**IMT Atlantique**
Technopôle de Brest-Iroise - CS 83818
29238 Brest Cedex 3
Téléphone : +33 (0)2 29 00 13 04
Télécopie : +33 (0)2 29 00 10 12
URL : **www.imt-atlantique.fr**

**DASCI Fundamental of Machine Learning**
FISE A2

# Challenge 1 : Spotify

Achraf ESSALEH
Sara ELBARI
Nada ALEIAN
Eva LANSALOT
Houda DAOUAIRI
Kalis KRAÏFI

Date d'édition : 17 octobre 2025
Version : 1.5

**IMT Atlantique**
Bretagne-Pays de la Loire
École Mines-Télécom

# Sommaire

# 1. Introduction

Pour ce travail, nous avons étudié un dataset fourni par Spotify, contenant les 100 morceaux les plus populaires par année de 2010 à 2022, soit 2 300 morceaux. Chaque morceau est décrit par 23 variables, incluant des informations sur la playlist et l'année, des caractéristiques audio (comme danceability, energy, valence, tempo...), l'album, ainsi que des informations sur l'artiste (nom, genre, popularité).
L'objectif principal de cette analyse était de comprendre la structure et les relations au sein de ce jeu de données, en appliquant des méthodes d'exploration de données, des techniques factorielles et au moins un algorithme de clustering, afin d'identifier des groupes de morceaux présentant des similarités. Cette étude permet également de mettre en pratique les notions d'ingénierie des fonctionnalités et d'interprétation des résultats dans un contexte d'apprentissage non supervisé.

## Structure du Répertoire de Projet

Afin de garantir une organisation professionnelle de l'analyse, la structure du répertoire suit les meilleures pratiques de la science des données :

```
Spotify-sHits-Unsupervised_Learning_Challenge/
|
|-- data/
|    |-- raw/              # Dataset original
|    |-- processed/        # Données nettoyées et prêtes pour la modélisation
|
|-- notebooks/ # Notebooks Jupyter (Exploration, Prétraitement, Modélisation)
|    |-- 01_EDA.ipynb
|    |-- 02_Preprocessing.ipynb
|    |-- 03_Clustering.ipynb
|    |-- 04_Analysis.ipynb
|
|-- src/  # Modules et scripts Python réutilisables (fonctions/clustering)
|    |-- data_preprocessing.py
|    |-- feature_engineering.py
|    |-- clustering_models.py
|    |-- visualization.py
|
|-- reports/              #  figures et présentations
|    |-- figures/
|    |-- slides/
|
|-- requirements.txt      # Liste des dépendances Python
|-- README.md             # Vue d'ensemble du projet
```

# 2. Analyse exploratoire des données : EDA

Le jeu de données comprend 2300 lignes et 23 colonnes. Il est donc suffisamment large pour faire émerger des tendances pertinentes.

Parmi les 23 variables, 13 ne présentent qu'une seule valeur manquante, ce qui correspond à un taux de données manquantes très faible. Après vérification, ces valeurs manquantes concernent une seule et même ligne, contenant 13 valeurs nulles sur 23. Cette observation représentant un seul échantillon sur 2 300, elle peut être supprimée sans impact significatif sur l'analyse.

De plus, le jeu de données ne contient aucun doublon de morceaux, ce qui confirme sa qualité et sa cohérence.

Ainsi, le jeu de données final est propre, sans doublons ni valeurs manquantes significatives, et prêt pour les étapes suivantes de l'analyse exploratoire.

