Rapport de Projet : Traitement de Données

Ferre Gabriel, Patel Lucah, Prouzet Charlotte

November 1, 2025

Contents

1. Business Goal

1.1. Contexte

Au cours de ce projet, nous avons décidé d'utiliser le dataset *Spotify Tracks Dataset*. Il s'agit d'un ensemble de données de titres Spotify couvrant 125 genres différents. Chaque titre est associé à des caractéristiques audios. Les données sont au format CSV, sous forme de tableau, et peuvent être chargées rapidement.

1.2. Quelle est l'utilité de notre dataset ?

Notre dataset référence des caractéristiques comme l'énergie, la valence, l'acoustique, le tempo etc, ce qui nous a permis de chercher les corrélations entre plusieurs caractéristiques, comme par exemple : « *y a t-il un lien entre l'énergie et la valence ?* ».
Il offre des données sur la popularité d'un son, permettant de faire le lien entre caractéristiques musicales et succès d'un titre. Nous pourrions ainsi prédire si un son pourrait, ou non, devenir un "hit" en se basant uniquement sur ses caractéristiques musicales.
Enfin, notre dataset couvre un très grand nombre de genres (125 au total) et de pistes (114 000), ce qui nous permet d'analyser la structure musicale des genres, et d'identifier des groupes similaires de morceaux en se basant sur les caractéristiques communes de ces derniers.

1.3. Objectif

À travers ce projet, nous cherchons à comprendre comment Spotify parvient à générer des transitions musicales aussi fluides que inattendues, comme passer du rap américain à la pop disco sans rupture perceptible.

Tout au long de ce projet, notre objectif sera d'explorer deux axes principaux : d'une part, le rôle de la popularité d'un son dans les suggestions faites à l'utilisateur, et d'autre part, la proximité musicale entre les titres, notamment à travers des caractéristiques comme l'énergie, la valence ou le tempo. Nous viserons ainsi à modéliser ces relations.

Enfin, une réussite consisterait à mettre en évidence les facteurs déterminants dans le fonctionnement des recommandations, à observer des regroupements cohérents entre morceaux, et à simuler des transitions musicales pertinentes à partir des données disponibles.

2. Team Management

2.1. Répartition des rôles

Charlotte : Définir l'objectif business, interpréter les résultats, proposer des idées d'application, rendre les résultats compréhensibles et visuellement impactants, faire de la data storytelling.

Gabriel : analyser le dataset, identifier les patterns, corrélations, tendances intéressantes, implémentation de la régression.

Lucah : construire les features et les modèles de prédiction, normalisation, réduction de dimension, implémentation de la régression et du système de recommandation.

2.2. Workflow

Notre projet s'est construit au fil de plusieurs séances de cours et de travaux dirigés. Nous avons travaillé sur ce projet pendant **4 séances de cours d'1 heure**, suivies de **6 séances de TD de 2 à 3 heures**, ce qui nous a permis de progresser étape par étape, de la recherche de données à la construction d'un système de recommandation.

2.3. Recherche de dataset

Au départ, chacun d'entre nous a exploré différents jeux de données disponibles sur *Kaggle*, en essayant de trouver un dataset à la fois riche, pertinent et motivant. Après discussion, nous avons tous eu un coup de cœur commun pour un dataset lié à **Spotify**, contenant plus de 114 000 morceaux avec des caractéristiques audio détaillées. C'était à la fois intéressant techniquement, et parlant pour nous tous en tant qu'utilisateurs de la plateforme.

2.4. Définition de la problématique

Une fois le dataset sélectionné, nous avons échangé sur ce que nous voulions faire avec ces données. Nous avons choisi de travailler autour d'une problématique qui nous semblait concrète et actuelle :

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

2.5. Analyse exploratoire des données

Nous avons ensuite plongé dans le dataset pour mieux comprendre les différentes variables (`danceability`, `energy`, `valence`, `tempo`, etc.), repérer les valeurs manquantes, les corrélations, et obtenir une première intuition sur les données. À ce stade, nous avons utilisé des outils comme `pandas`, `matplotlib` et `seaborn` pour créer des visualisations (histogrammes, heatmaps, boxplots, etc.) et mieux comprendre les relations entre les caractéristiques musicales.

