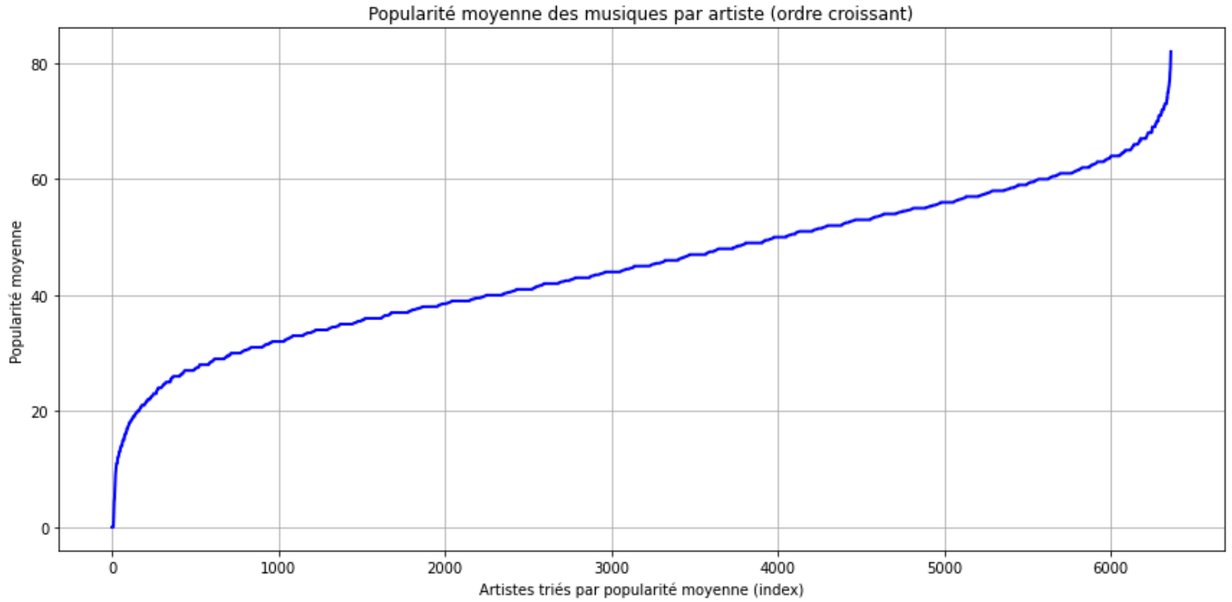


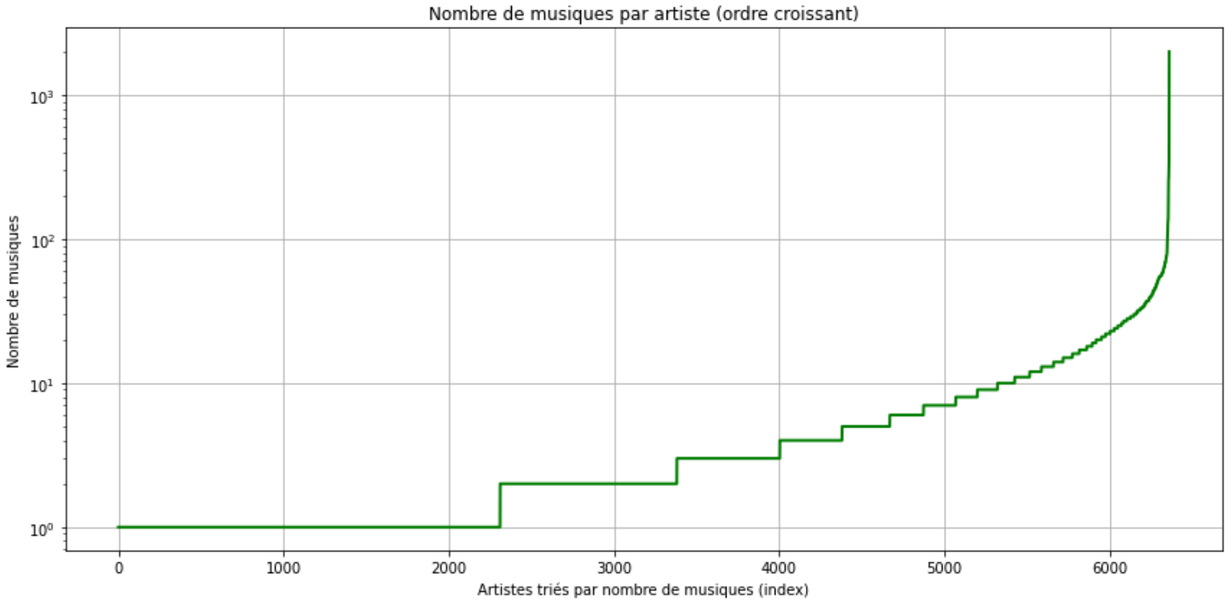


POPULARITÉ ET TITRE

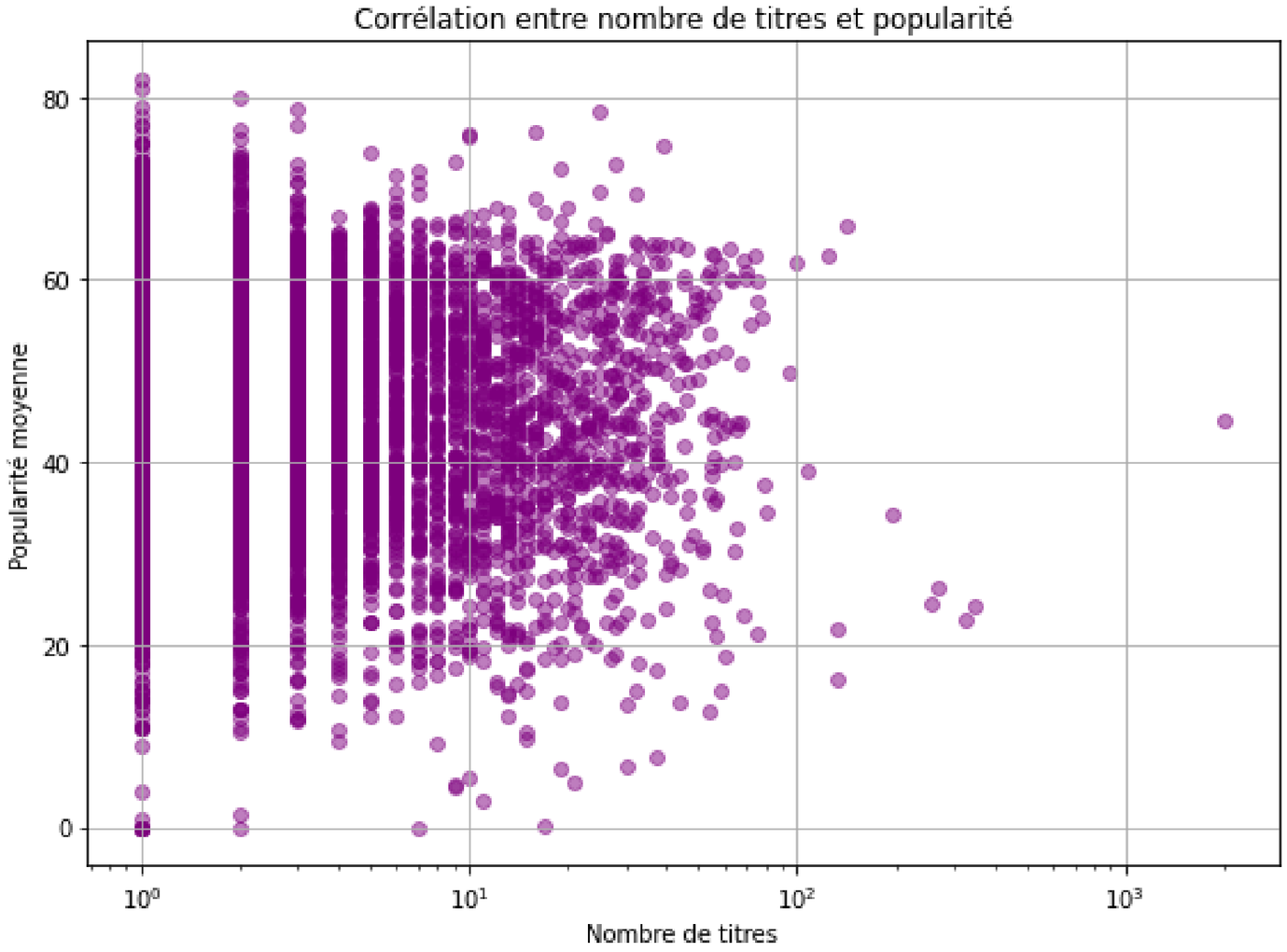


POPULARITE ET ARTISTE

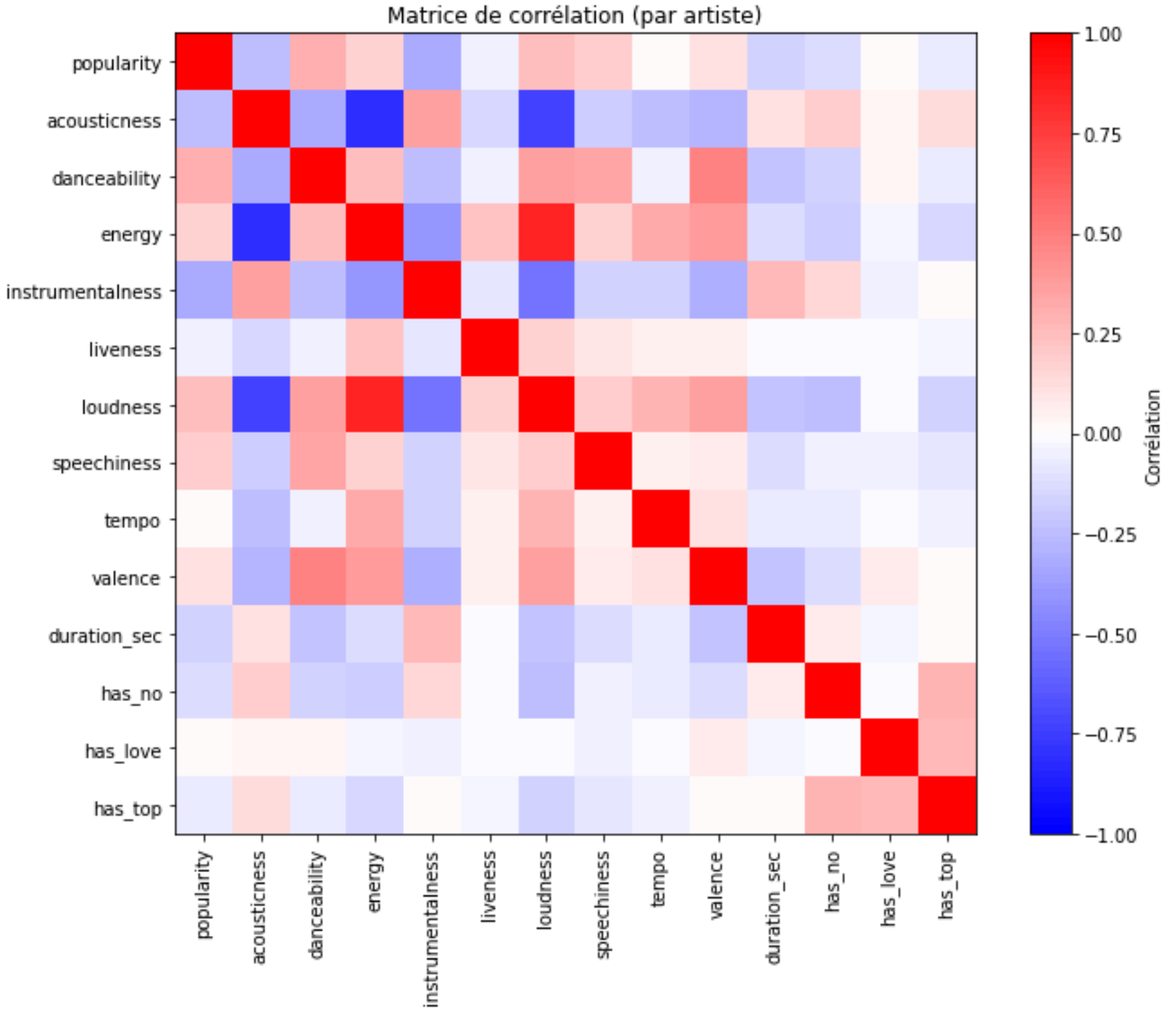


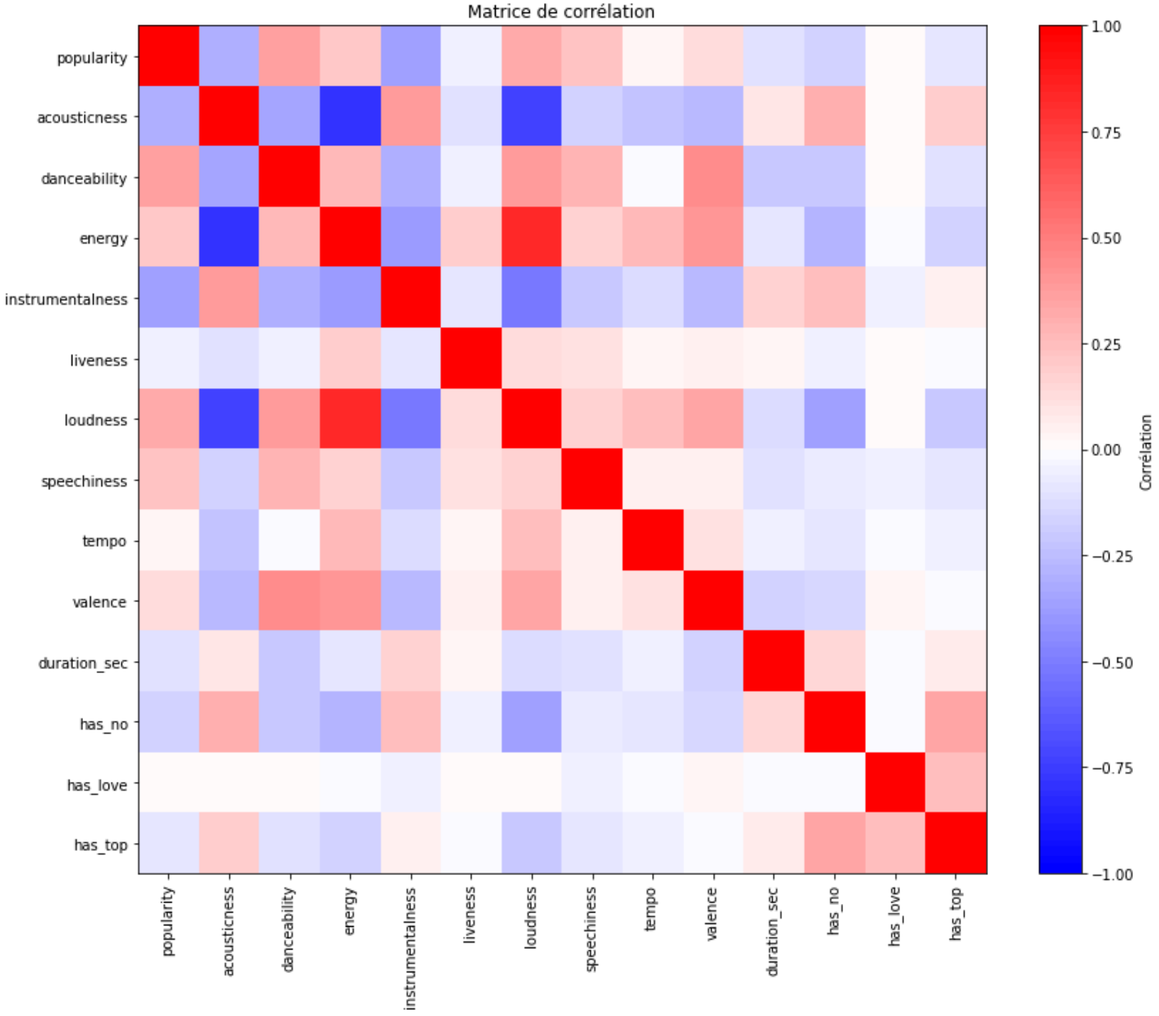


POPULARITE ET ARTISTE



POPULARITE ET ARTISTE





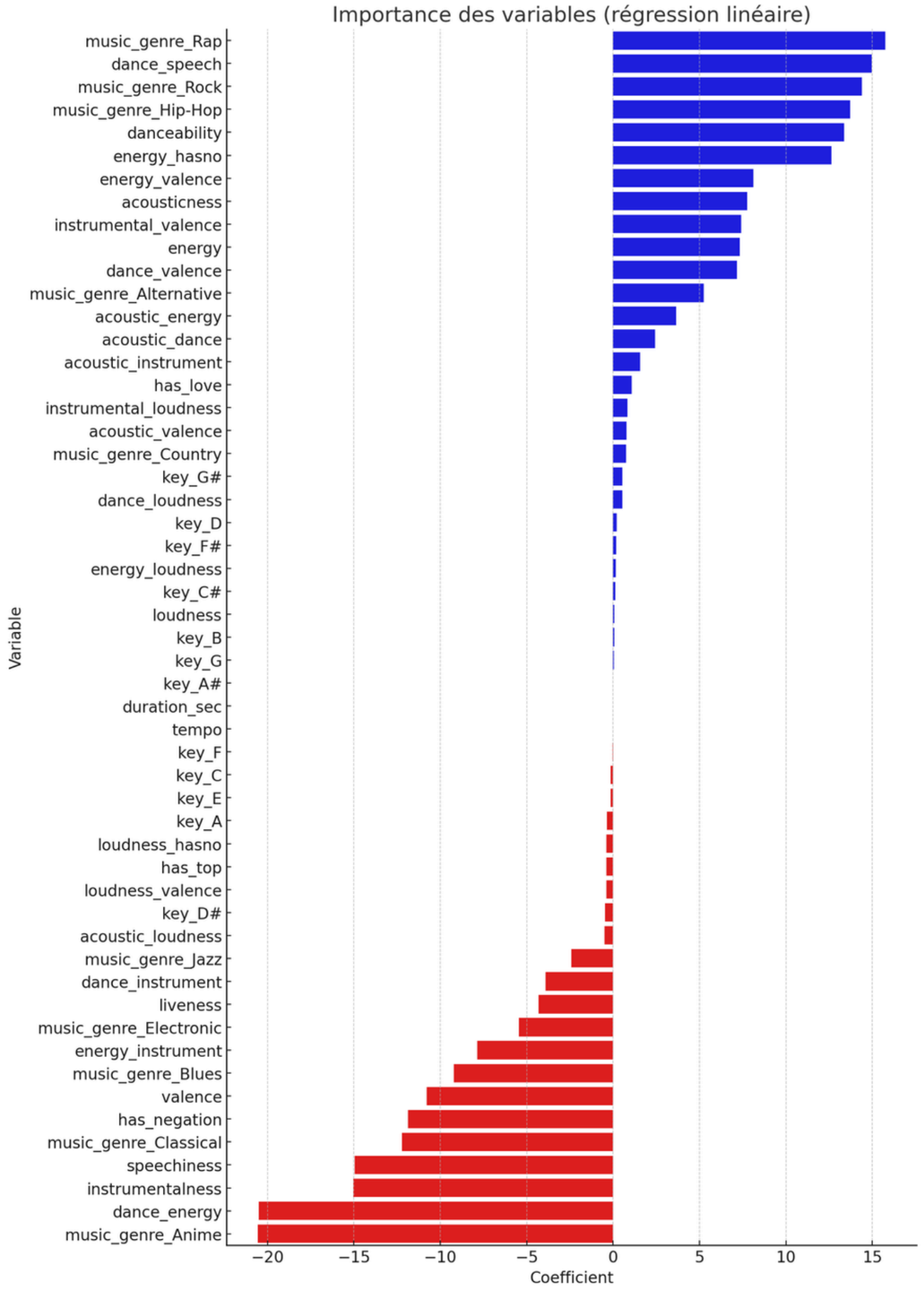