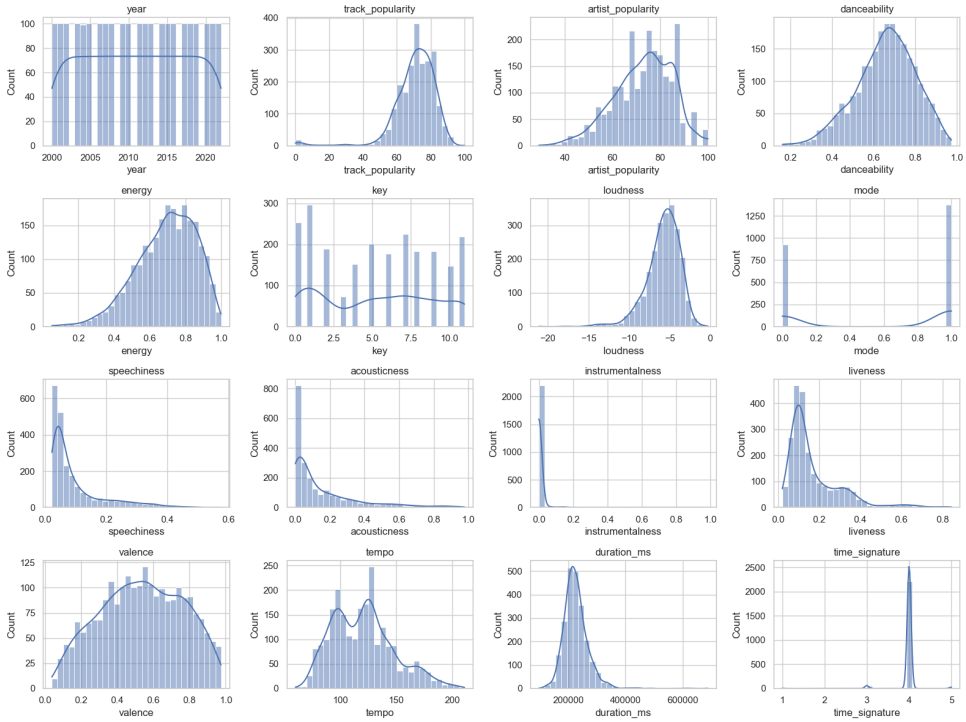


FIGURE 1 – Distribution de chaque feature numérique

La plupart des variables ne suivent pas une distribution normale. Certaines sont fortement asymétriques (par exemple acousticness ou speechiness), tandis que d'autres sont presque normales (energy, valence, loudness). Cette non-normalité, bien que naturelle pour des données musicales, impose une standardisation préalable avant toute analyse multivariée. Elle montre également que le dataset est assez homogène (notamment sur tempo, duration, danceability), ce qui pourra réduire la capacité du clustering à séparer clairement les styles, d'où l'intérêt d'une réduction de dimension (PCA) avant la classification.

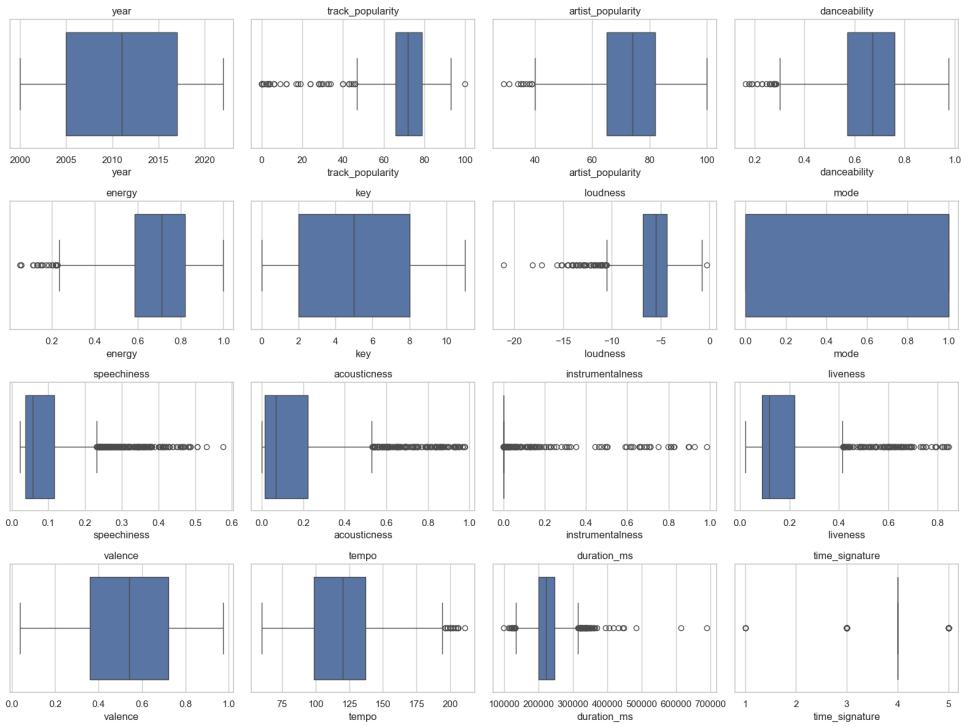


FIGURE 2 – Boxplots pour visualiser les outliers de chaque feature

Plusieurs variables présentent des outliers à longue traîne, notamment instrumentalness, acousticness, speechiness, liveness et duration_ms. Ces valeurs extrêmes reflètent vraisemblablement la diversité des morceaux plutôt que des erreurs. Track_popularity et artist_popularity contiennent également quelques outliers, mais ils restent cohérents avec les données musicales réelles. Les valeurs les plus extrêmes semblent traduire des variations légitimes de style et de production, et non du bruit ou des anomalies. De plus, loudness ne présente que quelques outliers sur le bas de l'échelle (morceaux plus calmes), ce qui correspond à des différences musicales plausibles, sans constituer d'erreurs.

