Rapport : Traitemant des Données

Ansel Camille, Catteau Elsa, Elfarsi Anas

October 29, 2025

Sommaire

1 Buisness Goal

Nous sommes un groupe d'amis passionnés de musique et nous envisageons de lancer un label de musique indépendant. Rapidement, une question s'est imposée : quels facteurs ont une influence sur la popularité d'une musique? Est-ce que le rythme, la tonalité, le genre, la durée et le tempo ont une forte influence sur la popularité d'une musique? Nous avons décidé d'analyser rigoureusement un jeu de données de plus de 50 000 titres, comprenant 18 variables telles que la `danceability`, l'`energy`, le `tempo`, ainsi que la variable qui est pour nous la plus importante, la popularité de chaque titre de musique que nous avons à disposition.

Objectif principal :

- Identifier les facteurs qui influencent la popularité d'une musique.

2 Team management

- **Mardi matin :** Choix et analyse du dataset, analyse du dataset.
- **Mardi après-midi :** Analyse du dataset, on a fait le datacleaning et on a commencé à créer des handcrafted features.
- **Mercredi matin :** Réalisation des premiers graphiques (boîtes à moustaches, barplots) et détermination des corrélations de notre dataset.
- **Mercredi après-midi :** Implémentation de la régression linéaire multiple.

- **Jeudi matin :** On a corrigé notre régression linéaire, et on a implémenté une régression logistique. De plus, on a commencé la présentation.
- **Jeudi après-midi :** On a fini la partie sur la régression linéaire et multiple et on a terminer notre présentation.
- **Vendredi matin :** Entraînements pour la présentation orale
- **Vendredi après-midi :** Rédaction et mise en forme finale du rapport.

3 Data visualisation

3.1 Dataset

Notre base de données contient environ **50 000 morceaux**, chacun décrit par :

- `instance_id` : Identifiant unique du morceau (entier)
- `artist_name` : Nom de l'artiste (texte)
- `track_name` : Titre du morceau (texte)
- `music_genre` : Genre musical (texte)
- `mode` : Majeur / Mineur (texte)
- `key` : Tonalité (A, A#, B, ...) (texte)
- `obtained_date` : Date d'extraction des données (texte)
- `popularity` : Score de popularité (0–100, entier)
- `duration_ms` : Durée en millisecondes (entier)
- `acousticness` : Acoustique (0–1, réel)
- `danceability` : Danceabilité (0–1, réel)
- `energy` : Énergie (0–1, réel)
- `instrumentalness` : Instrumentalité (0–1, réel)
- `liveness` : Vivacité (0–1, réel)
- `loudness` : Intensité sonore en dB (réel)
- `speechiness` : Proportion de voix/paroles (0–1, réel)
- `valence` : Valence émotionnelle (0–1, réel)
- `tempo` : Tempo en BPM (réel ou "?")

3.2 Data Cleaning

Avant l'analyse, nous avons appliqué un nettoyage en plusieurs étapes :

1. **Colonnes supprimées :** `instance_id` (identifiant unique) et `obtained_date` (date où la musique a été mise dans le dataset) n'apportaient rien à la prédiction.
2. **Lignes vides ou invalides:** Tout enregistrement comportant une valeur manquante a été retiré.
   - Si `duration_ms` = –1, la durée du morceau n'est pas renseignée.
   - Si `tempo` = "?", le tempo n'est pas renseigné.

3. **Conversion de la durée :** Création de `duration_sec` (durée en secondes) en divisant `duration_ms` par 1000, puis suppression de `duration_ms`.
4. **Nettoyage des titres :** À partir de `track_name`, génération de `title_clean` :
   - Passage en minuscules
   - Suppression des accents
   - Retrait de la ponctuation
5. **Encodage catégoriel :**
   - `key` (tonalité) → colonnes binaires (`key_A`, `key_A#`, ...).
   - `music_genre` → colonnes binaires (`music_genre_Rock`, `music_genre_Electronic`, ...).
6. **Suppression des doublons :** Recherche et retrait des lignes strictement identiques.

À l'issue de ce nettoyage, notre jeu de données est homogène, sans valeurs manquantes ni redondantes.