2.6. Préparation des données et modélisation

Nous avons nettoyé nos données pour supprimer les valeurs manquantes ou aberrantes, puis nous avons appliqué une normalisation sur les variables numériques afin d'obtenir des résultats plus cohérents lors de l'analyse.

2.7. Visualisation et interprétation des résultats

Notre visualisation s'est concentrée sur les liens de dépendance entre les différentes variables musicales, notamment à travers des cartes de chaleur et des graphiques de dispersion.

3. Data Visualization

3.1. Dataset Description

Le dataset utilisé provient de Kaggle et contient environ **114 000 morceaux musicaux**, chacun décrit par plusieurs variables numériques et catégorielles issues de l'analyse audio de Spotify. Chaque ligne du dataset représente un morceau unique.

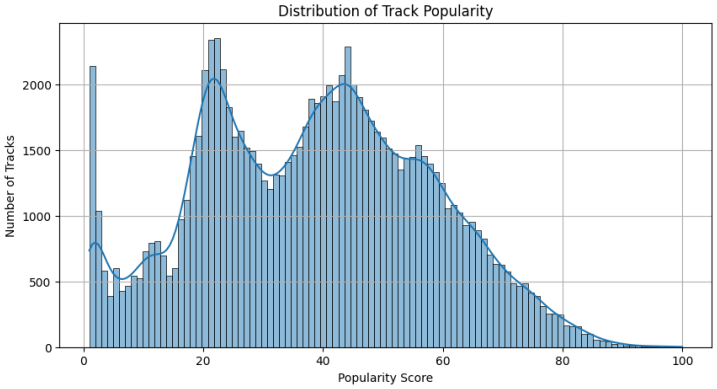
Voici quelques-unes des principales colonnes :

- `track_id`, `track_name`, `artists`, `album_name` : informations d'identification.
- `duration_ms`, `explicit`, `popularity` : durée, caractère explicite, popularité.
- `danceability`, `energy`, `acousticness`, `instrumentalness`, `liveness`, `valence`, `speechiness` : descripteurs audio liés à l'ambiance ou au style du morceau.
- `tempo`, `loudness`, `key`, `mode`, `time_signature` : variables musicales plus techniques.
- `track_genre` : genre musical du morceau.

3.2. Visualisations produites

Afin de mieux comprendre les caractéristiques de notre dataset musical, nous avons réalisé une première série de visualisations, ce qui nous a permis d'explorer la distribution des variables et leurs relations.

- **Histogramme** pour visualiser la distribution de popularité des morceaux de notre dataset.



Cet histogramme montre que la majorité des morceaux ont une popularité comprise entre 20 et 50%. Cela suggère une tendance vers une popularité modérée. Si l'on se concentre sur les morceaux qui ont une forte popularité (plus de 90%), on observe qu'ils sont très peu nombreux (environ 98 sur 89 741 sons).

Figure 1: Histogramme de la distribution de la popularité des morceaux

- **Diagramme en bâtons et diagramme circulaire** pour comparer le nombre de morceaux réalisés par un artiste à la popularité de cet artiste.