L'analyse des variables qualitatives et temporelles apporte plusieurs informations importantes. Le dataset contient 23 valeurs uniques pour les playlists, ce qui correspond aux années 2000 à 2022, indiquant que les chansons sont regroupées par playlist annuelle. Chaque track-id et track-name est presque unique, garantissant une bonne intégrité des données, tandis que certaines chansons partagent des albums (1663 albums au total), ce qui est normal. Avec 891 artistes différents, on observe à la fois diversité et la possibilité que le genre ou la popularité des artistes constitue un critère de regroupement pertinent. Comme la variable artist_genres compte 698 valeurs uniques, il est logique d'envisager un regroupement par genre.

L'analyse des genres montre que pop et dancepop, pop sont les plus fréquents, avec plus de 120 occurrences, tandis que les huit autres genres ne dépassent pas 40 occurrences. Cette distribution déséquilibrée pourrait biaiser un modèle, et un regroupement des genres rares pourrait permettre d'obtenir des insights plus significatifs.

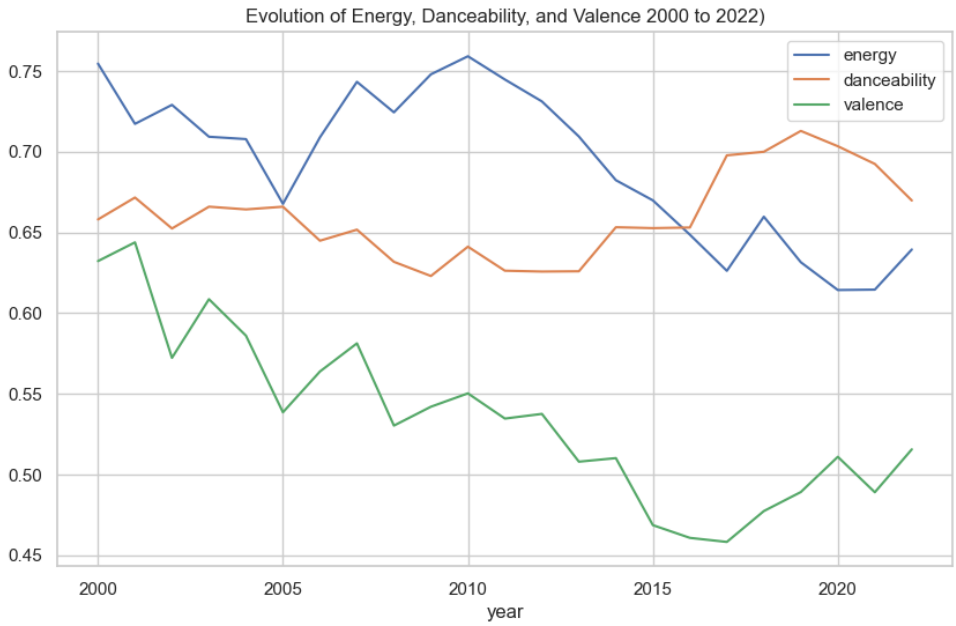


FIGURE 3 – Evolution de la danceability, energy et valence dans les playlists 2000-2022

L'évolution temporelle des caractéristiques musicales révèle des tendances intéressantes : l'energy montre une légère diminution depuis 2015, suggérant que les chansons deviennent moins intenses sur certains aspects, la danceability a augmenté, indiquant que les morceaux modernes sont plus rythmiques, et la valence a baissé, ce qui traduit une tendance vers des chansons plus mélancoliques au fil du temps.

# 3. Preprocessing

L'étape de prétraitement a pour but de rendre les données exploitables par les modèles d'apprentissage automatique. Les données brutes initialement extraites du fichier CSV contenaient des colonnes hétérogènes (textuelles et numériques), des valeurs manquantes identifiées lors de la phase EDA, ainsi que des échelles différentes entre variables. Ces irrégularités peuvent fausser les résultats d'un modèle non supervisé comme K-Means. Le prétraitement vise donc à nettoyer, homogénéiser et normaliser les données.

Le jeu de données a été scindé en deux sous-ensembles :

- df_meta : contient les métadonnées descriptives (titre, artiste, genre, année, etc.), conservées uniquement pour l'interprétation et la visualisation ;
- df_model / df_model_proc : contient les variables numériques exploitables pour la modélisation (par exemple : danceability, energy, valence, tempo, loudness, acousticness, etc.).