3.3 Let's focus on popularity

1. Popularité moyenne par plage de tempo (BPM)

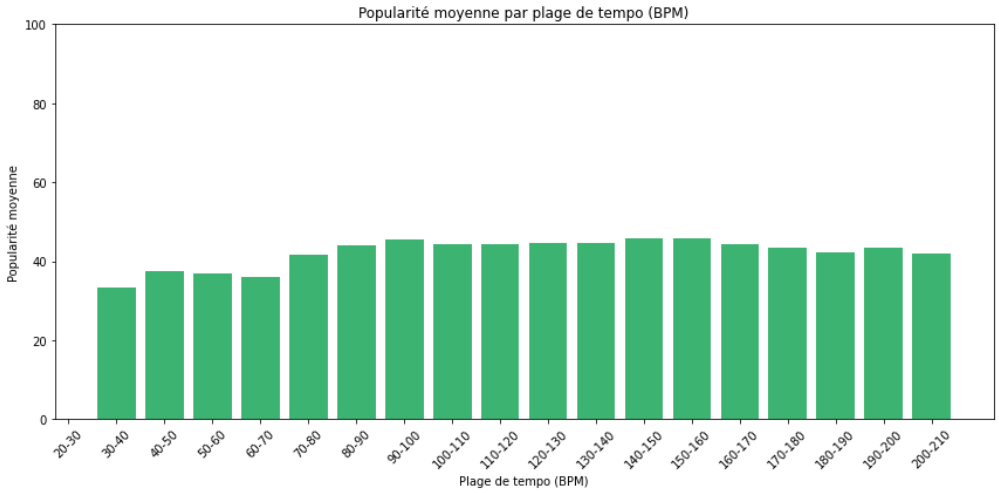

Figure 1: Popularité moyenne par plage de tempo (BPM).

Comme le montre la figure 1 :

- Les morceaux très lents (20–50 BPM) affichent une popularité moyenne plus faible (33–37 points).
- À partir de 60–70 BPM, la popularité remonte nettement (42 points), puis culmine autour de 45–46 points entre 80 et 160 BPM, avec un pic à 140–150 BPM (46 points). Cette plage correspond souvent aux tempos de la pop contemporaine et de certains styles électro (house, EDM).
- Au-dessus de 160 BPM, la popularité redescend légèrement (42–44 points), suggérant que les tempos extrêmes (très rapides) sont moins plébiscités.

On constate une corrélation modérée entre tempo et popularité : les tempos intermédiaires (80–150 BPM) sont légèrement favorisés, mais l'effet n'est pas très fort. C'est pourquoi, nous avons cherché d'autres types de features (textuelles et interactions) pour mieux expliquer la popularité.

2. Popularité moyenne par tonalité

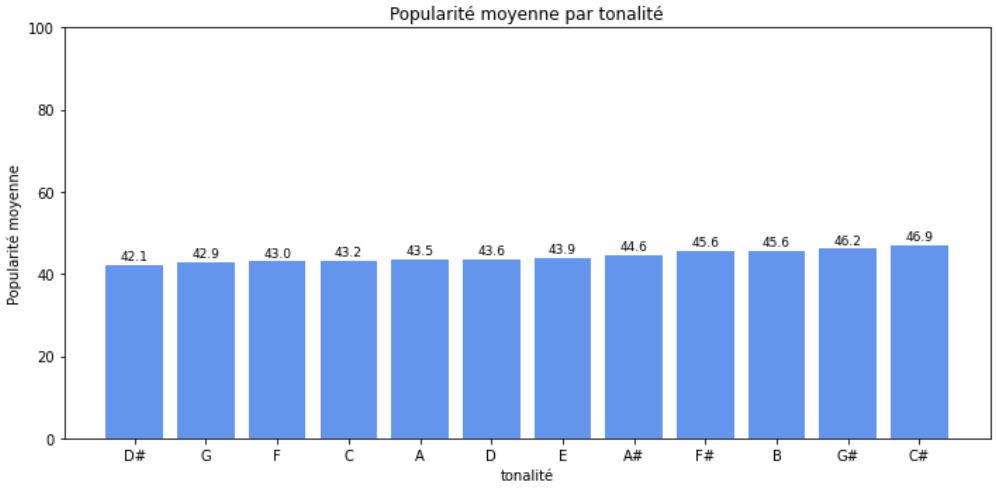

Figure 2: Popularité moyenne par tonalité.