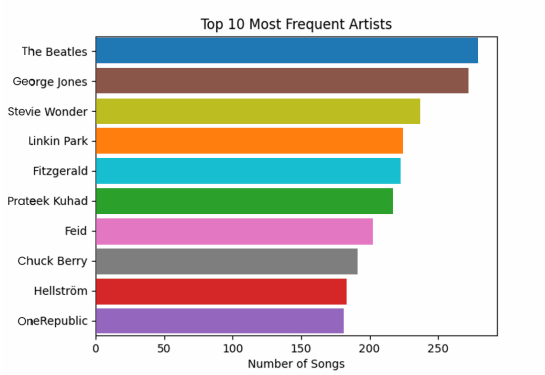


Figure 2: Classement des 10 premiers artistes ayant réalisé le plus de morceaux

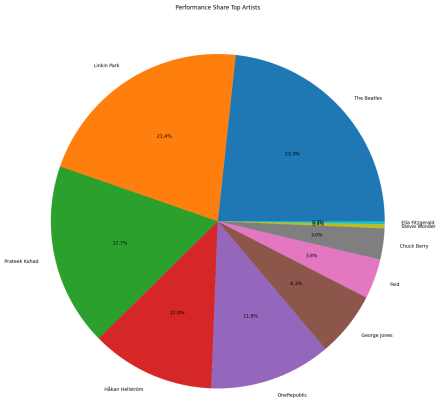


Figure 3: Popularité de des 10 premiers artistes ayant réalisé le plus de morceaux

On observe ici qu'il n'y a pas de forcément corrélation entre le nombre de morceaux réalisés par un artiste, et son taux de popularité. En effet, dans le cas de Stevie Wonder, classé 3ème du classement des artistes ayant réalisé le plus grand nombre de morceaux, sa popularité est inférieure à 1% par rapport aux 9 autres artistes.

- **Violin plots** pour comparer la popularité des 10 premiers genres musicaux.

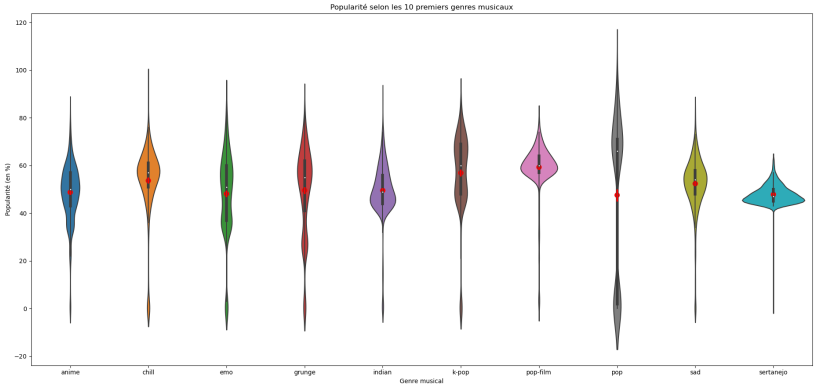


Figure 4: Popularité selon les 10 premiers genres musicaux

Genres avec une popularité moyenne élevée : Le violon est plus large en haut, ce qui indique qu'un grand nombre de morceaux dans les genres k-pop et pop-film sont populaires.

Genres avec une distribution très large : La pop, l'emo et le grunge ont des distributions très étalées, ce qui signifie qu'on retrouve, pour ces genres, à la fois des morceaux très populaires et d'autres beaucoup moins.

Genres plus homogènes : Le sertanejo ou encore l'anime ont une concentration plus resserrée, ce qui indique une certaine régularité dans la popularité des morceaux de ces genres.

Outliers : Certaines distributions ont des extrémités très longues (pop par exemple), ce qui suggère qu'il existe des morceaux extrêmement populaires (ou impopulaires), bien que rares.

Nous pouvons donc en déduire que certains genres présentent une popularité globalement plus élevée , tandis que d'autres montrent une grande hétérogénéité, voire une distribution plus concentrée. Cela suggère donc que la popularité ne dépend pas uniquement du genre, mais aussi de caractéristiques propres à chaque morceau. Ainsi, dans le cadre des recommandations, il est peu probable qu'un genre en soi soit un facteur déterminant : ce sont davantage des attributs audio précis comme le tempo, l'energy, etc.) qui pourraient expliquer la sélection d'un morceau. Par ailleurs, cette observation soutient l'idée que les recommandations Spotify se basent probablement plus sur la similarité audio entre morceaux que sur leur popularité globale d'un morceau lui-même.

- **Corrélation et Comparaisons** pour mieux comprendre le dataset.