Cette séparation permet de travailler uniquement sur les caractéristiques pertinentes pour les calculs statistiques, tout en gardant les informations descriptives pour l'analyse qualitative.

Plusieurs opérations de nettoyage ont été réalisées :

- Gestion des valeurs manquantes :
  - Les lignes contenant trop de valeurs nulles ont été supprimées.
- Filtrage des colonnes inutiles :
  - Les colonnes purement textuelles ou identifiants uniques (par ex. track_id) ont été retirées du DataFrame utilisé pour la modélisation.
  - Les colonnes catégorielles (comme genre_grouped2) ont été conservées dans df_meta pour l'analyse ultérieure, mais exclues de df_model_proc.

Ce travail garantit que le modèle ne traite que des données cohérentes et homogènes.

Lors de l'analyse exploratoire, il est apparu que la variable genre comportait un très grand nombre de catégories, souvent redondantes ou déséquilibrées : certains genres (comme "pop" ou "dancepop") comptaient plusieurs centaines de morceaux, tandis que d'autres (comme "edm" ou "rock") n'en contenaient qu'une poignée.
Un tel déséquilibre peut biaiser les analyses statistiques et rendre la visualisation confuse. Pour remédier à cela, un travail de regroupement sémantique a été effectué :

- Les morceaux ayant souvent plusieurs genres associés, on suppose qu'ils sont rangés par ordre de cohérence et ainsi retenons uniquement le premier qui serait le plus adéquat.
- Les genres trop rares ont été regroupés sous une catégorie "Autres" afin d'éviter une sur-segmentation.

Ce regroupement a donné lieu à une nouvelle variable *genre_grouped2* qui représente la version consolidée et équilibrée du champ genre. Cette transformation a permis :

- de réduire le nombre de classes à un niveau plus exploitable,
- d'améliorer la lisibilité des graphiques PCA,
- et d'éviter qu'un genre surreprésenté n'influence trop le modèle.

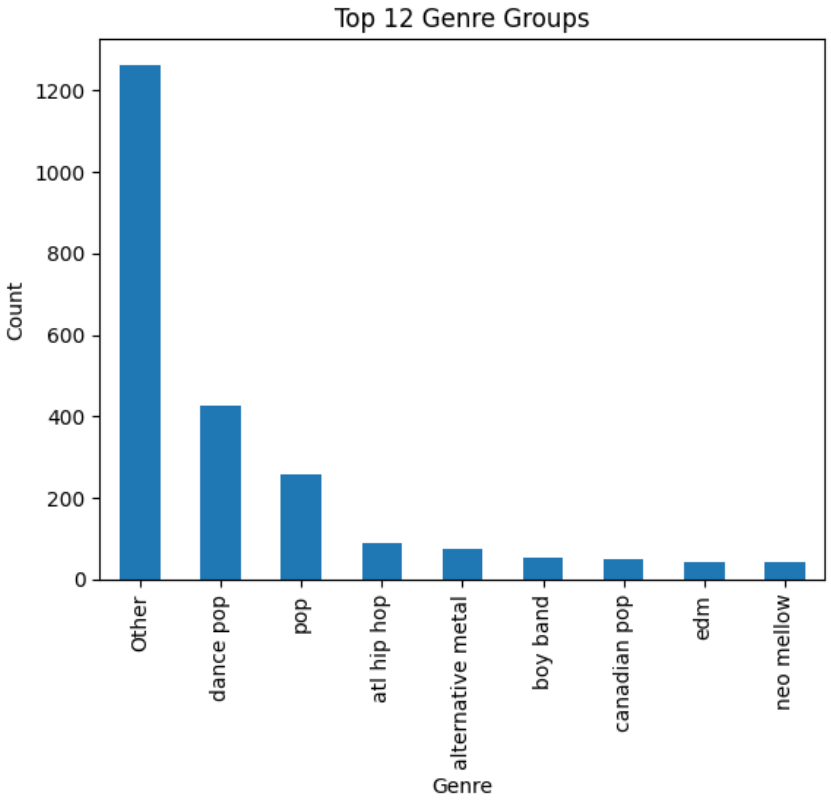


FIGURE 4 – Déséquilibre classes : genres musicaux

Les variables musicales (par exemple tempo, energy, loudness, valence, etc.) ne sont pas exprimées sur la même échelle :

- certaines varient entre 0 et 1 (ex. energy, valence),
- d'autres atteignent des valeurs élevées (ex. tempo en battements par minute, loudness en décibels négatifs).