D'après la figure 2 :

- Les tonalités graves sont parmi les moins populaires (D#: 42,1 %, G: 42,9 %).
- La popularité augmente progressivement en montant vers l'aigu (F: 43,0 % ; C: 43,2 % ; A: 43,5 % ; D: 43,6 %; E: 43,9%).
- Les tonalités aiguës obtiennent les scores les plus élevés (A#: 44,6 % ; F# et B: 45,6 % ; G#: 46,2 % ; C#: 46,9 %).

Il apparaît une tendance minimale : les tonalités aiguës semblent légèrement mieux notées, mais l'écart global reste faible (environ 4 à 5 points de D# à C#). Ce lien n'est pas très marquant et ne suffit pas à expliquer la majorité de la variance. Nous avons donc exploré d'autres variables (textuelles, interactions) pour compléter notre analyse.

3. Influence du genre musical sur la popularité

Pour voir l influence du genre musical sur la popularité, on a décidé de le visualiser avec des boîtes à moustache. Grâce à ce graphe on peut voir la moyenne de popularité pour chaque genre ainsi que la médiane. Ainsi, on peut observer que les 3 genres les plus populaires dans notre jeu de données sont : le rap, le rock et le hip hop.

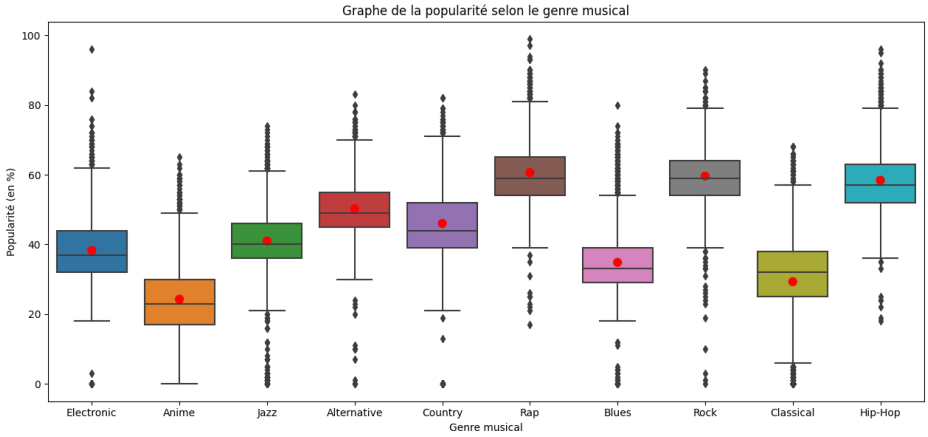

Figure 3: Graphe montre l'influence du genre musical sur la popularité

3.4 Genre characteristics

Nous observons un lien un peu plus important du genre avec la popularité. On peut observer les caractéristiques de chaque genre.

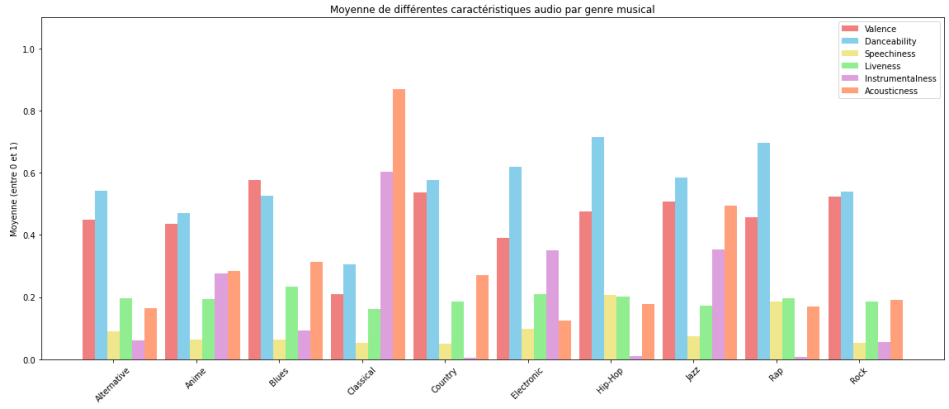

Figure 4: Moyenne des différentes caractéristiques audio par genre musical

Dans la figure ci-dessus certaines variables comme 'acousticness' varient beaucoup. Mais d'autres comme 'liveness' permettent difficilement d'établir des conclusions. De manière générale aucune variable seule ne semble expliquer le genre musical mais il serait sûrement possible de déduire celui-ci avec une méthode de machine-learning.