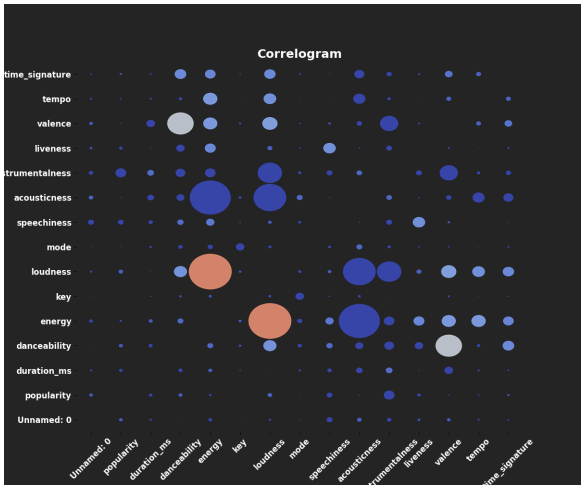


Figure 5: Heatmap de corrélation pour observer les relations linéaires entre variables numériques.

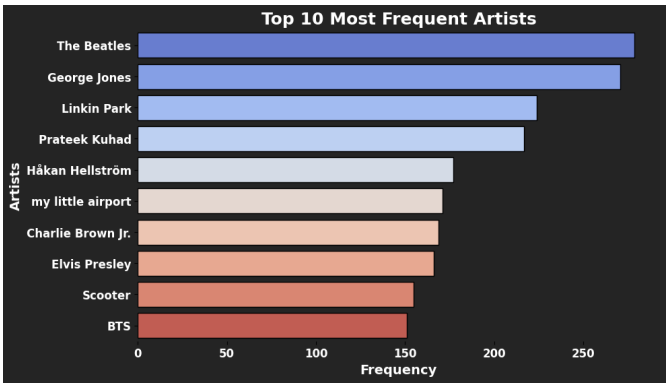


Figure 6: Barplot montrant les artistes les plus fréquents.

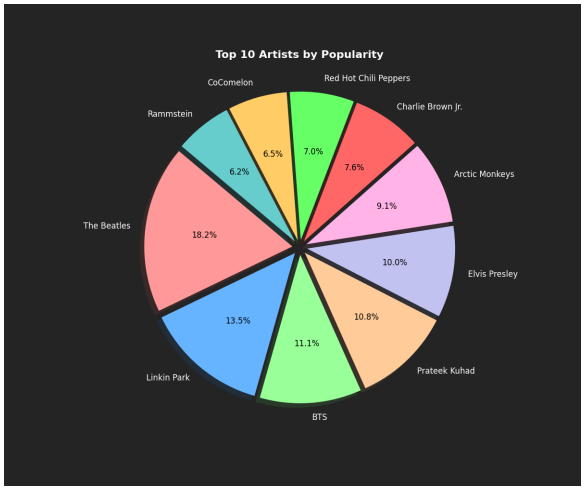


Figure 7: Diagramme circulaire illustrant la répartition de la popularité des morceaux selon les artistes les plus présents dans le dataset.

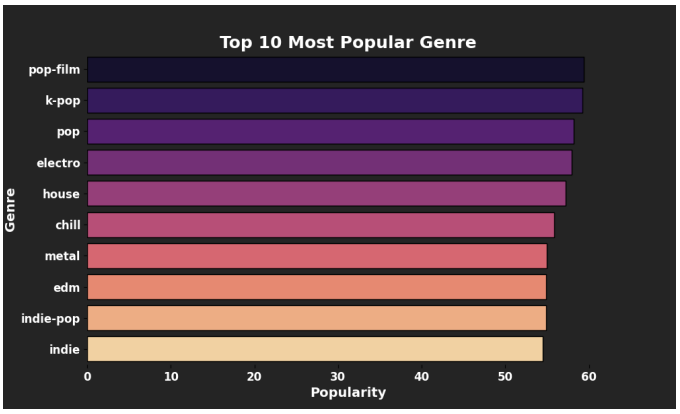


Figure 8: Barplot illustrant les genres les plus populaires dans le dataset.

Toutes ces visualisations nous ont permis de mieux comprendre les données et de poser des bases solides pour notre modèle de recommandation.