Afin d'uniformiser les échelles, une normalisation standardisée a été appliquée à l'aide de la classe StandardScaler du module scikit-learn *scaler.fit_ transform()*. Cette opération :

- centre les données (moyenne = 0),
- réduit leur variance (écart-type = 1),
- garantit que chaque variable contribue équitablement à la mesure de distance utilisée dans les algorithmes non supervisés (comme la PCA ou K-Means)

Après transformation :

- Le tableau X_scaled (données prêtes pour le modèle) a été obtenu sous forme de matrice NumPy
- Ce tableau contient exactement le même nombre de lignes que le DataFrame original après nettoyage, garantissant la cohérence entre les observations et les métadonnées.

Enfin, les données prétraitées ont été sauvegardées dans un nouveau fichier CSV (par exemple playlist_2000to2022_preprocessed_v2.csv), afin d'être réutilisées dans les étapes ultérieures du projet (réduction de dimension, clustering, visualisation).

Le prétraitement a permis :

- d'éliminer les incohérences du jeu de données,
- d'uniformiser les variables par normalisation,
- de préparer une base propre, stable et interprétable,
- et de conserver la structure logique entre les caractéristiques numériques et les métadonnées descriptives.

Afin de visualiser la structure des données et d'éliminer la redondance entre variables, une Analyse en Composantes Principales (PCA) a été réalisée. La PCA permet de projeter les données initiales sur un espace de plus faible dimension, tout en conservant le maximum de variance (information) possible.

L'analyse du pourcentage de variance expliquée par les composantes principales a montré que :

- les deux premières composantes (PC1 et PC2) expliquent une part significative de la variance totale (souvent entre 40 % et 60 % selon le jeu de données) ;
- les composantes suivantes apportent une information plus fine mais redondante.

Ces deux premières composantes ont donc été utilisées pour la visualisation des morceaux dans l'espace PCA, en colorant les points selon leur genre musical.

Le graphique obtenu permet de visualiser la proximité entre genres musicaux : les genres partageant des caractéristiques acoustiques similaires (énergie, rythme, valence...) ont tendance à se regrouper naturellement dans l'espace PCA. Par exemple, les genres électroniques et dance se positionnent souvent dans une même région du plan, tandis que les genres classiques ou acoustiques se regroupent ailleurs.

Les coefficients de corrélation entre les variables initiales et les composantes principales permettent d'interpréter les axes :

- PC1 semble corrélé avec les variables liées à l'énergie et à la valence : il peut être interprété comme un axe allant de la musique calme et mélancolique à la musique rythmée et joyeuse ;
- PC2 semble davantage associé à la dansabilité et à la présence rythmique, séparant les morceaux acoustiques des morceaux fortement produits et destinés à la danse.

Cette interprétation facilite la lecture des graphes et servira de base pour le clustering (section suivante du rapport).

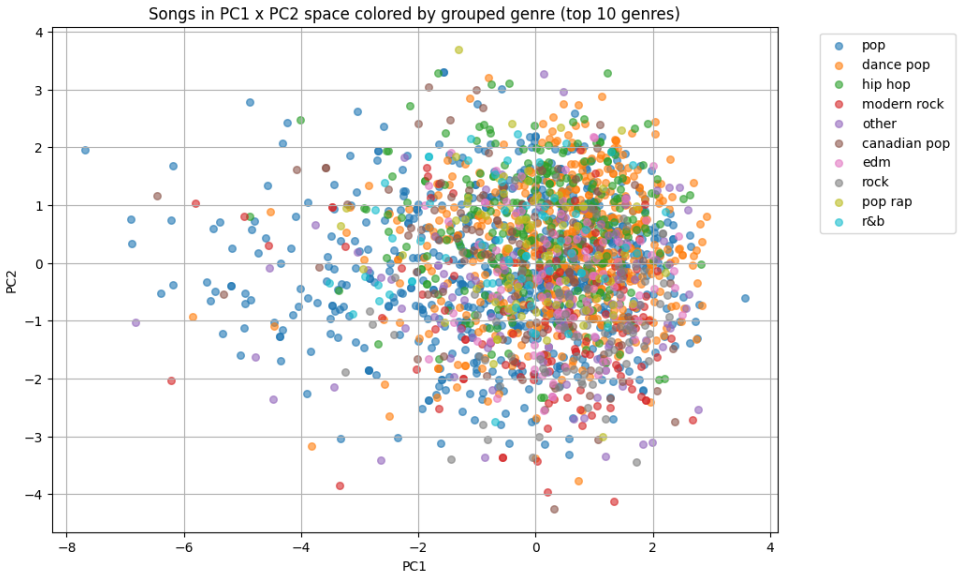


FIGURE 5 – PCA1 et PCA2

Cette phase de prétraitement a permis d'obtenir une base de données propre, normalisée et réduite en dimension, prête à être utilisée pour le clustering non supervisé (K-Means). La PCA a mis en évidence des régularités dans la structure acoustique des morceaux, montrant que les caractéristiques musicales suffisent déjà à distinguer certains styles. Le travail de clustering permettra d'examiner si ces regroupements observés visuellement se confirment mathématiquement.

