SOUTENANCE DE PROJET

INTRODUCTION AU TRAITEMENT DE DONNÉES

FERRE Gabriel
PATEL Lucah
PROUZET Charlotte

MAM3

I. BUSINESS PROBLEM AND DATA

“SPOTIFY TRACKS DATASET”

- Données de titres de Spotify
- 125 genres différents
- Différentes caractéristiques audios
- Format CSV

UTILITÉ DE NOTRE DATASET ?

- Diversité des features
- lien entre caractéristique musicale et succès d’un son
- Large nombre de pistes (114 000)

> analyse de la structure musicale possible

PROBLÉMATIQUE :

Sur quels critères est ce que les recommandations Spotify se basent-elles pour nous suggérer un nouveau son ?

II. DATA DESCRIPTION

> Structure générale :
- 21 colonnes (features)
- 114 000 lignes (titres)

> 4 Types de données :
- Variables numériques continues : float
- Variables numériques discrètes : int
- Variables catégorielles (texte) : object
- Variables booléenne : bool

> Les variables numériques continues : float

Features	Danceability	Energy	Loudness	Speechiness	Accousticness	Instrumentalness	Liveness	Valence	Tempo
Intervalle	[0 , 1]	[0 , 1]	[-49.5 , 4.53]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 1]	[0 , 243]

> Les variables numériques discrètes : int

Features	Unamed	Popularity	Duration_ms	Key	Mode	Time_signature
Intervalle	[0 , 114 000]	[0 , 100]	[0 , 5 400 000]	[0 , 11]	[0 , 1]	[0 , 5]

> Les variables catégorielles

Features	Track_id	Artists	Album_name	Track_name	Time_genre
Nb de valeurs uniques	89 741	31 438	46 590	73 609	114

> Les variables booléennes

Features	Explicit
Valeurs possibles	true, false

> Les variables numériques discrètes : int

Features	Unamed	Popularity	Duration_ms	Key	Mode	Time_siganture
Intervalle	[0 , 114 000]	[0 , 100]	[0 , 5 400 000]	[0 , 11]	[0 , 1]	[0 , 5]

> Les variables catégorielles

Features	Track_id	Artists	Album_name	Track_name	Time_genre
Nb de valeurs uniques	89 741	31 438	46 590	73 609	114

Les variables manquantes

> Les variables booléennes

Features	Explicit
Valeurs possibles	true, false

> Les variables numériques discrètes : int

Features	Unamed	Popularity	Duration_ms	Key	Mode	Time_siganture
Intervalle	[0 , 114 000]	[0 , 100]	[0 , 5 400 000]	[0 , 11]	[0 , 1]	[0 , 5]

> Les variables catégorielles

Features	Track_id	Artists	Album_name	Track_name	Time_genre
Nb de valeurs uniques	89 741	31 438	46 590	73 609	114

Les variables manquantes

Les variables supprimées

> Les variables booléennes

Features	Explicit
Valeurs possibles	true, false

> Les variables numériques discrètes : int

Features	Unamed	Popularity	Duration_ms	Key	Mode	Time_siganture
Intervalle	[0 , 114 000]	[0 , 100]	[0 , 5 400 000]	[0 , 11]	[0 , 1]	[0 , 5]

> Les variables catégorielles

Features	Track_id	Artists	Album_name	Track_name	Time_genre
Nb de valeurs uniques	89 741	31 438	46 590	73 609	114

Data cleaning { ***Les variables manquantes*** / ***Les variables supprimées*** }

+ Normalisation sur les variables numériques

> Les variables booléennes

Features	Explicit
Valeurs possibles	true, false

III. DATA STORYTELLING

NB ARTISTES :

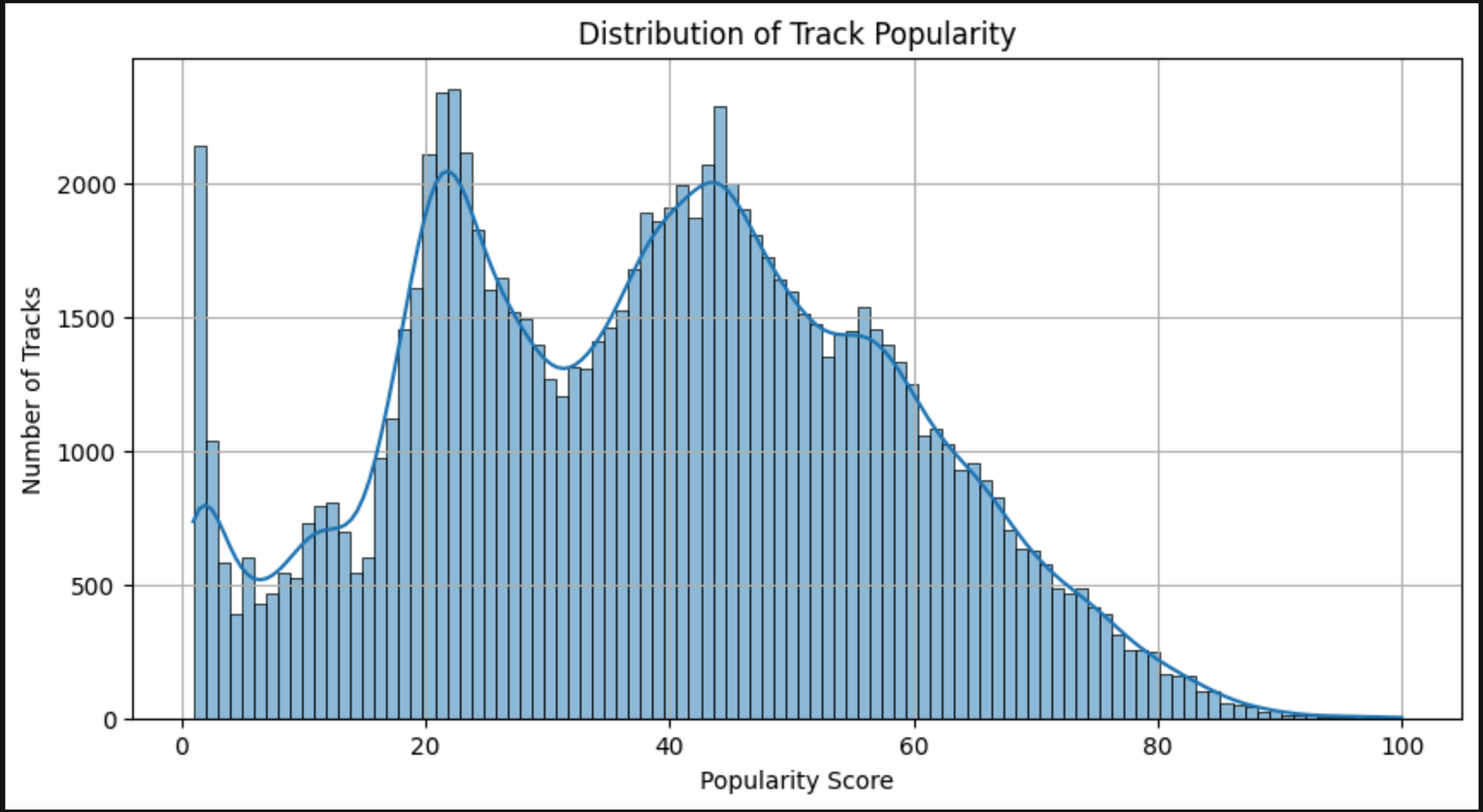
31 438

NB SONS :