4 Handcrafted features

4.1 Track name based features

Il y a t-il un lien entre le titre d'une musique et sa popularité ? Pour répondre à cette question nous avons cherché comment analyser le titre. Un moyen de faire cela est de le considérer comme une combinaison de mots et d'étudier les mots qui le composent.

Nous avons extrait et classé par nombre d'occurrences les mots qui composent les titres. Il a été nécessaire d'écarter les mots vides de sens (ex: the, mix, 2, concerto, etc.).

On observe que 'no' et 'love' ressortent premiers, nous avons donc créé deux features qui indiquent si une négation pour l'un et l'idée d'amour pour l'autre sont présentes dans le titre. Nous avons essayé de prendre en compte la diversité des langues et de leurs expressions.

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

Figure 5: Proportion de `has_love` et `has_no` dans le dataset

On voit sur la figure 1 que ces deux indicateurs ciblent une petite partie du dataset.

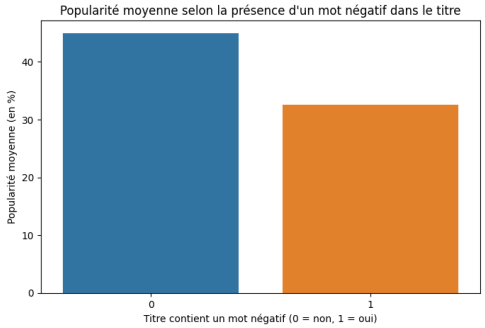

Figure 6: Popularité moyenne en fonction de `has_no`

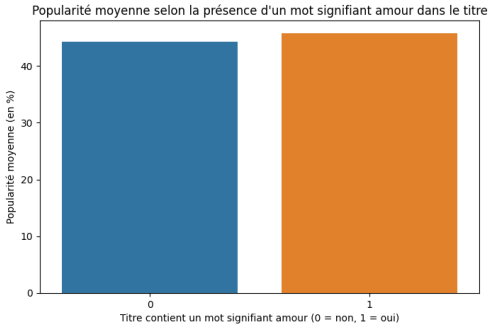

Figure 7: Popularité moyenne en fonction de `has_love`

On observe que has_love n'a pas un impact significatif sur la popularité contrairement à has_no qui modifie en moyenne de 10% la popularité d'une musique.
Faisons un pas de côté et différencions les musique en fonction de la présence ou non d'un mot appartenant au top 50 de notre classement.

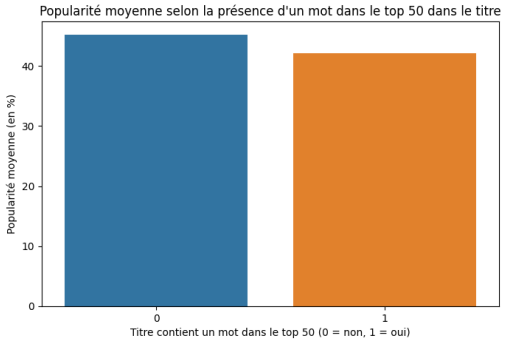

Figure 8: Popularité moyenne en fonction de `has_top`

En faisant varier has_top entre 10 et 100 on se rend compte que réduire la valeur tend à donner un graphe similaire à has_no tandis que l'augmenter tend à égaliser les résultats. On observera plus tard l'influence de ces nouvelles features.

4.2 Link between artist and track

Nous allons maintenant chercher à exploiter le nom de l'artiste. En supposant que la moyenne des popularités des musiques d'un artiste est proche de la popularité de sa musique que l'on souhaite prédire. On retourne donc le dataset pour que chaque ligne corresponde à un artiste en moyennant ses différents paramètres.