# 4. Clustering

Après la phase de prétraitement, plusieurs méthodes de clustering non supervisé ont été appliquées afin d'identifier d'éventuelles structures latentes dans les données musicales. L'objectif est de regrouper des morceaux similaires selon leurs caractéristiques audio, sans utiliser d'étiquettes externes (comme le genre).
Le principal algorithme utilisé est K-Means, évalué à l'aide de plusieurs métriques telles que :

- L'inertie (ou somme des distances intra-cluster) afin d'observer la courbe du coude (elbow method) ;
- Le Silhouette Score, qui mesure la cohésion et la séparation des clusters ;
- Et d'autres indicateurs tels que les indices de Davies-Bouldin et Calinski-Harabasz, souvent utilisés pour valider la qualité d'un clustering.

L'ensemble des figures et codes correspondants a été enregistré dans le répertoire /reports et les scripts dans /src.

Deux principales visualisations ont été réalisées :

- La courbe du coude (Elbow Curve) représentant la décroissance de l'inertie en fonction de k ;

- La courbe du Silhouette Score en fonction de k.

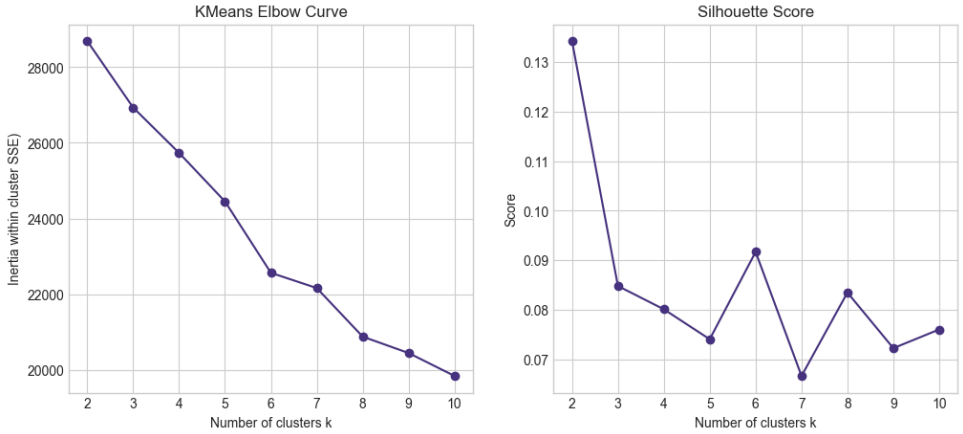


FIGURE 6 – Elbow Curve et Silhouette

Les observations sont les suivantes :

- Une chute significative de l'inertie est visible entre k = 2 et k = 6 ;
- Le point de coude semble se situer entre k = 6 et k = 7, indiquant un bon compromis entre compacité et séparation ;
- Le Silhouette Score atteint un maximum à k = 2, puis diminue brusquement pour k = 3 à 5, avec un léger rebond à k = 6.

Malgré tout, les scores restent globalement faibles, ce qui suggère que les clusters sont faiblement distincts et que les observations présentent une forte mélangeabilité dans l'espace PCA. Autrement dit, les chansons ont des caractéristiques proches, ne formant pas de groupes naturellement bien séparés.