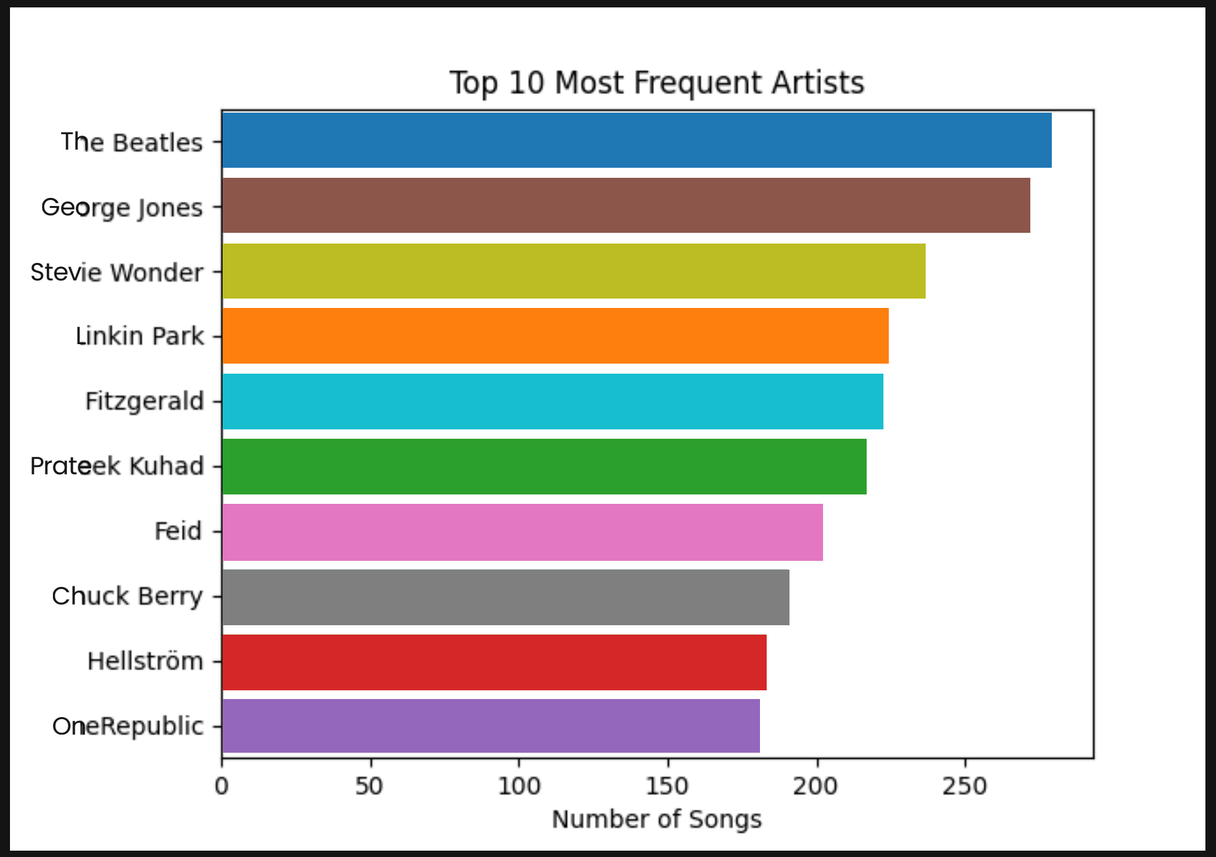
89 741

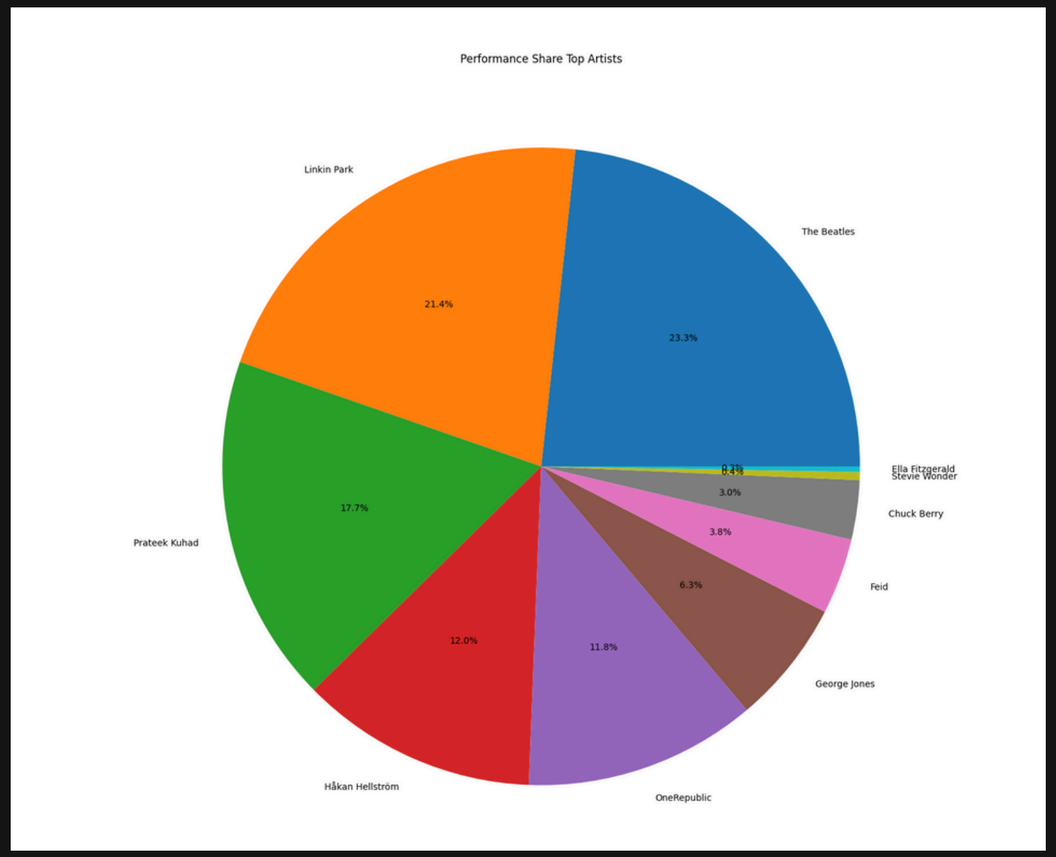
NB SUCCÈS :

98



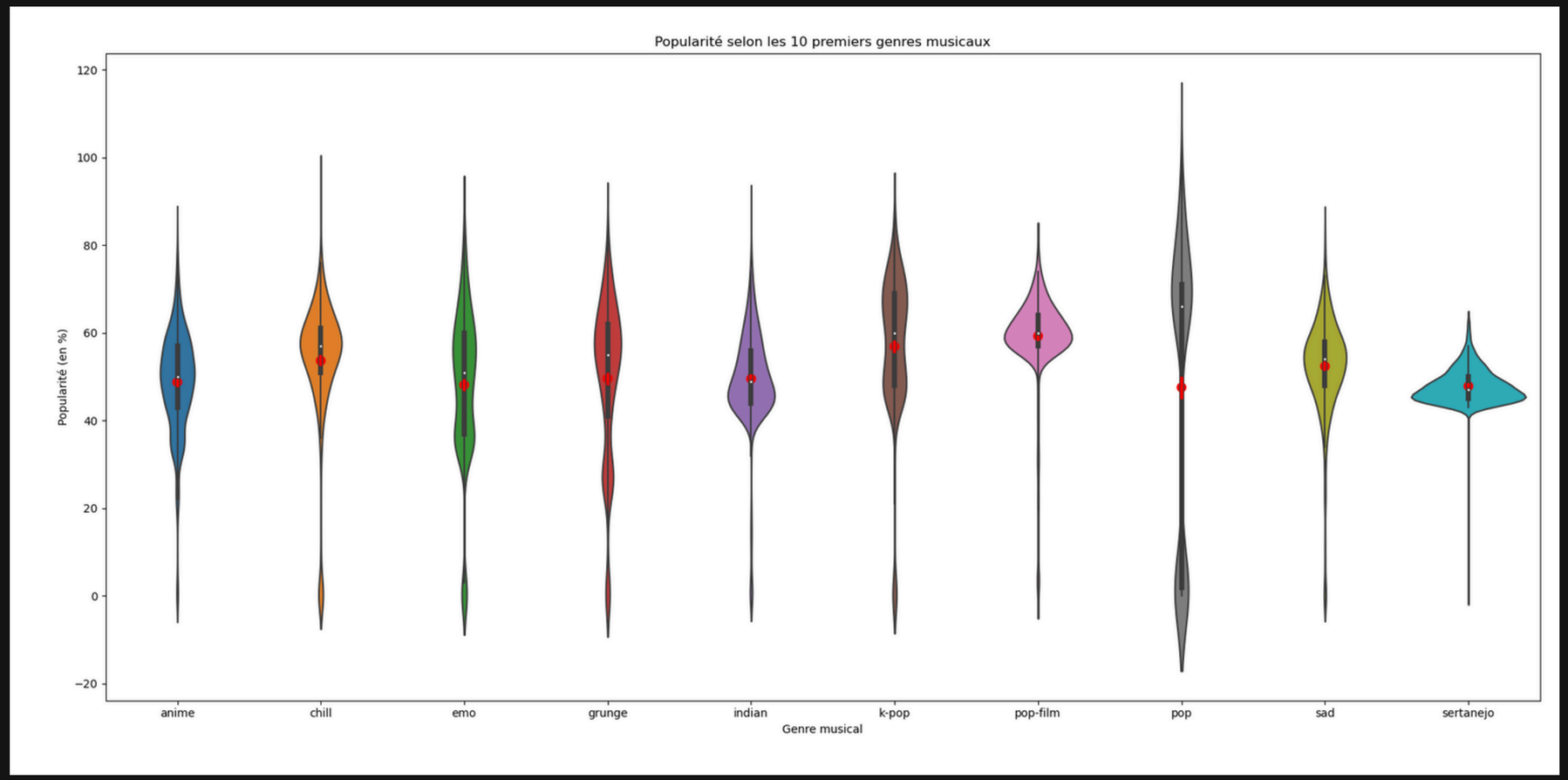
1. QUELS SONT LES FACTEURS QUI INFLUENCENT LA POPULARITÉ D'UNE MUSIQUE ?





> Nombre de sons élevé => forte popularité ?

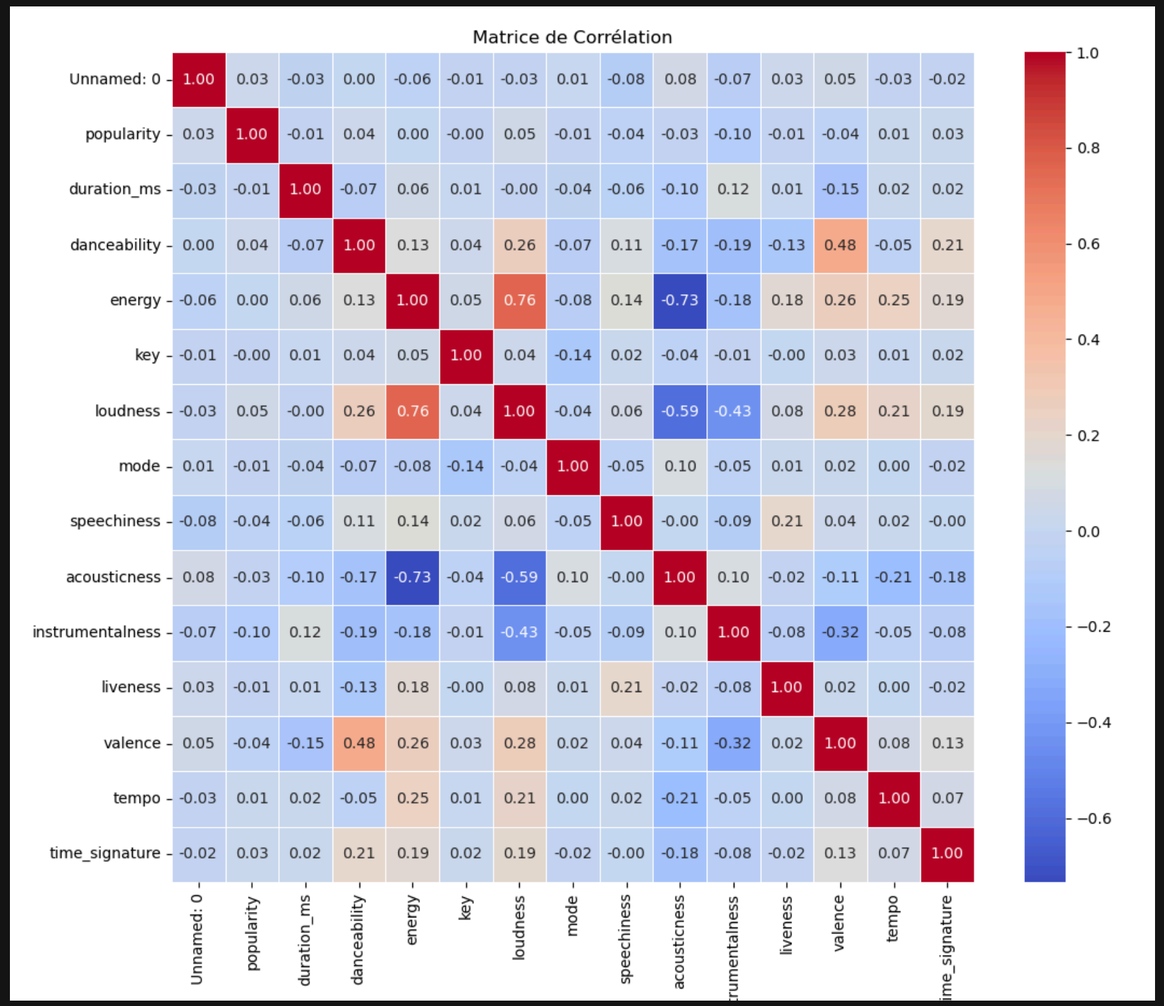
2. QUELS SONT LES LIENS ENTRE LES DIFFRENTS GENRES ?



> Peut-on les regrouper ? Est ce que l'on peut trouver des caractéristiques communes qu'ils partagent entre eux ?

