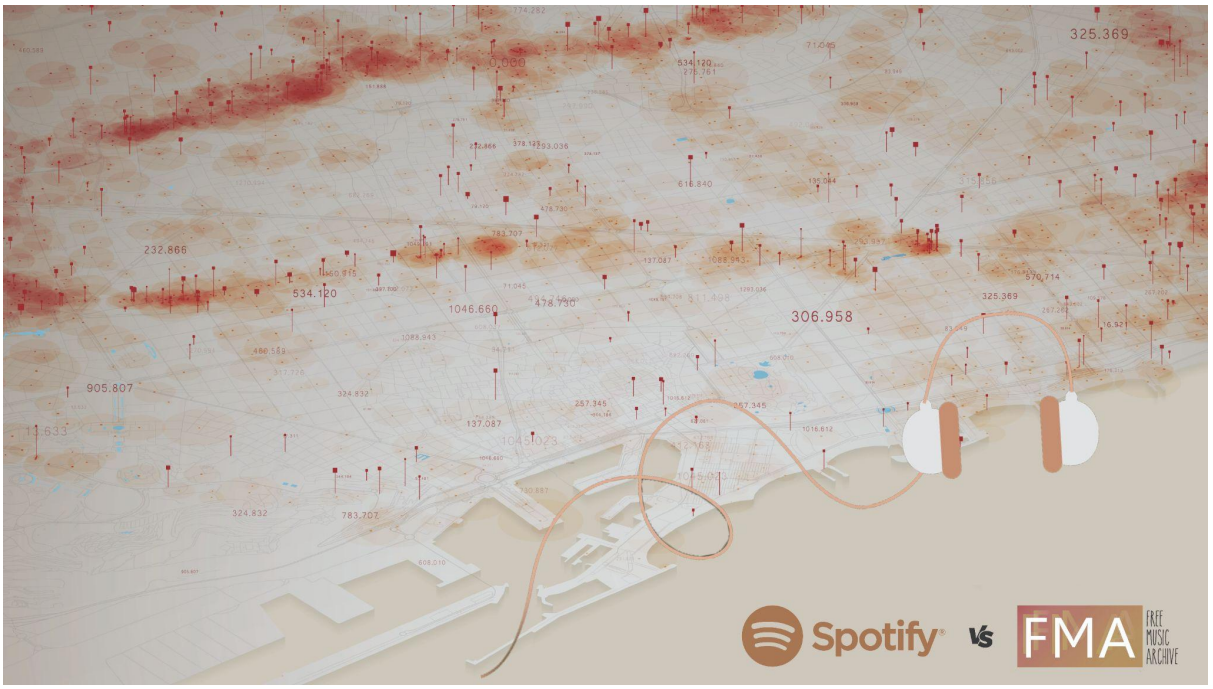
Mélodie FLEURY
Chaima HASDI
Fosio ULUTUIPALELEI

PROJET DE GÉOSTATISTIQUE

Quels sont les facteurs contribuant à la popularité d'une musique considérée comme un hit comparés à une chanson amateur selon différentes échelles spatiales ?



Enseignants : Paul CHAPRON, Yann MENEROUX, Juste RIMBAULT

ING3 Géo Data Science - Décembre 2023

SOMMAIRE

Introduction

A la fin des années 2000, la numérisation de l'industrie musicale s'intensifie avec l'émergence de plateformes d'écoutes payantes telles que Spotify ou gratuites telles que FMA (Free Music Archive). Alors, il est aujourd'hui possible de récolter de nombreuses informations notamment spatiales concernant le comportement des utilisateurs vis-à-vis des chansons hébergées sur ce type d'application.
Il est alors pertinent de se demander "quels sont les facteurs participant à la popularité d'une musique considérée comme un hit comparé à une musique amateur selon différentes échelles spatiales ?"
Pour répondre à cette question, il est décrit dans une première partie les principales caractéristiques de nos deux jeux de données. Puis, dans une seconde partie, les analyses statistiques sont illustrées avant de terminer dans une dernière partie par les analyses spatiales de ces mêmes jeux de données.

I/ Présentation des données et leurs (géo-)traitements

1.1. Présentation générale des jeux de données

Comme déjà dit dans l'introduction, nous avons deux jeux de données distincts portant sur des chansons avec leurs caractéristiques associées.
D'une part, il y a la table csv des 50 meilleures titres Spotify dans 70 pays différents appelée "universal_top_spotify_song" qui est mise à jour quotidiennement grâce à l'API de Spotify. Ces données sont disponibles librement sur le site *kaggle.com* [1]. La version utilisée pour les analyses date du 18 au 28 octobre 2023. La table regroupe 25 attributs tels que le nom et l'artiste de la chanson mais également sa place dans le classement du pays ou bien dans le classement mondial ainsi que des caractéristiques musicales qui seront détaillées plus loin.
D'autre part, il y a le jeu de données constitué de 106 574 chansons publiées sur FMA (Free Music Archive). FMA est une plateforme gratuite et ouverte où il est possible de publier ses musiques et de les réutiliser ensuite pour un usage personnel ou professionnel. Les données sont disponibles sur le répertoire github du chercheur Mr Michaël Defferrard [2] et elles datent d'Avril 2017. Le dataset est constitué de différentes tables csv (voir Annexe 1). Il existe également un site internet [3].