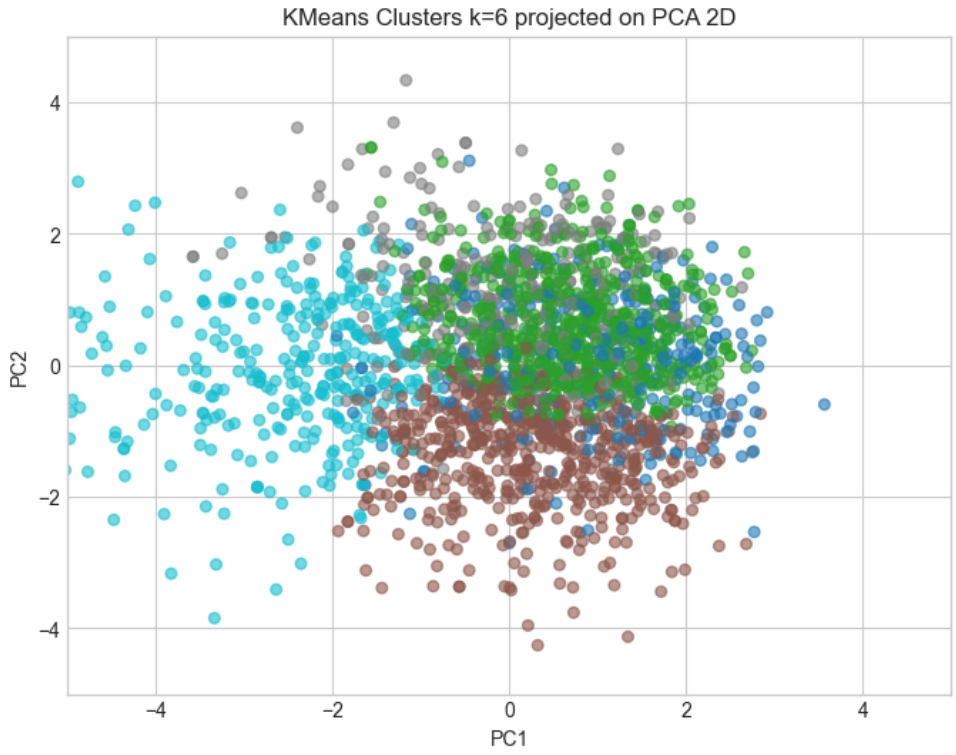


FIGURE 7 – Graphe K-Means K=6 clusters

L'analyse visuelle montre un chevauchement important entre les clusters, particulièrement autour du centre des axes principaux. Cela indique que les frontières entre groupes sont floues, et qu'il n'existe pas de séparation nette entre catégories sonores.
Une explication probable est que la PCA, utilisée pour la réduction de dimension, compresse la variance sur seulement deux composantes principales. De nombreuses séparations qui existent dans l'espace original à plus haute dimension peuvent être perdues lors de cette projection.
Ainsi, même si des regroupements existent dans l'espace complet, ils ne sont pas visibles dans le plan bidimensionnel.

Les faibles scores de silhouette et la forte superposition observée confirment que le dataset ne présente pas de structures de clusters bien marquées selon les seules caractéristiques audio. De plus, les genres musicaux apparaissent dans presque tous les clusters : cela signifie que le genre musical n'est pas strictement corrélé aux caractéristiques acoustiques. En effet, des chansons appartenant à des genres différents peuvent partager des propriétés sonores similaires (par exemple tempo, énergie ou tonalité).
Pour approfondir l'analyse, des techniques complémentaires ont été explorées, notamment le t-SNE (t-distributed Stochastic Neighbor Embedding). Cette méthode de projection non linéaire peut révéler des structures cachées que la PCA ne met pas en évidence. Les résultats du t-SNE ont confirmé que le jeu de données ne contient pas de catégories musicales nettement distinctes sur la base des seules caractéristiques audio.
Enfin, une moyenne des valeurs de chaque variable pour chaque cluster a été calculée

afin d'obtenir un profil type des chansons appartenant à chaque groupe. Cette analyse descriptive permet d'interpréter les clusters d'un point de vue musical, même en l'absence de séparation nette.