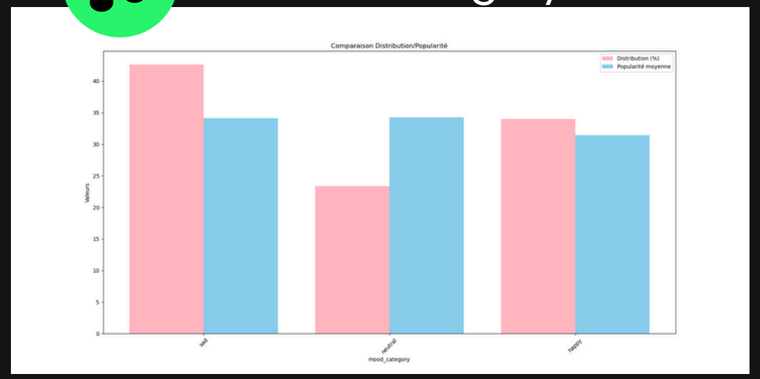
IV. HANDCRAFTED FEATURES

> Étude de la matrice de corrélation

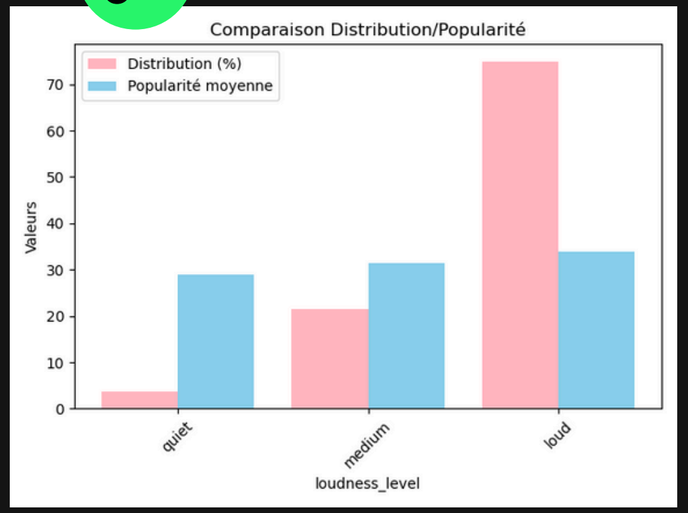


> Détermination des heandcrafted features

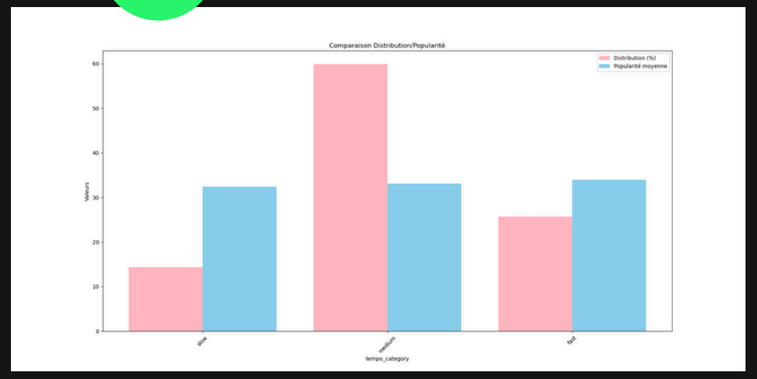
Mood_category



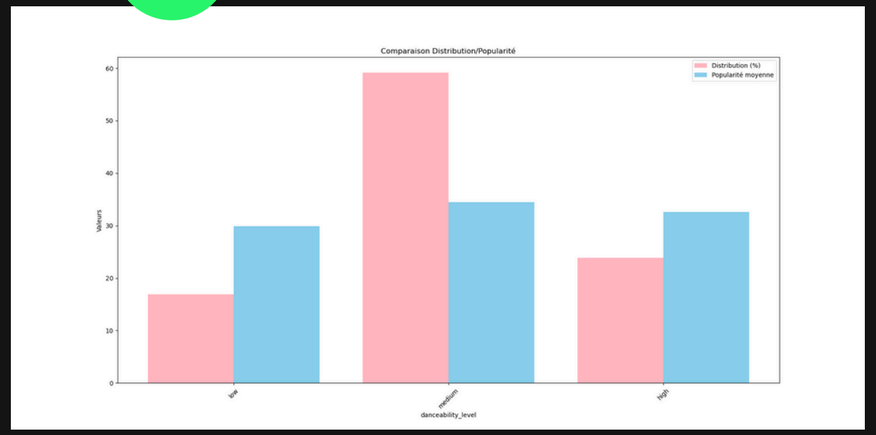
Loudness_level



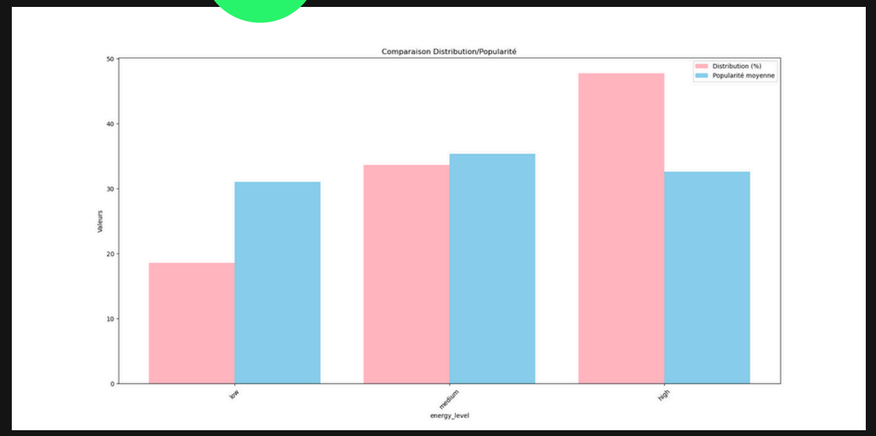
Tempo_category



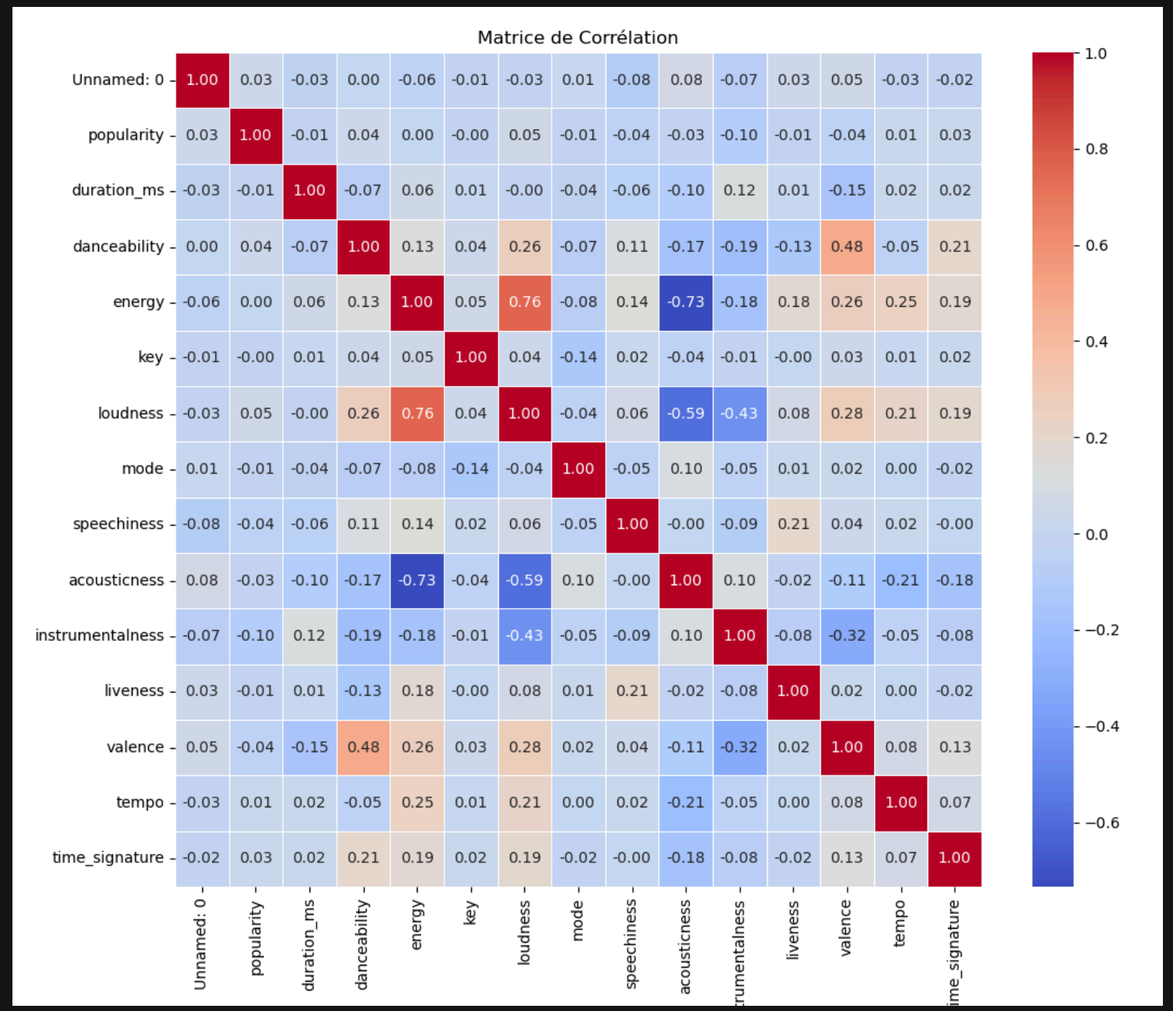