1.2. Sélection des attributs

Après avoir décrit grossièrement les données qui vont être utilisées dans cette étude, il est nécessaire de faire un tri dans les attributs afin de sélectionner les plus pertinents pour répondre à la problématique. Pour cela, une recherche bibliographique de chaque attribut a été effectuée. En effet, concernant les données de Spotify, le site contient les métadonnées du jeu de données contrairement au répertoire github où sont présentes les données de FMA.
Finalement, la sélection s'est faite sur les attributs seulement de la table "universal_top_spotify_song" pour les données Spotify et elle s'est faite à la fois sur les

tables et les attributs pour le jeu de données FMA. Un schéma résume cette sélection et leur type associé (Annexe 2) et un tableau donne les définitions de chaque attribut sélectionné (Annexe 3). A titre informatif, cette dernière table ne définit pas deux fois le même attribut s'il est présent dans 2 tables différentes. De plus, un dictionnaire des métadonnées des attributs est disponible sur le répertoire github [4].

1.3. Les jointures

Les données seules de nos deux jeux de données ne sont pas suffisantes pour répondre à la problématique. Il faut joindre d'autres données aux tables, c'est pour cela qu'il faut effectuer des jointures.
D'abord, concernant la table "universal_top_spotify_song", on peut compter 2 jointures : une avec la table csv des codes ISO de tous les pays du monde pour ajouter le nom des pays dans la table du Top 50 de Spotify [5] et une seconde avec la table csv des IDH (Indice de Développement Humain) afin de pouvoir effectuer les études spatiales [6].
Puis, concernant la table csv "tracks", on dénombre 3 jointures : une première avec la table csv "raw_tracks" pour joindre l'attribut "track_explicit"; une deuxième avec la table csv "echonest" pour joindre les attributs correspondant aux caractéristiques musicales (déjà présents dans la table "universal_top_spotify_song") et une dernière jointure qui est spatiale afin de faire correspondre la position géographique de l'artiste (attributs "latitude_artist" et "longitude_artist") avec un pays, une région du monde et un continent [7].

1.4. Création de subsets à partir des données initiales

Lorsque l'on commence à visualiser les données, on se rend compte de la grande hétérogénéité de nos données et de la difficulté d'obtenir des résultats pertinents lors de l'ACP. Il est alors nécessaire de réduire le jeu de données d'étude pour homogénéiser davantage les données.
En premier lieu, en ce qui concerne les données provenant de Spotify, le jeu de données passe ainsi de plus de 40 000 à 1 000 chansons. Ensuite, en analysant les histogrammes des différents attributs du dataset, il faut alors supprimer les chansons présentant des caractéristiques musicales trop "extrêmes" (outliers) soit celles qui entraînent un "écrasement" de l'histogramme vers des valeurs faibles, comme illustré ci-dessous par l'attribut "instrumentalness" (voir Figure 1). Il devient donc essentiel d'analyser les autres attributs et de filtrer les chansons en établissant des seuils. Ainsi, après le filtrage impliquant l'application de seuils sur les 14 variables numériques, on passe de 1 000 chansons à un nouveau dataset constitué d'environ 800 chansons.

En ce qui concerne les données extraites de la table "tracks" de FMA, la même approche a été adoptée que précédemment. Initialement, le dataset a été réduit de 13 000 à 1000 chansons. Ensuite, en appliquant des seuils aux 11 variables numériques (à l'exclusion de "latitude_artist", "longitude_artist", et "track_id") lors du filtrage des attributs, le jeu de données final est d'environ 750 chansons.

Ces sous-ensembles ont été générés à plusieurs reprises en utilisant des quantités variables de chansons (1000, 5000, 7000...) et les résultats sont plutôt équivalents.