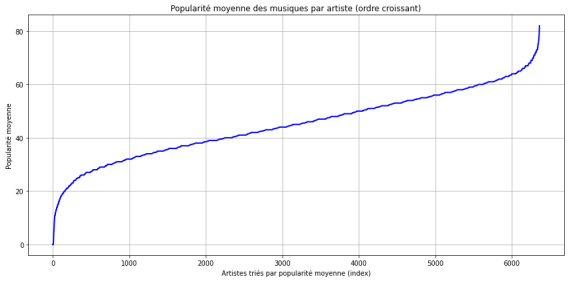

Figure 9: Popularité des artistes dans l'ordre croissant

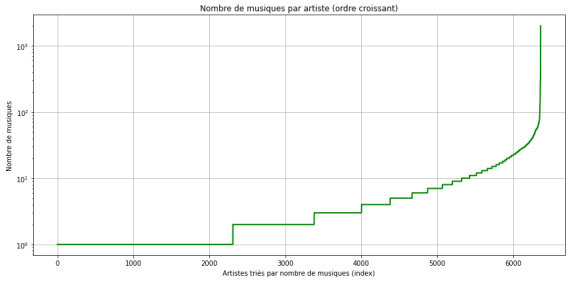

Figure 10: Nombre de musique des artistes par ordre croissant

Dans le graphe à gauche, on observe que tout le spectre de popularité est rempli par les artistes. Il serait donc possible de différencier la popularité d'une musique en fonction de son artiste. Mais comme pour le titre il est évidemment impossible de faire une régression linéaire sur le nom de l'artiste directement. Nous devons trouver une variable de l'artiste ayant un lien avec sa popularité. Essayons avec le nombre de titres de l'artiste présents dans le dataset.
[1]

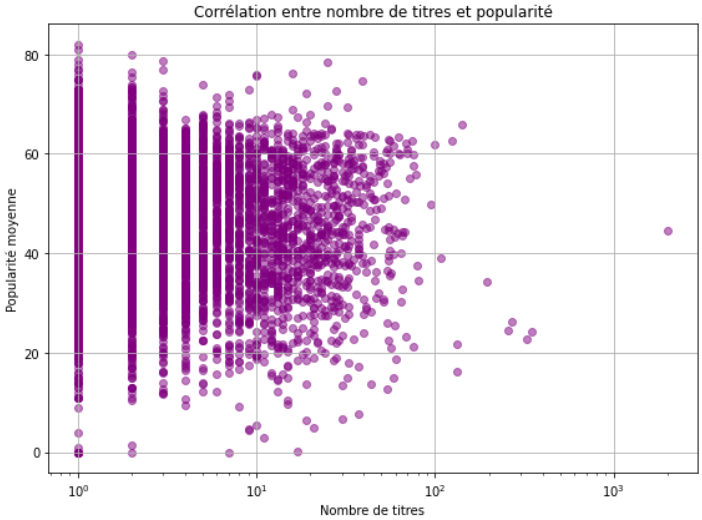

Figure 11: Nombre de musique des artistes par ordre croissant

Sur le graphe de la popularité d'un artiste en fonction du nombre de titre, on n'observe un nuage de point qui ne révèle pas de lien logique. Nous avons donc traçé la matrice de corrélation de notre dataset orienté artiste. Celle-ci est très similaire à la matrice de corrélation des variables du dataset orienté musique. Nous avons décidé de ne pas continuer sur cette piste que semble être un simple détour.

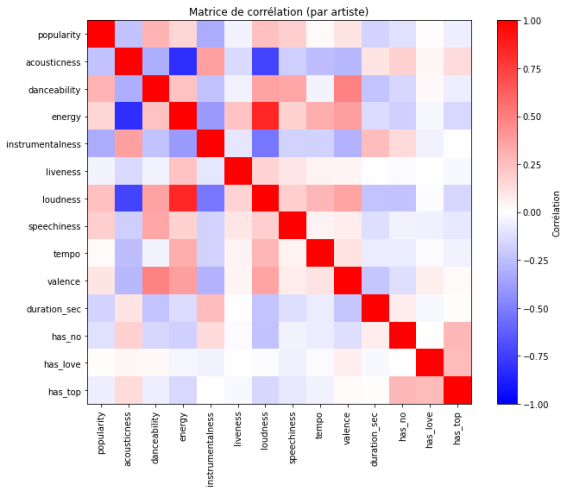

Figure 12: Matrice de corrélation (par artiste)