Danceability_level



Energy_level

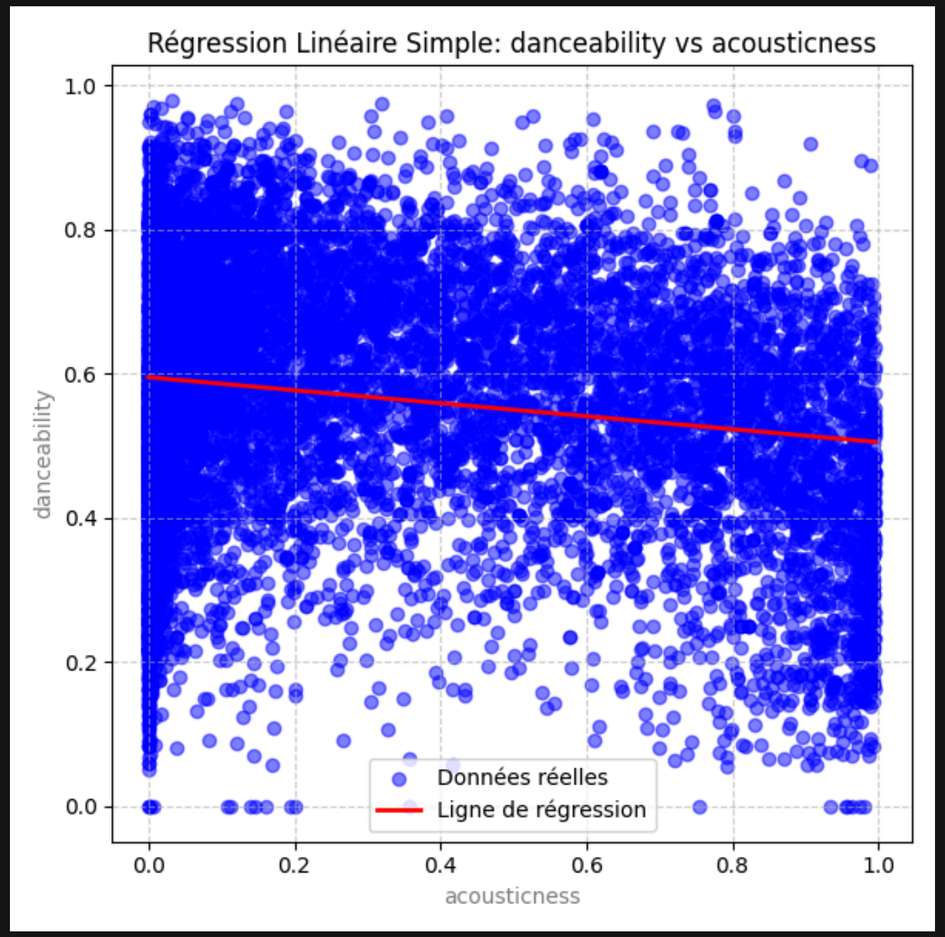


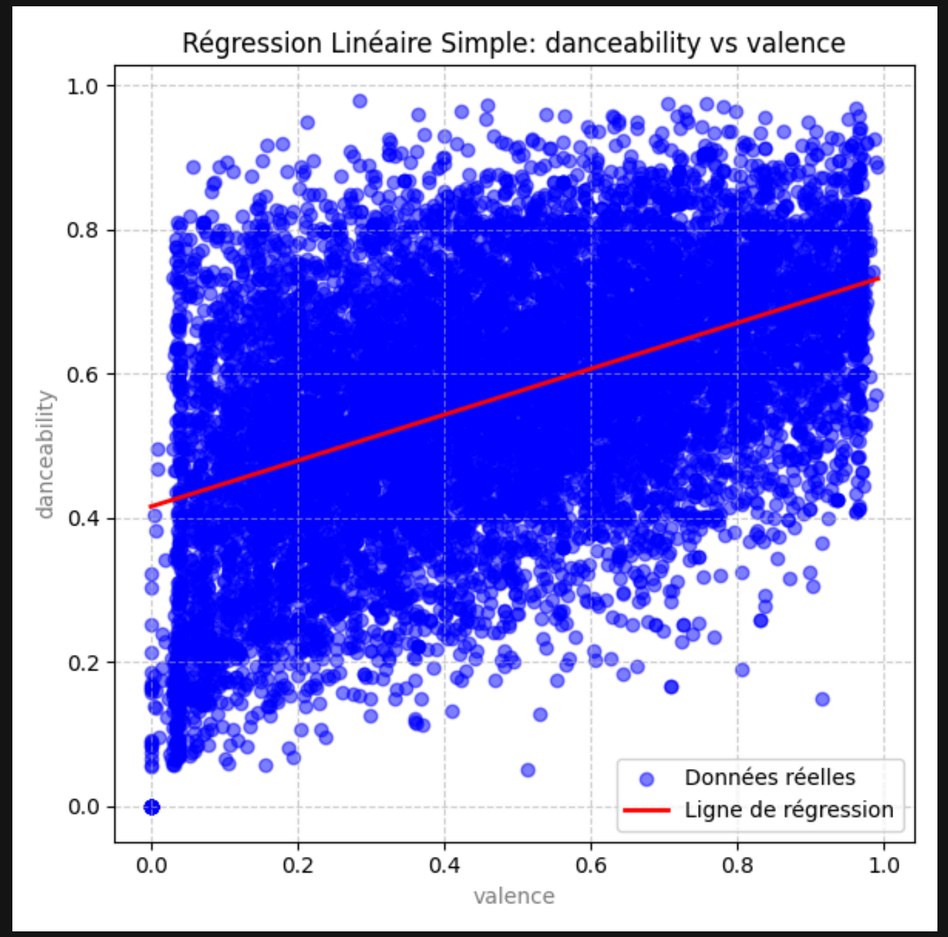
V. RÉGRESSION LINÉAIRE



1. Régressions linéaires

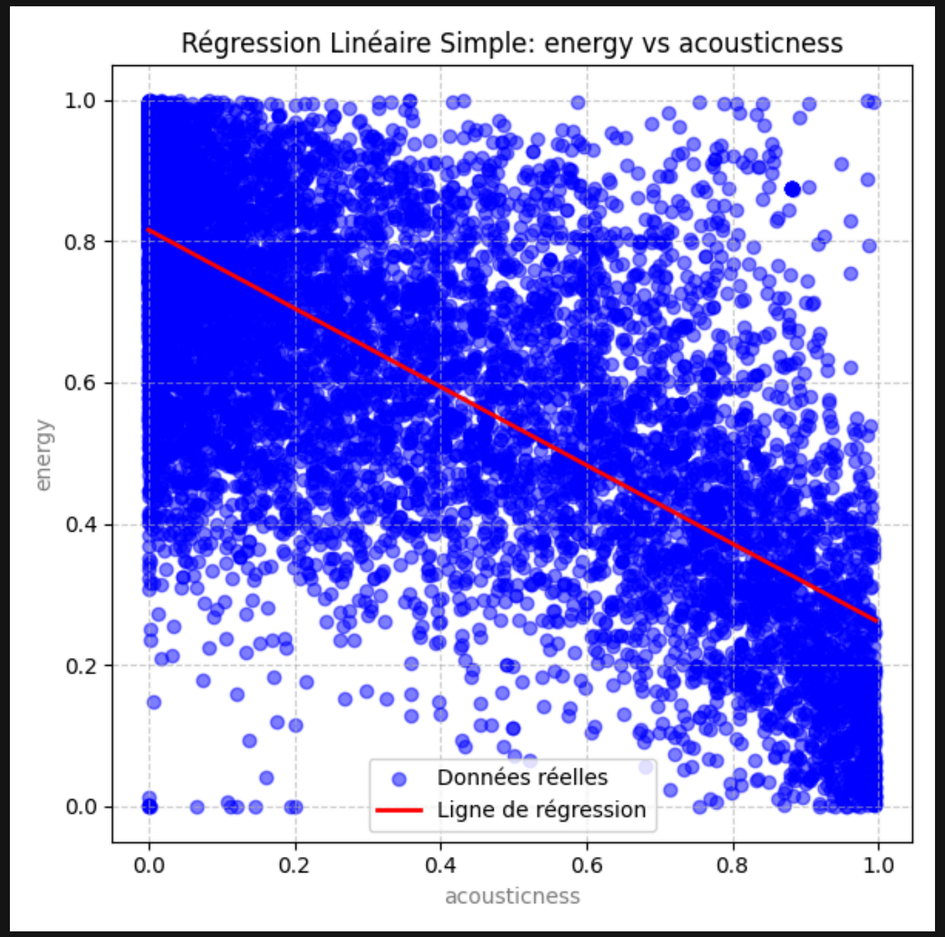
DANCEABILITY

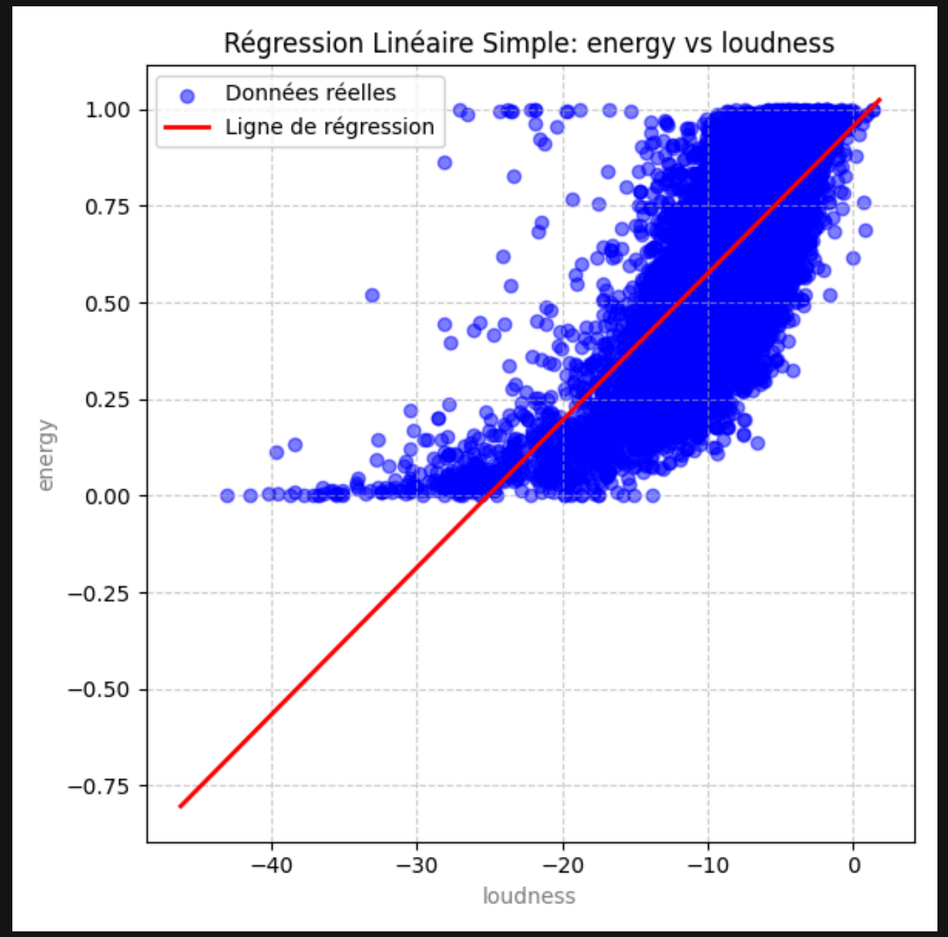




1. Régressions linéaires

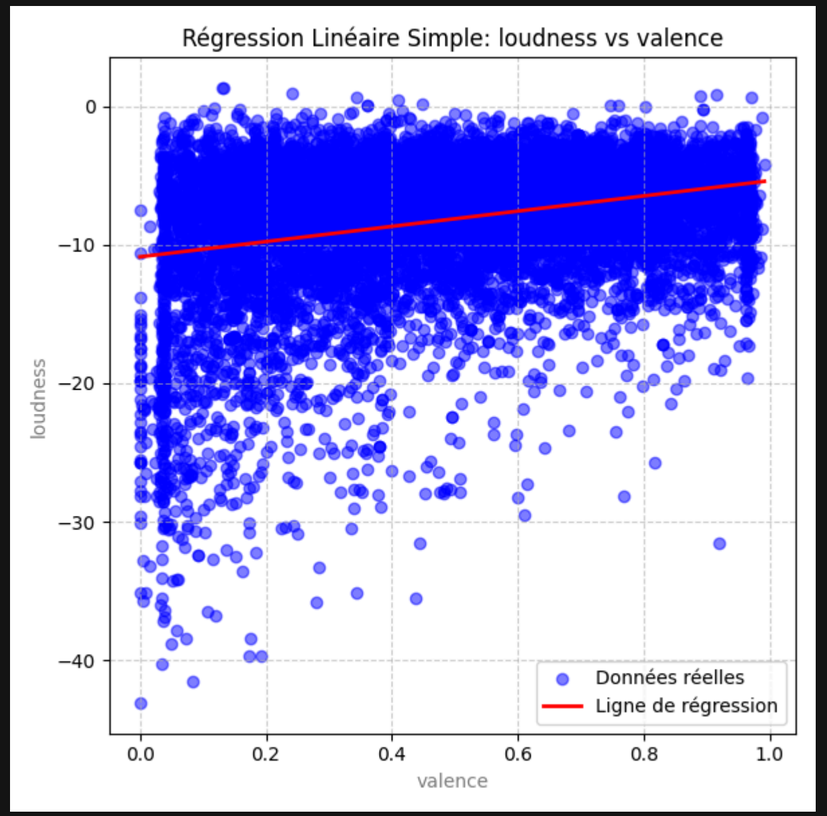
ENERGY

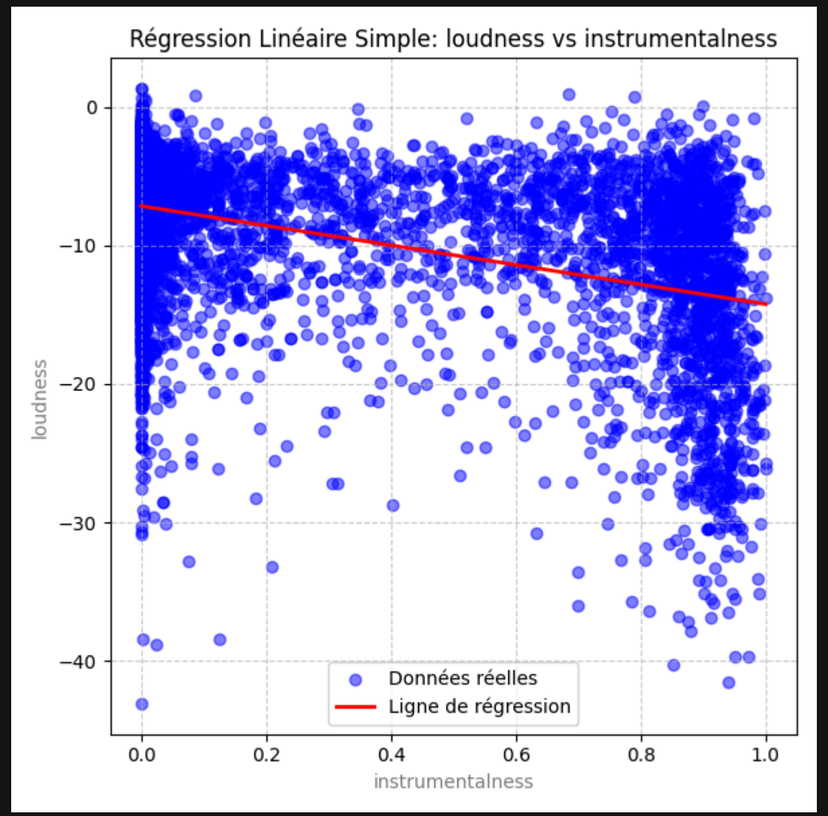