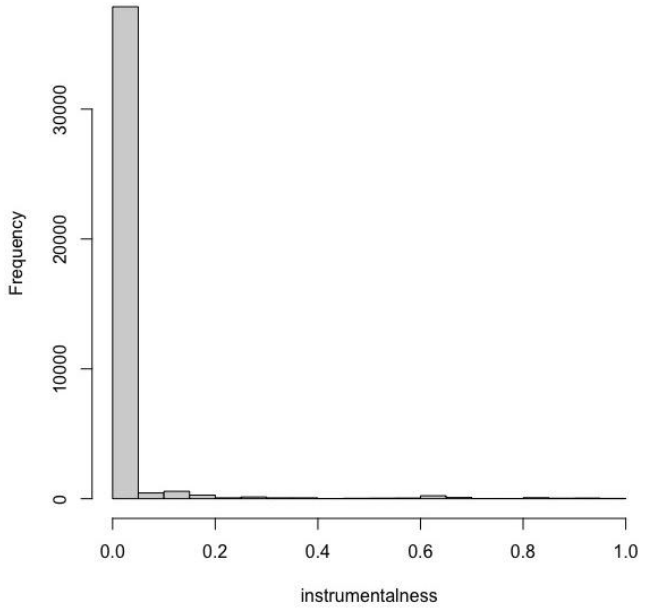


Figure 1 : Histogramme de l'attribut "instumentalness" (données Spotify)

II/ Analyses statistiques

Après avoir trouvé et mis en forme les données, il est nécessaire de les explorer en les décrivant et en les analysant, c'est l'objet de cette partie.
Tout d'abord, une analyse globale des attributs et de leurs relations est proposée. Puis, l'ACP est réalisée sur ces mêmes données afin de confirmer ou non les résultats précédents voire d'apporter un nouvel angle d'analyse. Enfin, cette partie se termine par une analyse tenant compte de la position géographique de la chanson.
Cependant, avant d'aller plus loin, il est nécessaire de définir l'ACP. En effet, dans les deux jeux de données, il y a un grand nombre de variables, malgré la sélection en amont, qu'il est difficile d'analyser de manière univariée ou bivariée. Alors, il est nécessaire de les approcher avec une méthode multivariée : "L'Analyse en Composantes Principales" ou ACP.
Cette méthode a pour objectif de réduire la colinéarité et le nombre de dimensions qui décrivent les chansons afin de résumer l'information tout en la restituant le plus fidèlement possible. Pour cela, il faut déterminer les axes composantes capturant le plus d'inertie possible des deux jeux de données des chansons, soit la majorité des relations existantes entre les attributs sélectionnées.

2.1. Données Spotify ("universal_top_spotify_song")

2.1.1. Analyse globale des données Spotify

D'abord, une étude générale des attributs du jeu de données de Spotify est essentielle.

En tenant compte de l'hétérogénéité des données, cette étude se réfère systématiquement à la médiane plutôt qu'à la moyenne car elle est moins sensible aux valeurs extrêmes et est plus adaptée aux données qui ont une distribution asymétrique. Pour comparer plusieurs variables quantitatives, un "graphique en radar" (ou "radarchart" en anglais) est pertinent (voir Figure 2). En effet, cela se présente comme sous la forme d'une toile d'araignée où "chaque variable est dotée d'un axe qui part du centre. Tous les axes sont disposés radialement, avec des distances égales entre eux, tout en maintenant la même échelle entre tous les axes. Les lignes de la grille qui relient les axes entre eux sont souvent utilisées comme guides. Chaque valeur de variable est tracée le long d'un axe individuel et toutes les variables d'un ensemble de données sont reliées pour former un polygone" [8]. L'axe du radarchart va de 0 à 1 et son pas est de 0,25. Plus la valeur médiane de la variable s'approche de 1, plus les chansons ont la tendance représentée par la variable.

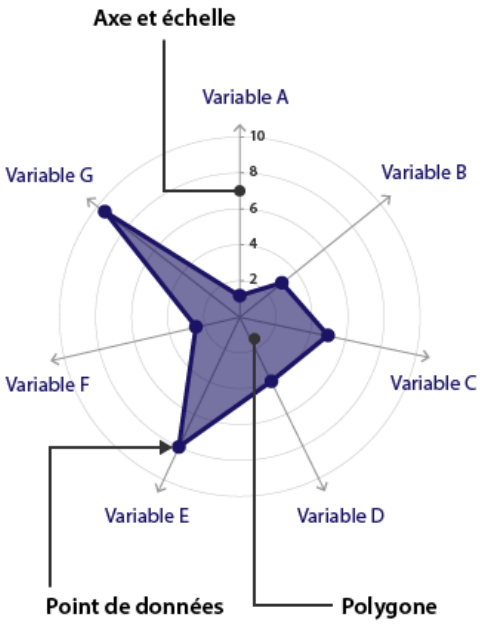


Figure 2 : Schéma d'un "graphique en radar"

Un premier "graphique en radar" a été réalisé montrant le profil médian d'une chanson populaire de Spotify. Il a été choisi d'exclure les attributs dont leur valeur médiane se situaient en dehors de l'intervalle 0 et 1. En effet, la normalisation de l'ensemble des données a été envisagée mais de nombreux attributs avaient déjà des valeurs comprises entre 0 et 1 donc cela pourrait fausser l'interprétation des résultats.
Alors, il est possible de constater que les chansons populaires de Spotify ont tendance à être entraînantes, dynamiques et énergiques comme en témoignent les valeurs des attributs "danceability"[1] et "energy" (voir Figure 3) qui sont environ égales à 0,75. Ces caractéristiques sont souvent associées aux genres musicaux tels que la pop, le rock et la musique électronique. Par ailleurs, il semble que les préférences des auditeurs de Spotify se dirigent