3.3. Création de nouvelles variables

Nous avons dans un premier temps créé une matrice de corrélation, ce qui nous a permis de relever les *features* dont l'analyse semblait la plus pertinente. Nous avons donc crée de nouvelles variables explicatives issues de notre dataset, qui reflètent des caractéristiques musicales importantes, plausibles comme critères dans les recommandations Spotify :

- **Mood_category** à partir de *valence*
- **Tempo_category** à partir de *tempo*

- **Energy_level** à partir de *energy* (faible, moyen, fort)
- **Danceability_level** à partir de *danceability* (faible, moyen, élevé)
- **Loudness_level** à partir de *loudness*

Pourquoi à t-on fait ces choix ?

Spotify semble utiliser des caractéristiques audio objectives (rythme, énergie, ambiance) pour proposer des morceaux similaires ou correspondant aux goûts de l'utilisateur. Les variables que nous avons créées résument ces propriétés acoustiques et sont donc des candidats logiques pour expliquer la sélection ou la recommandation d'un morceau.

Nous avons donc représenté les histogrammes suivants qui comparent la popularité d'un morceau à sa distribution pour chacunes des 3 catégories (au sein d'une même feature) que nous avons crée.

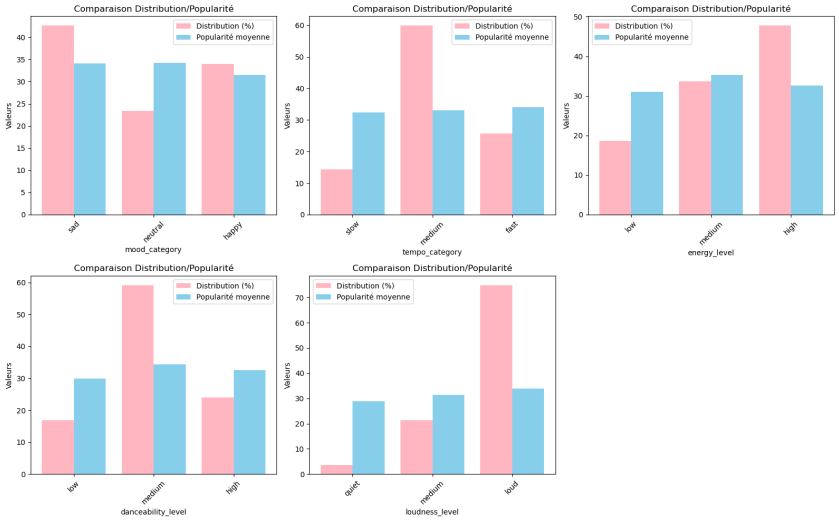


Figure 9: Histogramme de la distribution de la popularité des morceaux

Nous avons regroupé nos observations dans le tableau suivant :

Feature	**Populaire si...**	**Interprétation**
Mood	Sad	Les chansons émotionnelles plaisent, même si elles sont moins "joyeuses".
Danceability	Medium	Un morceau légèrement dansant est privilégié car polyvalent.
Energy	Medium	Un son ni trop calme, ni trop intense semble optimal pour l'écoute quotidienne.
Tempo	Fast	Les rythmes rapides attirent, surtout avec les tendances actuelles.
Loudness	Loud	Un volume élevé traduit un son "pro", plus présent dans les playlists populaires.

Table 1: Lien entre caractéristiques musicales et popularité

On peut donc en déduire que les morceaux à tendance rapide, relativement forts, modérément dansants et énergétiques sont les plus populaires, et donc potentiellement les plus susceptibles d'apparaître dans les recommandations Spotify.

4. Analyse par Régression Linéaire

Pour comprendre au mieux les différentes caractéristiques des morceaux et leurs liens entre leurs différents attributs, nous avons employé des méthodes de régressions linéaires. Dans un premier temps, nous avons exploré des modèles de régressions simple (Y X) puis, des modèles multiples pour comprendre des interactions plus complexes notamment avec les variables catégorielles (définit avec les *handcrafted features*). Notre objectif ici est d'utiliser la régression comme un outil exploratoire pour identifier certaines dépendances et leurs forces.