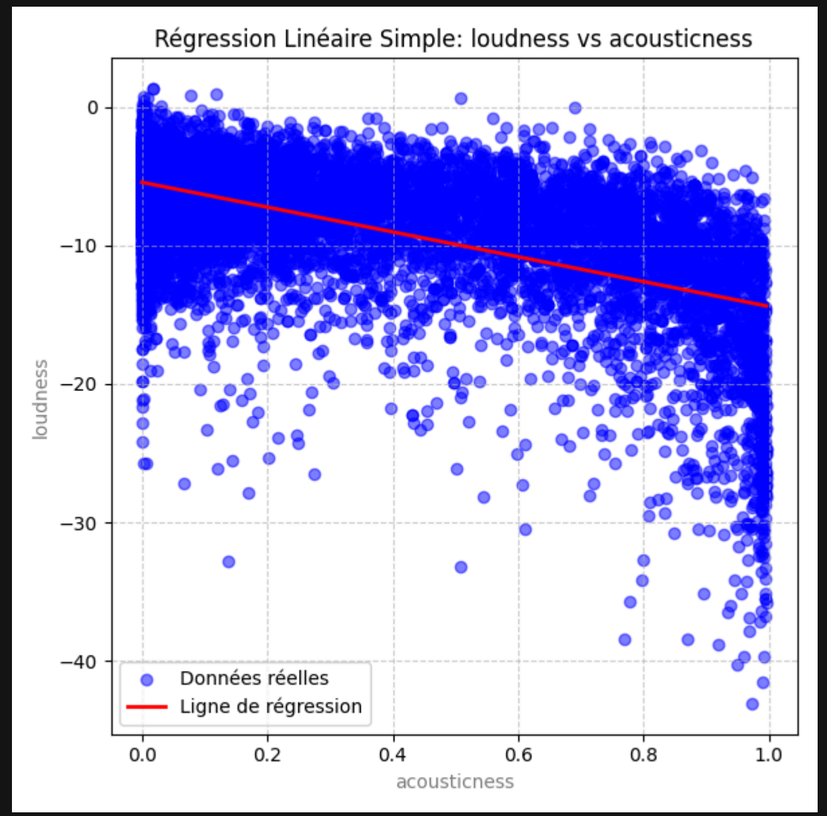


1. Régressions linéaires

LOUNDNESS

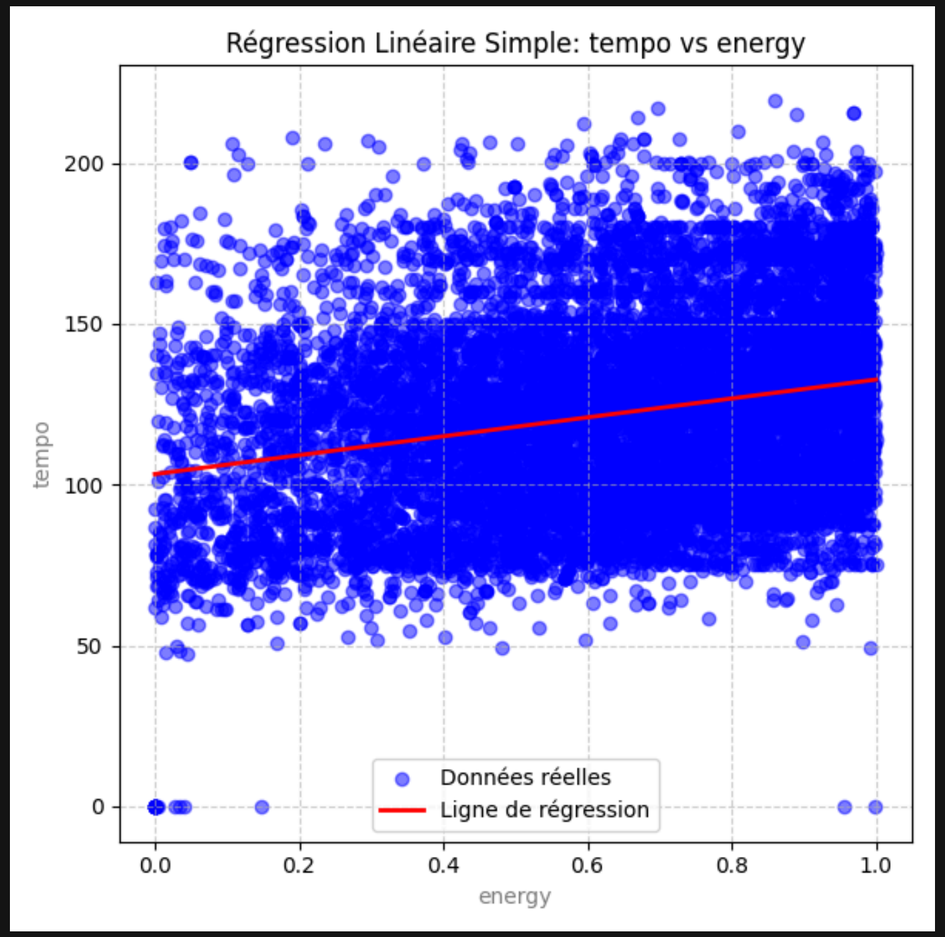


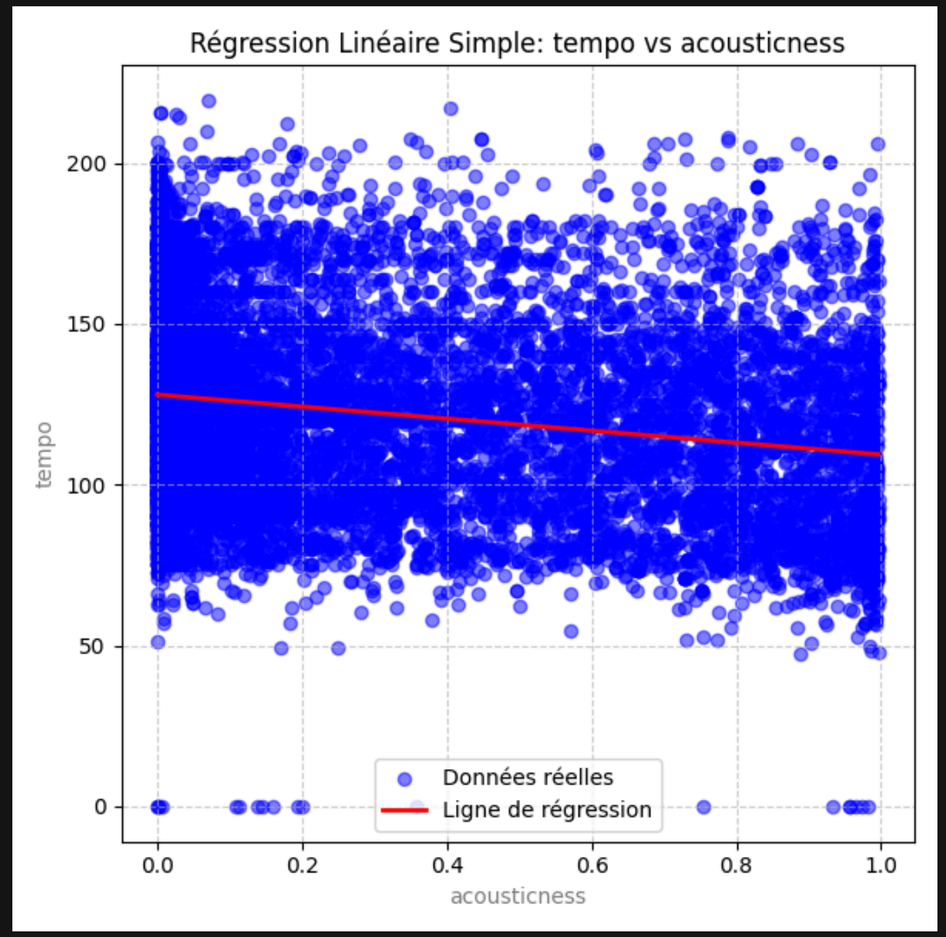




1. Régressions linéaires

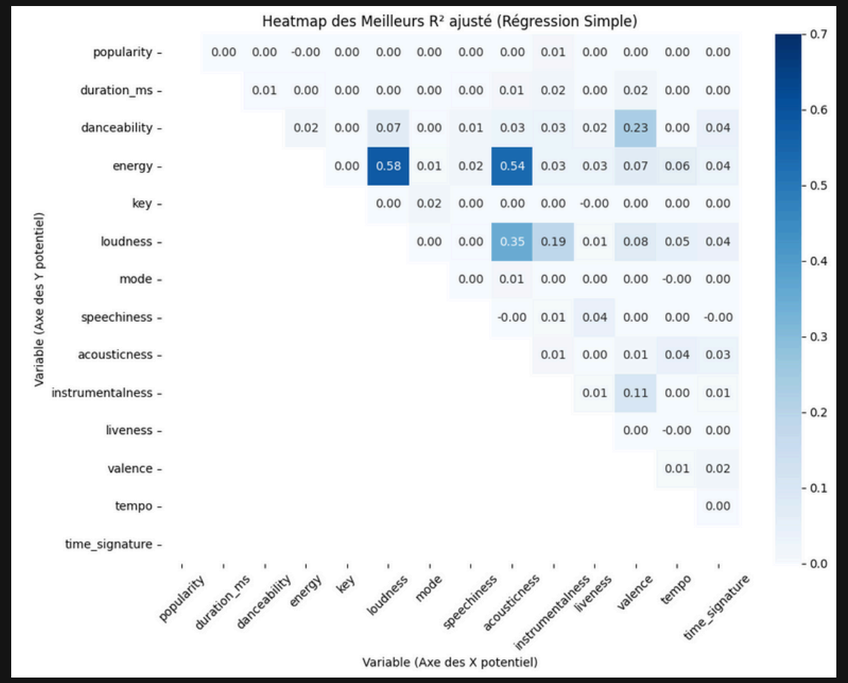
TEMPO



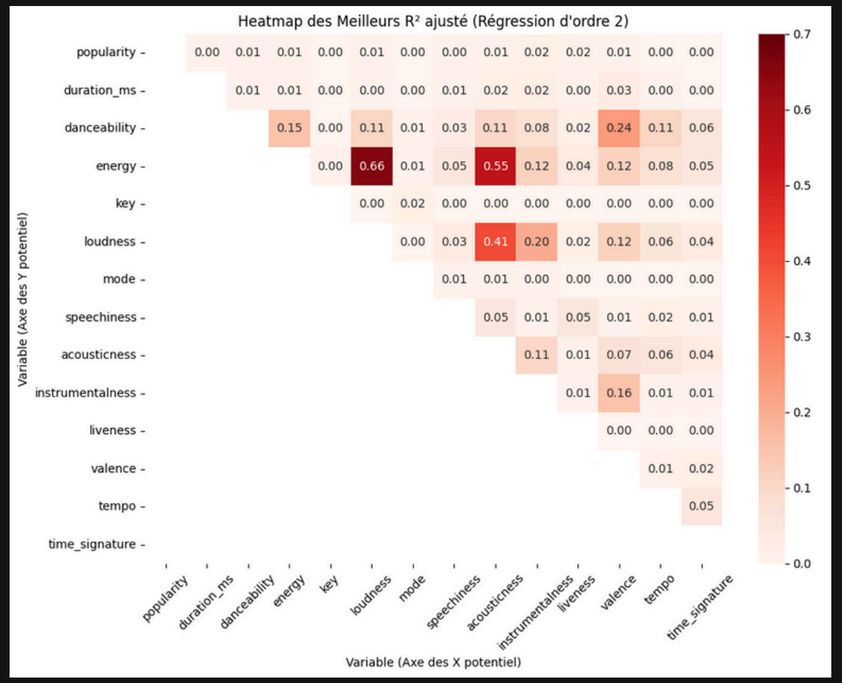


2. Analyse du R²

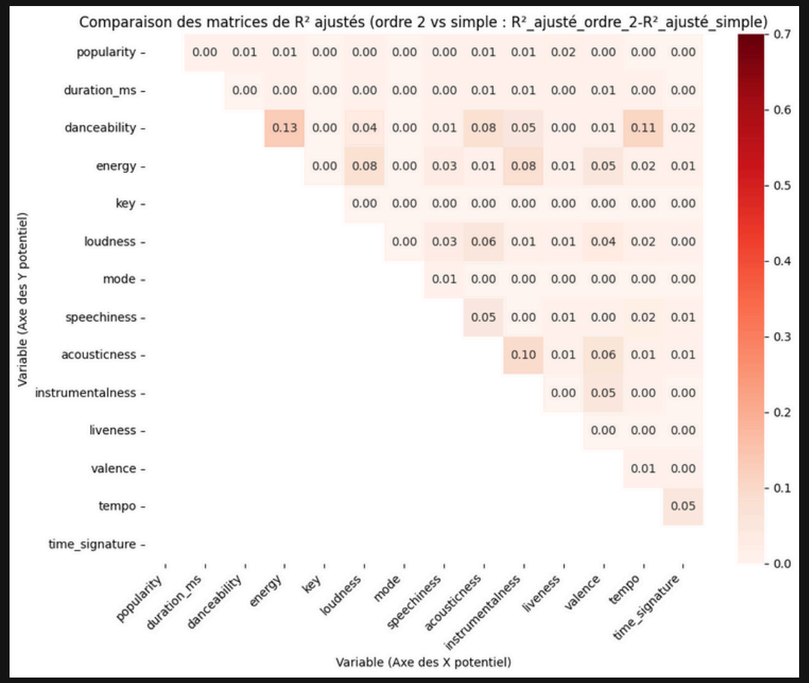
Heatmap des meilleurs R² ajusté en Régression linéaire simple