RÉGRESSION LINÉAIRE

Régression linéaire multiple :

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

On a trouvé un $R^2 = 0.63 < 1$

Avec conversion de key et music_genre en one hot


Importance des variables (régression linéaire)
music_genre_Rap
dance_speech
music_genre_Rock
music_genre_Hip-Hop
danceability
energy_hasno
energy_valence
acousticness
instrumental_valence
energy
dance_valence
music_genre_Alternative
acoustic_energy
acoustic_dance
acoustic_instrument
has_love
instrumental_loudness
acoustic_valence
music_genre_Country
key_G#
dance_loudness
key_D
key_F#
energy_loudness
key_C#
loudness
key_B
key_G
key_A#
duration_sec
tempo
key_F
key_C
key_E
key_A
loudness_hasno
has_top
loudness_valence
key_D#
acoustic_loudness
music_genre_Jazz
dance_instrument
liveness
music_genre_Electronic
energy_instrument
music_genre_Blues
valence
has_negation
music_genre_Classical
speechiness
instrumentalness
dance_energy
music_genre_Anime
Variable
−20
−15
−10
−5
0
5
10
15
Coefficient

RÉGRESSION LOGISTIQUE

Matrice de confusion :

6768	532
1195	3673

	precision	recall	f1-score	support
0	0.85	0.93	0.89	7300
1	0.87	0.75	0.81	4868
accuracy			0.86	12168
macro avg	0.86	0.84	0.85	12168
weighted avg	0.86	0.86	0.86	12168

CONCLUSION

- Amélioration de la régression possible en ayant plus de variables
- Méthodes alternatives (K Nearest Neighbors)

MERCI POUR VOTRE ATTENTION !