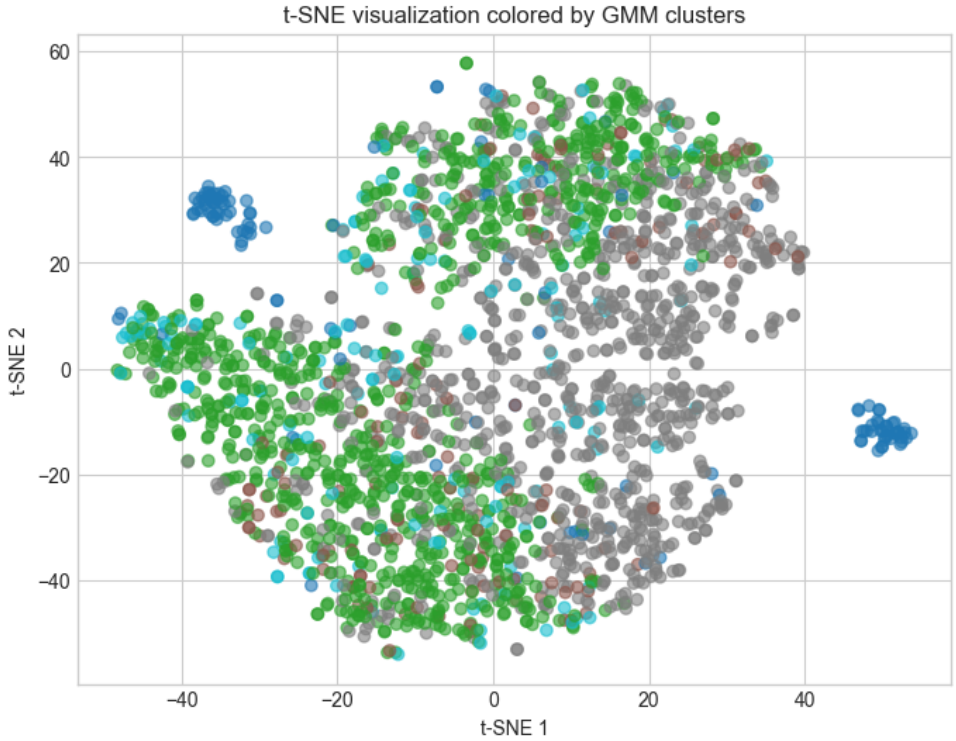


FIGURE 8 – Graphe T-SNE

# 5. Conclusion

Ce projet valide l'ensemble de la démarche non supervisée, de la préparation des données à l'interprétation. Nous avons réussi à traiter les données (Nettoyage, **Macro-Mapping de Genres**, transformation log) pour ensuite les simplifier via l'**ACP**, identifiant les axes majeurs de variation .

## 5.1. Résultats Clés du Clustering

Notre analyse du **K-Means** et l'évaluation par **GMM/t-SNE** nous amènent à une conclusion fondamentale : la musique forme un **continuum** de caractéristiques. Le score de Silhouette faible ($\approx 0.07$) le confirme, montrant que les frontières entre genres sont floues.

Malgré cela, nous avons isolé **5** archétypes sonores distincts, tels que :

- L'**Archétype Dance/Pop** (haute **Danceability**), qui correspond aux titres les plus récents et populaires (**Validation Externe**).
- L'**Archétype Spoken Word** (haute **Speechiness**), clairement séparé.
- L'**Archétype Acoustique** (basse **Energy**, haute **Acousticness**), souvent associé aux titres plus anciens.

L'utilisation d'outils comme le **GMM** et le **t-SNE** nous a permis comprendre leur **nature continue**.

## 5.2. Ouverture et Perspectives : Si la Contrainte Non Supervisée Était Levée

Si nous n'étions pas limités au non supervisé, la suite logique de ce travail serait de passer à une approche supervisée pour des applications telles que :

- **Prédiction de Popularité (Régression) :** Utiliser les caractéristiques audio (et les composantes PCA) comme variables prédictives pour estimer la **track_popularity** ou le **track_popularity** d'un nouveau morceau.
- **Classification de Genre (Classification) :** Utiliser le **genre_grouped2** comme étiquette cible pour entraîner un modèle (avec le **Random Forest** par exemple ) à classer un nouveau titre uniquement à partir de ses caractéristiques audio. Cela permettrait d'automatiser le taguage musical de manière plus précise que les tags manuels de Spotify.

OUR WORLDWIDE PARTNERS UNIVERSITIES - DOUBLE DEGREE AGREEMENTS

3 CAMPUS, 1 SITE


IMT Atlantique Bretagne–Pays de la Loire – **http://www.imt-atlantique.fr/**

**Campus de Brest**
Technopôle Brest-Iroise
CS 83818
29238 Brest Cedex 3
France
T +33 (0)2 29 00 11 11
F +33 (0)2 29 00 10 00

**Campus de Nantes**
4, rue Alfred Kastler
CS 20722
44307 Nantes Cedex 3
France
T +33 (0)2 51 85 81 00
F +33 (0)2 99 12 70 08

**Campus de Rennes**
2, rue de la Châtaigneraie
CS 17607
35576 Cesson Sévigné Cedex
France
T +33 (0)2 99 12 70 00
F +33 (0)2 51 85 81 99

**Site de Toulouse**
10, avenue Édouard Belin
BP 44004
31028 Toulouse Cedex 04
France
T +33 (0)5 61 33 83 65


**IMT Atlantique**
Bretagne-Pays de la Loire
École Mines-Télécom



Imprimé à IMT Atlantique
Dépôt légal : Septembre 2017
ISSN : 2556-5060