Heatmap des meilleurs R² ajusté en Régression d'ordre 2

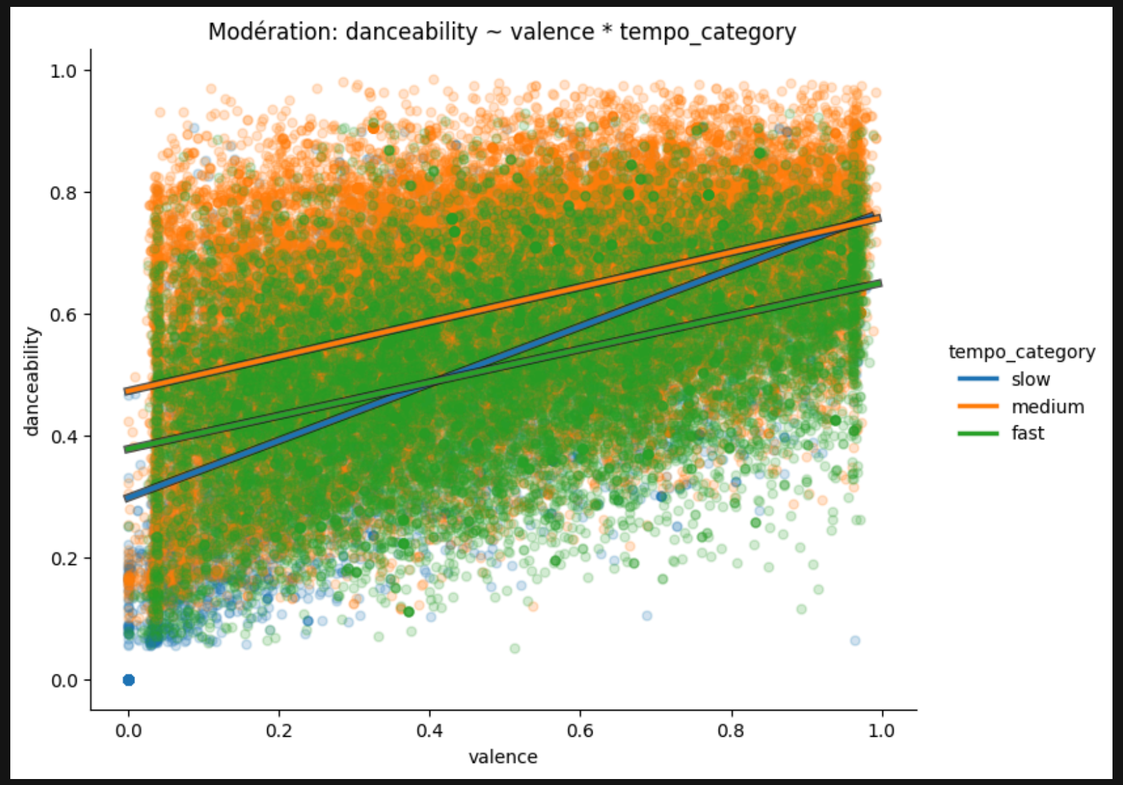


Comparaison des matrices R² ajustés



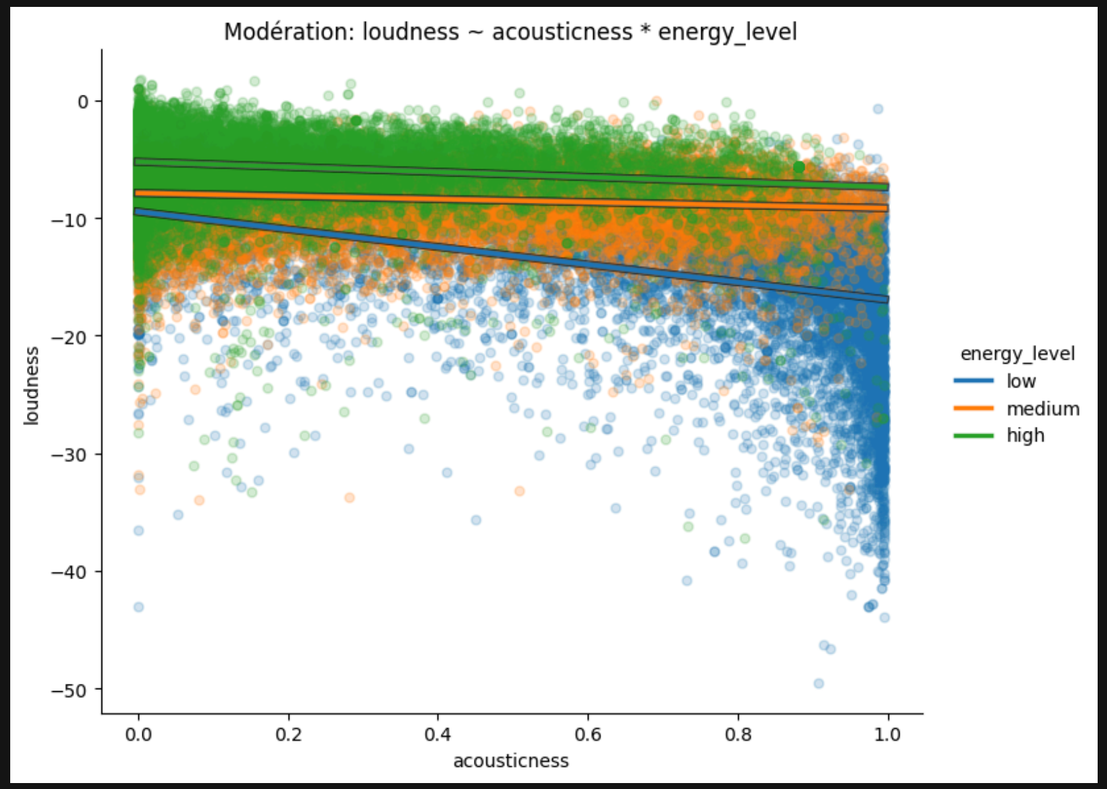
3. Handcrafted features

1. DANCEABILITY - VALENCE X TEMPO_CATEGORY



R^2 adj (Interaction) : 0.3175
R^2 adj (Effets Principaux) : 0.3082
ΔR^2 adj (dû à l'intéraction) : 0.0093
ΔR^2 adj (différence avec la Régression simple) : 0.0675