4.1. Régression Linéaire Simple et d'ordre 2 entre les variables explicatives

Tout d'abord, nous avons exploré les relations entre toutes les paires de variables explicatives. En particulier, par des modèles linéaires simples ($Y = \beta_0 + \beta_1 X$), et ensuite par des modèles polynomiaux (degré 2, à savoir $Y = \beta_0 + \beta_1 X + \beta_2 X^2$) pour identifier d'éventuelles relations non linéaires entre les variables. Par ailleurs, nous avons déterminer l'orientation optimale de chaque paire ($Y \sim X$ ou $X \sim Y$) qui maximise le coefficient R^2 ajusté, tout en vérifiant la significativité statistique des coefficients (via la p-value respective).

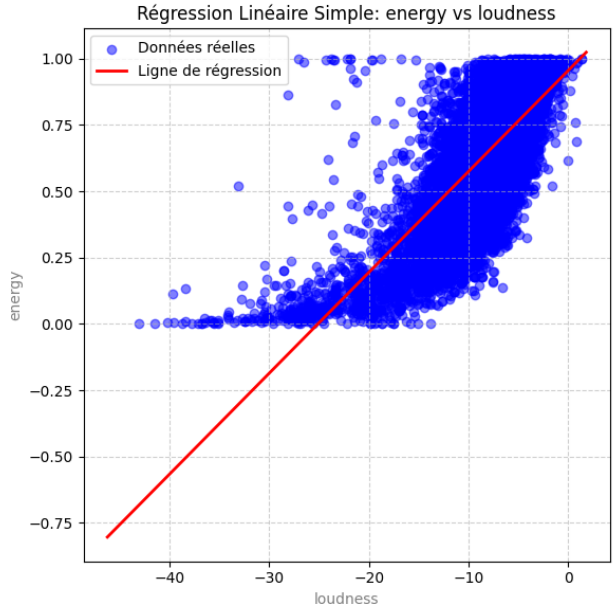


Figure 10: Exemple de régression simple entre `energy` et `loudness`. La courbe montre une tendance positive forte, légèrement curviligne.

Relation entre énergie et volume perçu : Par exemple, la relation entre l'énergie (`energy`) et le volume perçu (`loudness`) a été identifiée comme forte et positive. La figure 10 illustre cela. Le modèle simple a montré un R^2 ajusté de 0.58, indiquant qu'une part importante de la variance du volume peut être expliquée par l'énergie du morceau, et inversement.

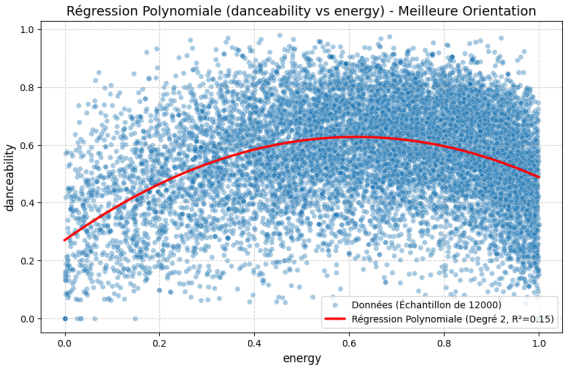


Figure 11: Exemple de régression polynomiale entre `danceability` et `energy`. La forme indique un lien parabolique entre les variables.

Relation entre `danceability` et `energy` :
Une relation non-linéaire intéressante a été observée entre la `danceability` et la `valence` (figure 11). Un modèle simple a permis d'obtenir un R^2 ajusté de 0.02 (l'orientation optimale) tandis qu'un modèle d'ordre 2 nous donne un R^2 de 0.15, suggérant que les morceaux les plus dansants sont souvent (mais pas systématiquement) ceux avec une énergie moyenne. À noter que certains morceaux avec une énergie élevée ou basse sont aussi très dansants. Pour autant, cela montre bien qu'il existe des relations non-linéaires intéressantes à identifier.

Ces différentes analyses nous permettent alors de nous rendre compte que des liens linéaires, mais aussi plus complexes comme polynomiales d'ordre 2 sont aussi à l'oeuvre entre les différentes variable ce qui nous encourageant donc à creusé la question (notamment avec les Handcrafted Features), tout en nous donnant alors une première cartographie des interdépendances entre les variables.