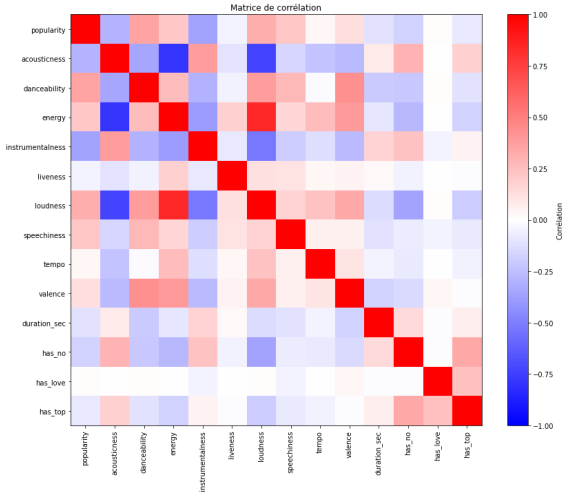

Figure 13: Matrice de corrélation sur le dataset

[1] À posteriori, nous nous sommes rendu compte qu'un tiers des artistes n'ont pas de musique, d'après la figure 9. Par manque de temps, nous nous sommes concentrés sur la suite de notre analyse. La valeur la plus élevée correspond quant à elle à `unknown_artist`.

5 Linear regression

Pour la régression linéaire multiple, on a la formule suivante qui nous permet d'avoir y en fonction de coefficients associés à chaque variables.

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

Les indicateurs qui nous permettent de voir si notre modèle est bon sont R^2 et R^2 ajusté. Pour la régression linéaire multiple, le R^2 ajusté est plus adapté mais dans notre cas, on a trouvé la même valeur pour R^2 et R^2 ajusté. Pour effectuer la régression linéaire, on doit choisir les variables qu'on prend en compte, on a fait le choix de prendre toutes les variables qu'on a gardé après le data cleaning et d'ajouter celles qu'on a créé précédemment.

De plus, on a créé de nouvelles variables, on a multiplié certaines variables entre elles suivant si elles étaient corrélées entre elles. Voici la liste des variables qu'on a créé :

- `acoustic_dance` : corrélation négative entre `acousticness` et `danceability`
- `acoustic_energy` : corrélation négative entre `acousticness` et `energy`
- `acoustic_instrument` : corrélation positive entre `acousticness` et `instrumentalness`
- `acoustic_loudness` : corrélation négative entre `acousticness` et `loudness`
- `acoustic_valence` : corrélation négative entre `acousticness` et `valence`
- `dance_energy` : corrélation positive entre `danceability` et `energy`
- `dance_instrument` : corrélation négative entre `danceability` et `instrumentalness`
- `dance_loudness` : corrélation positive entre `danceability` et `loudness`
- `dance_speech` : corrélation positive entre `danceability` et `speechiness`
- `dance_valence` : corrélation positive entre `danceability` et `valence`
- `energy_instrument` : corrélation négative entre `energy` et `instrumentalness`
- `energy_loudness` : corrélation positive entre `energy` et `loudness`
- `energy_valence` : corrélation positive entre `energy` et `valence`
- `energy_hasno` : corrélation positive entre `energy` et `has_negation`
- `instrumental_loudness` : corrélation négative entre `instrumentalness` et `loudness`
- `instrumental_valence` : corrélation négative entre `instrumentalness` et `valence`
- `loudness_hasno` : corrélation positive entre `loudness` et `has_negation`
- `loudness_valence` : corrélation positive entre `loudness` et `valence`

Pour notre modèle, on a donc trouvé un R^2 de 0.63 et on a affiché les valeurs de chaque coefficient bêta qu'on a affiché sous forme de graphe.