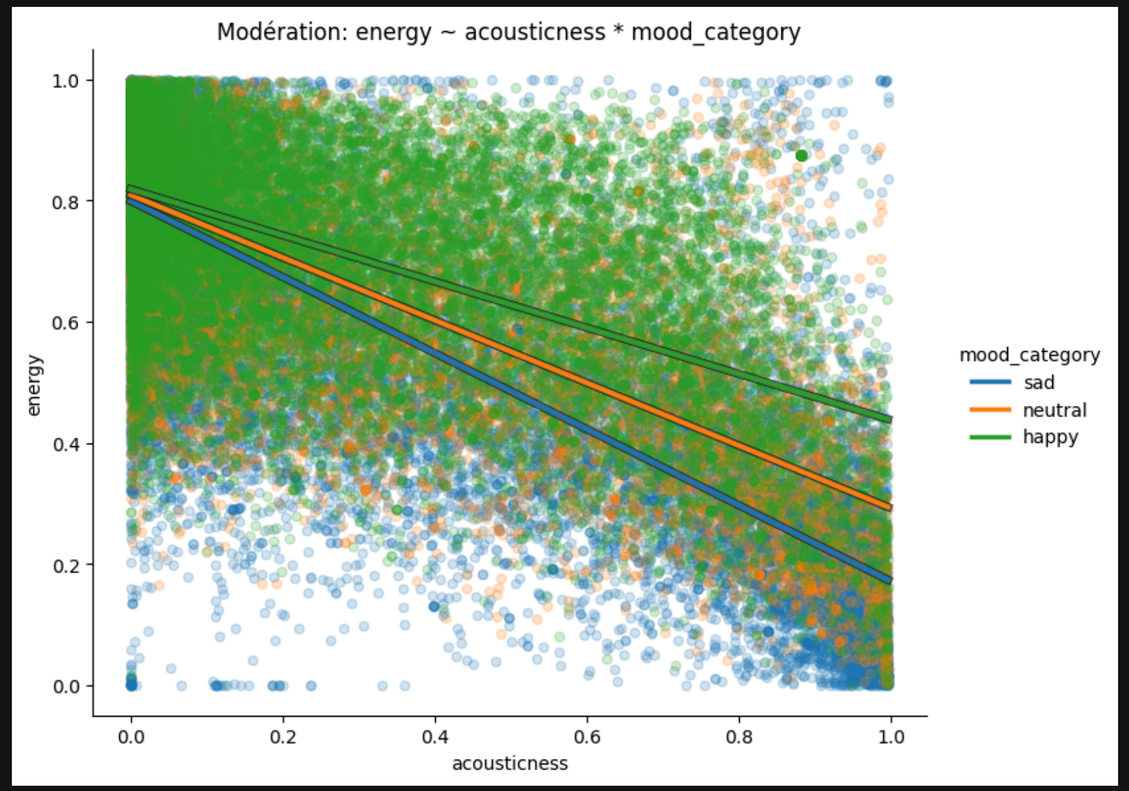
2. LOUDNESS - ACOUSTICNESS X ENERGY_LEVEL



R^2 adj (Interaction) : 0.5260
R^2 adj (Effets Principaux) : 0.5130
ΔR^2 adj (dû à l'intéraction) : 0.0130
ΔR^2 adj (différence avec la Régression simple) : 0.176

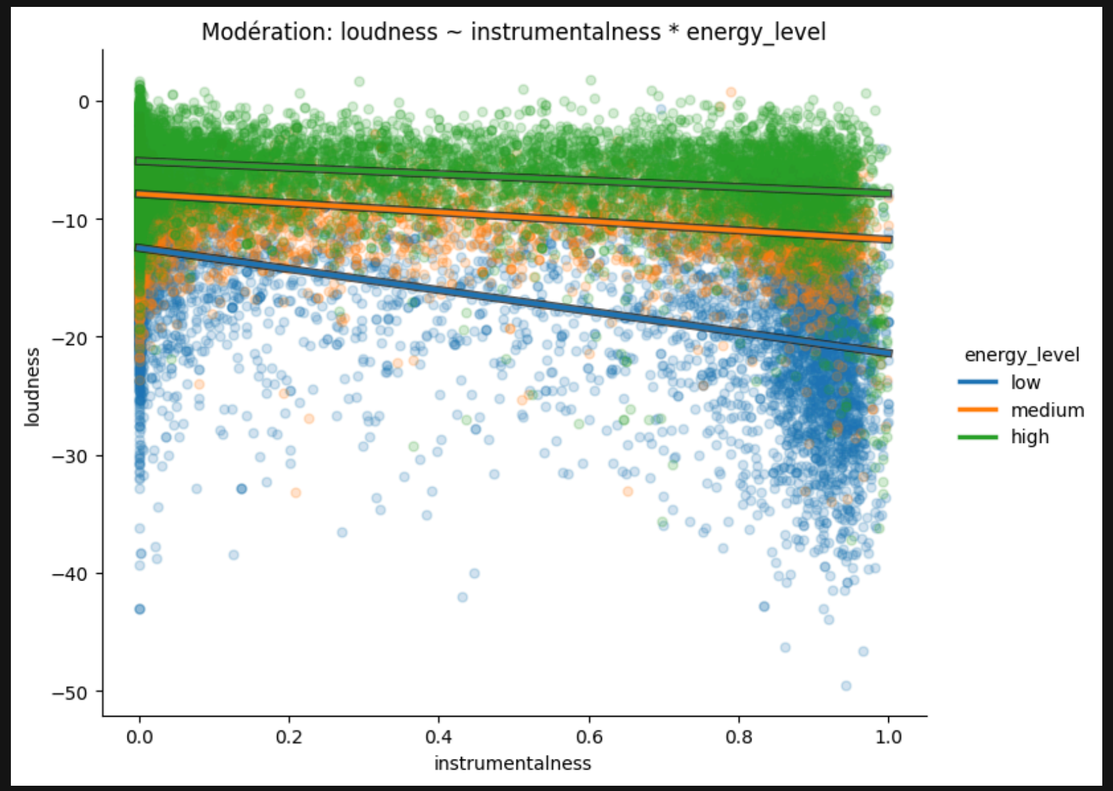
3. Handcrafted features

3. ENERGY - ACOUSTICNESS X MOOD_CATERGORY



R^2 adj (Interaction) : 0.5840
R^2 adj (Effets Principaux) : 0.5656
ΔR^2 adj (dû à l'intéraction) : 0.0184
ΔR^2 adj (différence avec la Régression simple) : 0.044

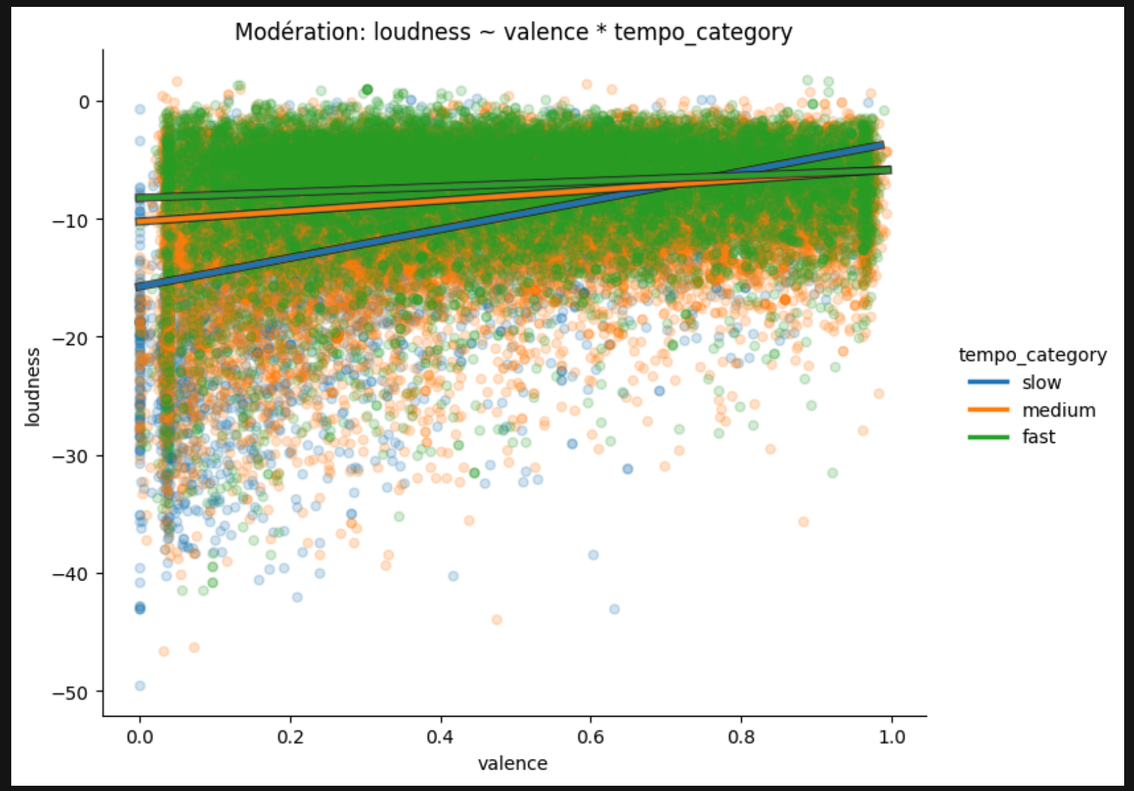
4. LOUDNESS - INSTRUMENTALNESS X ENERGY_LEVEL



R^2 adj (Interaction) : 0.6146
R^2 adj (Effets Principaux) : 0.5876
ΔR^2 adj (dû à l'intéraction) : 0.0270
ΔR^2 adj (différence avec la Régression simple) : 0.41