4.2. Modèles de Régression Multiple avec les Handcrafted Features

L'étape suivante a donc consisté en l'exploration de modèles plus complexes ($Y \sim X \cdot HF$) pour voir si les relations entre les variables numériques (qui paraissaient intéressantes à étudier d'après notre partie précédente) étaient modérées, liés par nos handcrafted features, ou si des interactions plus complexes entre différentes features amélioraient l'explication d'une variable cible Y. Nous avons principalement utilisé des modèles de la forme $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon$, où X_2 représente une handcrafted feature.

4.2.1 Interaction entre les caractéristiques audio et Handcrafted Features

Nous avons testé si l'effet d'une caractéristique audio numérique (comme la `valence`) sur la variable `loudness` différait selon les catégories d'une handcrafted feature (comme `energy_level`). La figure 12 montre la relation entre la `valence` et `loudness` pour chaque niveau de `energy_level`.

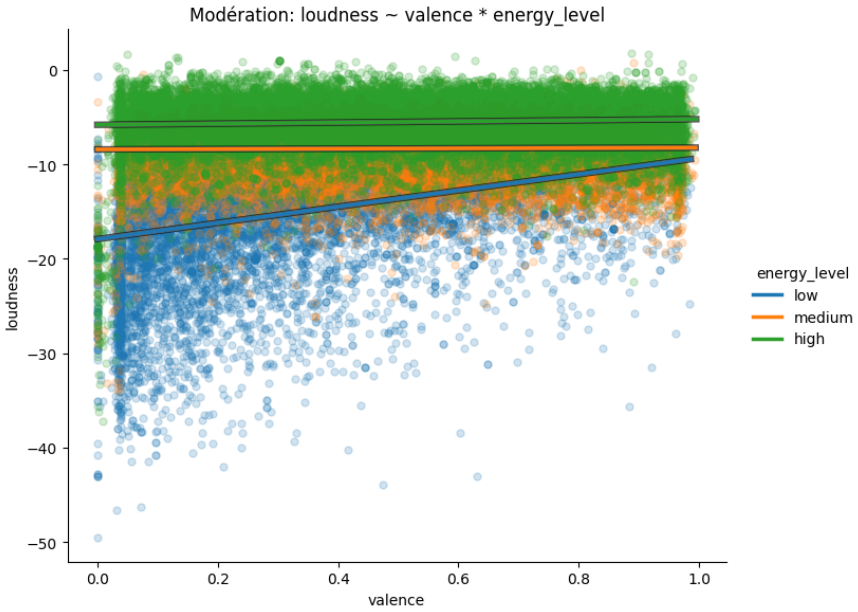


Figure 12: Interaction entre `energy_level` et `valence` pour prédire la variable `loudness`. Les lignes de couleurs distinctes représentent les différents niveaux d'énergie. On observe que la relation entre la valence et loudness est plus forte pour les morceaux ayant une haute énergie.

L'analyse statistique a révélé un terme d'interaction significatif (avec une p-value inférieure à 0.01). Le modèle complet (incluant l'interaction) a atteint un R^2 ajusté de 0.5225, ce qui représente une amélioration de 44% (ΔR^2_{adj} vs simple de 0.44) par rapport au modèle simple `loudness ~ valence`, et une amélioration de 50,16% (ΔR^2_{adj} vs principaux de 0.5016) par rapport à un modèle incluant seulement les effets principaux de `valence` et `energy_level` (évalué par `loudness ~ valence + energy_level`). Cette amélioration importante suggère que l'interaction est forte entre ces 3 variables.

4.2.2 Modération des Relations entre Caractéristiques Audio

Nous avions également examiné si les relations linéaires simples que nous avions identifiées précédemment étaient modérées par nos handcrafted features. Par exemple, nous avons testé si la relation entre `energy` et `loudness` variait en fonction de la `tempo_category`.