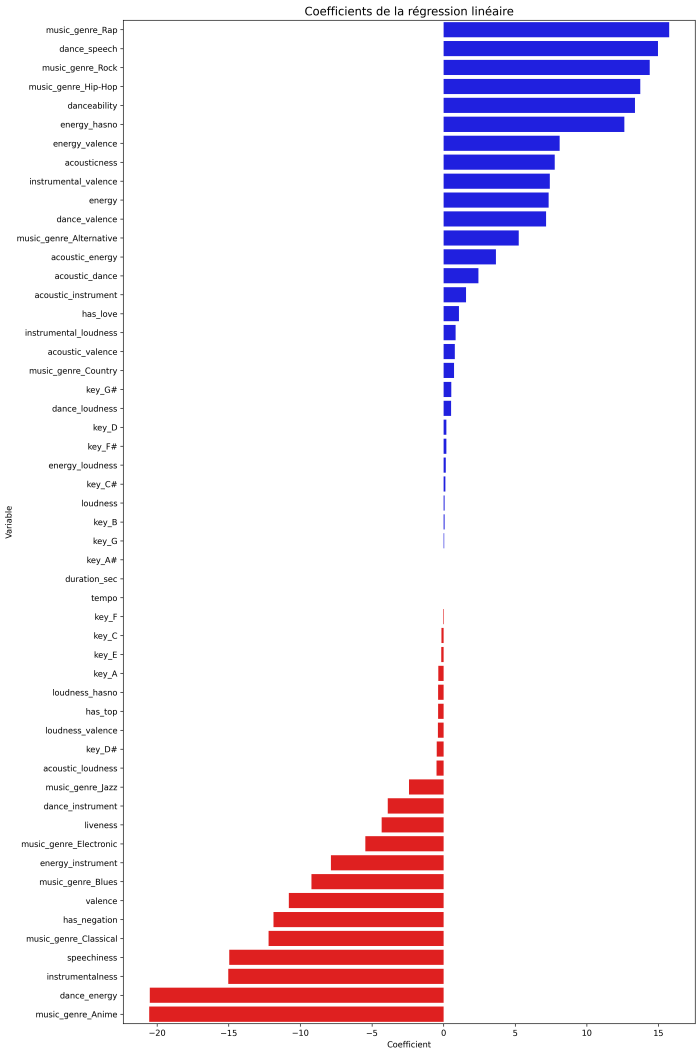

Figure 14: Coefficients bêta de la régression linéaire

On peut donc observer que les variables qui ont une influence positives importantes sont: le Rap, la capacité de danser et la présence de parole, le rock et le hiphop. Les variables qui ont une fortes influence négatives sont : le genre Animé, le fait que la musique soit dansante et énergique. Les variables citées sont donc celles qui ont le plus d'influence sur la régression linéaire.

Observons nos handcrafted features :

- has_no, acoustic_loudness, dance_instrument, energy_instrument, dance_energy : apparaîssent comme des variables influençant négativement la popularité d'une musique
- has_love, dance_speech, energy_hasno, energy_valence, instrumental_valence, dance_valence, acoustic_energy, acoustic_dance, acoustic_instrument, instrumental_loudness, acoustic_valence : quant à elles augmentent la popularité d'une musique
- has_top, dance_loudness, energy_loudness, duration_sec, loudness_hasno, loudness_valence : semblent assez anecdotique comparé aux autres

On a pu également observer que le grenre musical à une forte influence sur la popularité car en ne mettant pas cette variable pour la régression linéaire, on avait un R^2=0.26.

De plus, on a voulu tester notre modèle en prenant une musique aléatoirement pour comparer la popularité réelle avec celle prédite. Pour le titre Lost & Not Found - Acoustic de l'artiste Chase &

Status, on a une valeur réelle de popularité de 30,0% et une valeur prédite de 34,06%. Il y a donc une différence non négligeable c'est pourquoi nous avons voulu tester la régression logistique.

5.1 Logistic Regression

Comme notre R^2 n'est pas proche de 1, on a essayé de faire une régression logistique où on a défini une musique comme étant populaire si sa popularité était supérieur ou égal à 50 %. Ainsi, si la musique est considérée comme étant populaire elle prend la valeur 1 sinon elle prend la valeur 0. On a alors obtenu ce tableau avec des valeurs assez convaincante car elle sont tous proche de 1.

	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

Figure 15: Résultat de la régression logistique

6 Conclusion

La régression linéaire a montré qu'environ 63 % de la variance de la popularité peut s'expliquer par des mesures audio (énergie, danceability, loudness, etc.), des interactions entre ces mesures et quelques indicateurs textuels extraits des titres. Les 37 % restants correspondent vraisemblablement à des facteurs externes (marketing, notoriété de l'artiste, viralité, etc.), non présents dans notre dataset.

Cette analyse nous a donc montré que de nombreux paramètres ont une influence sur la popularité d'une musique. Le genre semble être le facteur qui a l'effet le plus significatif sur la popularité de la musique.

Pour améliorer nos résultats, on pourrait :

- Intégrer des données externes : données sur l'artiste (nombre de followers, volume de streams), placements en playlists influentes.
- Expérimenter d'autres modèles tel que le K-Nearest Neighbors par exemple.