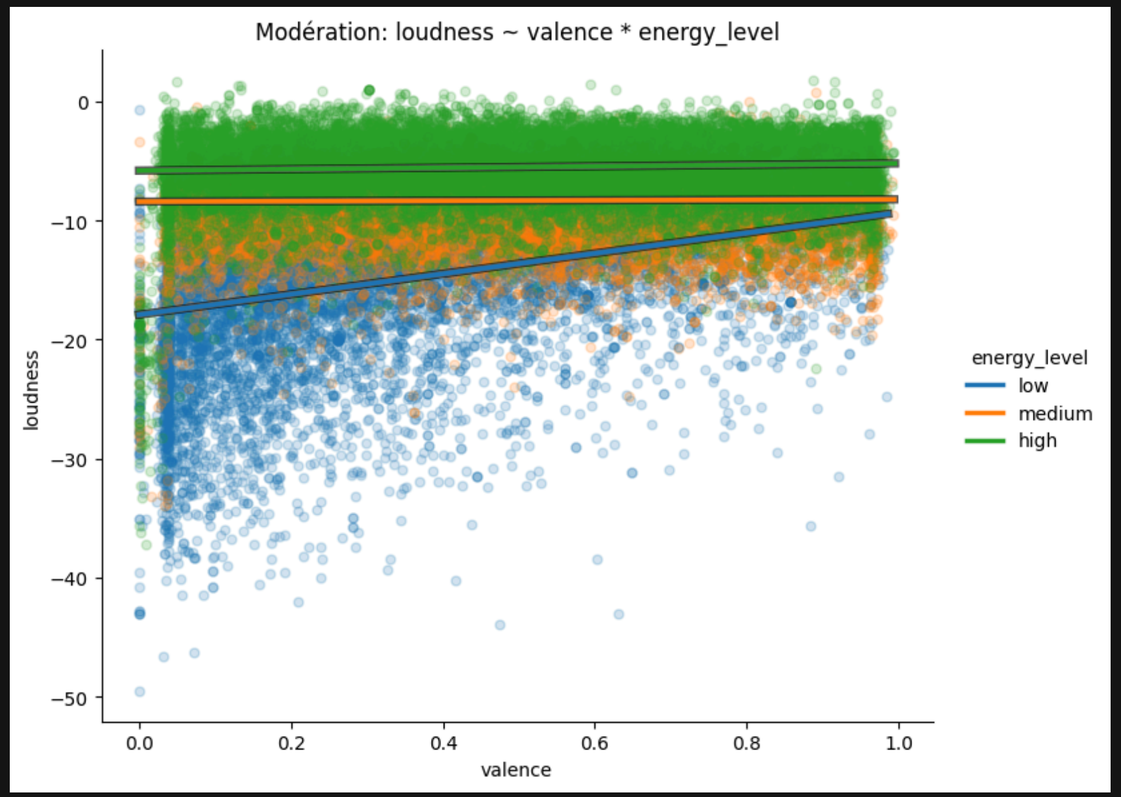
3. Handcrafted features

5. LOUDNESS - VALENCE X TEMPO_CATEGORY



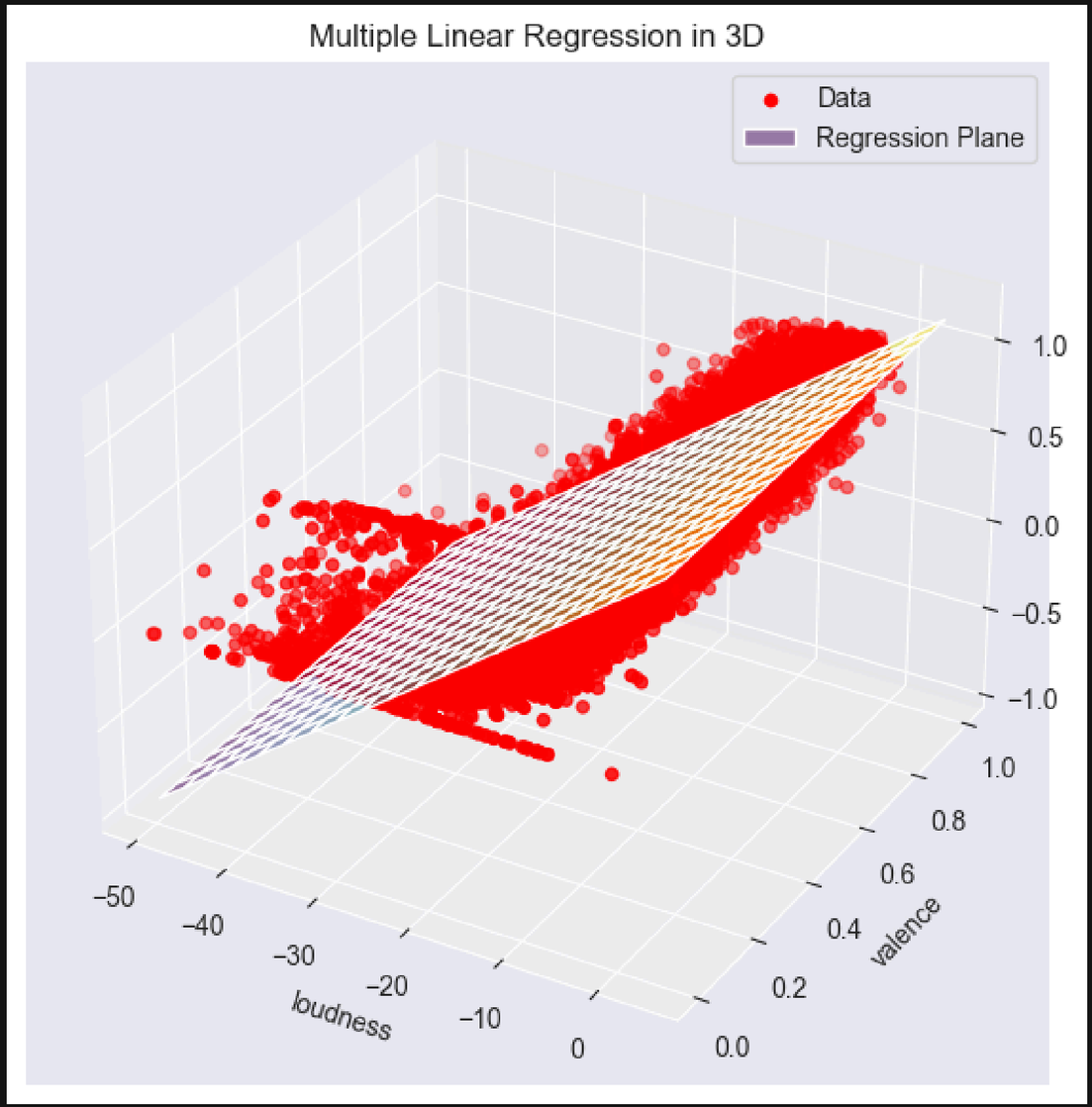
R^2 adj (Interaction) : 0.1444
R^2 adj (Effets Principaux) : 0.1205
ΔR^2 adj (dû à l'intéraction) : 0.0239
ΔR^2 adj (différence avec la Régression simple) : 0.064

6. LOUDNESS - VALENCE X ENERGY_LEVEL

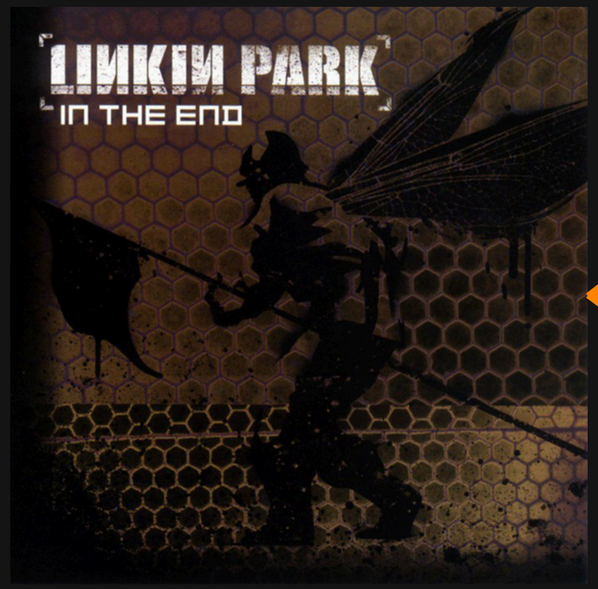


R^2 adj (Interaction) : 0.5225
R^2 adj (Effets Principaux) : 0.5016
ΔR^2 adj (dû à l'intéraction) : 0.0209
ΔR^2 adj (différence avec la Régression simple) : 0.44

4. Régression linéaire multiple



VI. SIDEQUEST



'DANCEABILITY'

'ENERGY'

'VALENCE'

'ACOUSTICNESS'

'INSTRUMENTALNESS'

'LIVENESS'

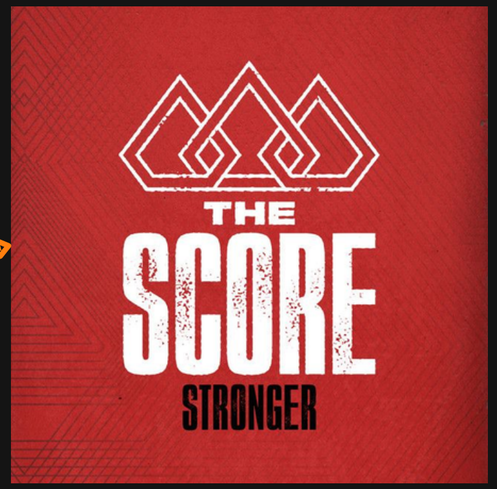
'SPEECHINESS'

'TEMPO'

8D

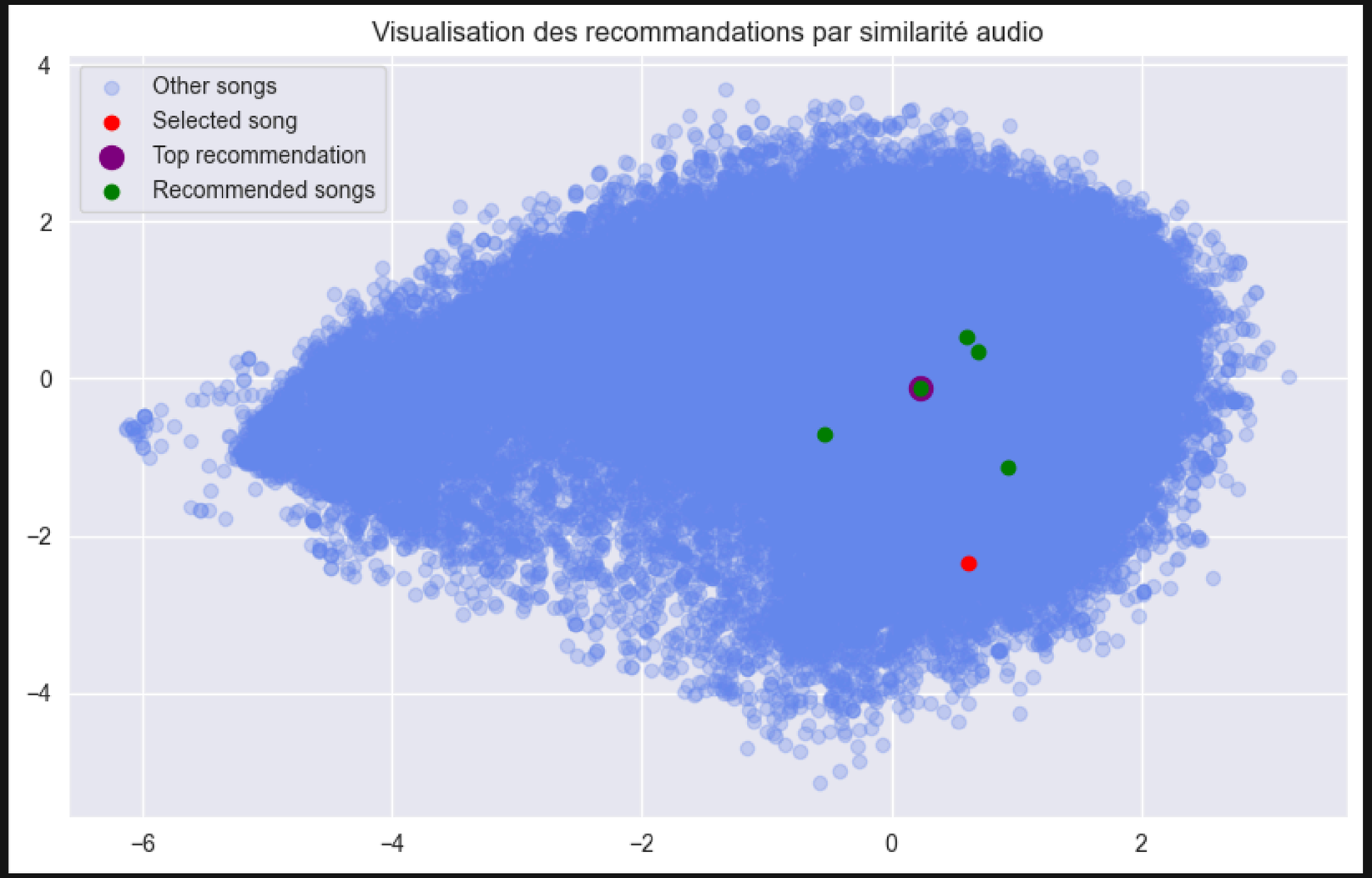
FILTRAGE DES DONNÉES PAR GENRE

track_name	artists
Stronger	The Score
Rain (from The Suicide Squad)	grandson;Jessie Reyez
The Fear	The Score
Nervous	The Neighbourhood
Scary Love	The Neighbourhood



SIMILARITY SCORE : 73%

SIMILARITY SCORE : 69%


Visualisation des recommandations par similarité audio
Other songs
Selected song
Top recommendation
Recommended songs
4
2
0
−2
−4
−6
−4
−2
0
2

MERCI POUR VOTRE ATTTNTION !