Les résultats ont montré que l'interaction n'était que très légèrement meilleur indiquant une relation entre 3 variables, mais suffisamment significatif par rapport à la relation simple (ΔR^2_{adj} vs simple était de 0.018). Par contre d'autre variable comme `instrumentalness` et `energy_level` explique mieux la variance de `loudness` que le modèle simple avec 41% (ΔR^2_{adj} vs simple était de 0.41)

4.3. Discussion des Résultats de Régression

Ainsi, les analyses par régression nous ont permis de quantifier plusieurs dépendances. Et bien que les variables seules n'expliquent qu'une fraction des liens entre ces dernières, les interactions révélées sont toutes fois instructives pour l'étude de nos données et la recommandation de musiques en fonction de leurs variables. En effet, elles suggèrent que les relations entre les attributs musicaux ne sont souvent pas simplement et systématiquement linéaire, mais que l'effet d'une caractéristique peut dépendre du niveau d'une autre.

Par exemple, le fait que le volume moyen (loudness) d'un morceau pourrait être un levier que les systèmes de recommandation exploitent implicitement ou explicitement pour suggérer des morceaux qui, bien que possiblement différents sur une dimension, partagent une relation contextuelle similaire sur une autre (comme loudness et instrumentalness avec l'energy level). Il est aussi clair que les handcrafted features se sont avérées utiles pour segmenter les données et révéler des liens importants et complexes entre certaines variables.

De plus, il faut noter que la causalité ne peut être expliquée par la corrélation, par contre les dépendances et interactions observées fournissent des pistes essentielles pour comprendre la relation entre les différentes variables et comment différents types de sons pourraient être perçus comme "proches" ou "cohérents" dans une séquence de lecture.

5. Conclusion

Finalement, nous avons exploré les propriétés des morceaux et les dynamiques de recommandation propres à Spotify. Grâce à l'analyse de plus de 114000 morceaux issus de différents genres, nous avons pu mettre en lumière l'importance de certaines variables audio (énergie, valence, danceability, tempo, loudness) dans la recommandation des titres. De plus, la création de *handcrafted features* nous a permis de résumer efficacement des propriétés musicales complexes et de mieux comprendre leur impact sur la perception et le succès d'un morceau. Les analyses de corrélation et de régression linéaire, simples ou multiples, ont montré que les relations entre les attributs musicaux ne sont pas toujours linéaires et que ces interactions jouent un rôle significatif dans la structure des recommandations. Nous avons également mis en évidence que c'est avant tout la combinaison de caractéristiques audio spécifiques qui déterminent la probabilité d'apparition d'un morceau dans les suggestions et non leur popularité seule dans le cas pratique.

Les Limites : Malgré la richesse du dataset, notre analyse reste très superficielle, limitée par la nature statique des données (pas d'accès aux historiques d'écoute réels, ni aux algorithmes internes de Spotify). Par ailleurs, l'approche statistique utilisée ici ne permet pas de démontrer un lien direct entre caractéristiques musicales et popularité, mais met en évidence des tendances et des corrélations utiles.

Perspectives d'améliorations : Pour aller plus loin, il serait pertinent d'intégrer des données supplémentaires (comme les playlists, les transitions réelles entre morceaux, ou le contexte d'écoute), d'appliquer des méthodes de machine learning avancées (forêts aléatoires, ACP) pour explorer et détecter de nouveaux clusters de morceaux similaires. Enfin, un axe d'amélioration pourrait consister à analyser la dimension temporelle (évolution des goûts, tendances musicales), qui n'est pas disponible avec notre dataset actuel, permettant alors d'affiner la modélisation des préférences utilisateur.

En conclusion, ce projet nous a permis de mieux comprendre les liens sous-jacents aux recommandations musicales, tout en développant des compétences en analyse de données, en modélisation statistique et en data storytelling. Les résultats obtenus préparent la voie à de possibles futurs études sur le domaine de la musique en science des données.

References

- Analyse d'un dataset de vidéos de trafic routier (Medium)
- Régression linéaire en PyTorch (Kaggle Notebook)
- Guide de la régression logistique avec le dataset Titanic (Medium)
- Projet Machine Learning Hackathon Bradesco (GitHub)
- Analyse et prédiction des ventes Black Friday (Kaggle Notebook)