[1] Consulter l'annexe 3 du rapport pour avoir accès aux définitions précises des attributs utilisées ici

vers des chansons équilibrée en termes de volume (valeur médiane de “loudness" = -6.210[2]) (voir Annexe 4) avec une composante acoustique moins prononcée (valeur médiane de "acousticness" = 0.20) et une probabilité très faible d'être purement instrumentale (valeur médiane de “instrumentalness” $=1,7 \times 10^{-6}$). Autrement, l’attribut "liveness" est relativement basse (valeur médiane inférieure à 0,25), indiquant que les chansons populaires ont tendance à être enregistrées en studio plutôt que jouées en direct. De plus, les valeurs des attributs “valence”, “key” et “tempo” indiquent respectivement une variété d'émotions, une distribution relativement uniforme des tonalités et une diversité de rythme dans les chansons populaires.

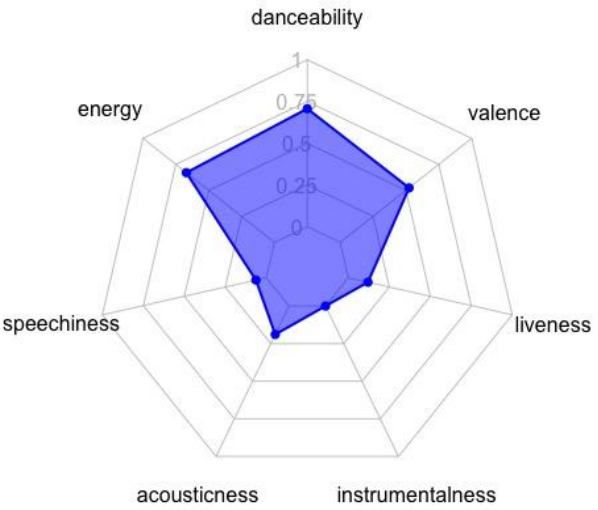


Figure 3 : Radarchart du profil médian des chansons provenant de Spotify

Après avoir analysé globalement chaque attribut, il est intéressant d’étudier les corrélations entre les différents attributs en mettant particulièrement l'accent sur leur relation avec l'attribut "popularity". Ce dernier représente la popularité, attribut étudié dans la problématique de l’étude, d'un morceau musical dans le dataset de Spotify. La valeur de "popularity" est déterminée par différents facteurs (nombre d’écoute et de sauvegarde, activité récente, etc.) qui reflètent la manière dont les utilisateurs interagissent avec la musique. On constate qu’il est difficile de trouver de bonnes corrélations entre l’attribut “popularity” et les autres attributs, la plus haute corrélation étant avec l’attribut “loudness” (0.14) ce qui présente quand même une faible corrélation. (voir Annexe 5)
Par rapport à la majorité des valeurs de corrélations de la table, on peut voir de “bonnes” corrélations entre l'attribut "danceability" et les attributs "valence" (0.36) et "energy" (0.23). Cette observation est cohérente car une musique perçue comme joyeuse et énergique peut naturellement inciter à la danse. Par ailleurs, une corrélation négative avec entre “danceability” et "acousticness" (-0.28) est relevée. Cette anti-corrélation suggère que les musiques populaires présentant une forte valeur de l’attribut “acousticness” ont tendance à être moins propices à la danse.

[2] Les valeurs de “loudness” peuvent varier de manière significative, mais en général, elles se situent souvent dans une fourchette de -60 dB à 0 dB. Une musique avec -6 dB en loudness est perçue comme forte.

Ces résultats soulignent alors la complexité des relations entre les attributs musicaux et la popularité. En effet, la musique est une forme d'art complexe qui peut être influencée par de nombreux facteurs tels que la mélodie, l'instrumentation ou bien la structure rythmique par exemple et les attributs musicaux peuvent ne pas avoir de relations linéaires simples entre eux. Pour appréhender au mieux cette grande inertie présente dans le jeu de donnée, il peut être utile d'utiliser des méthodes statistiques plus avancées telles que l'analyse en composantes principales (ACP).

2.1.2. ACP données Spotify

D'abord, selon le scree plot ci-dessous (voir Figure 4), les deux premières composantes (ou dimensions) principales sont conservées car le "point de coude" indique que l'ajout de composantes supplémentaires contribue moins significativement à l'explication de la variance totale du dataset. Cependant, les deux premières composantes principales sélectionnées capturent seulement 30% de l'inertie totale, donc il est important de faire attention à l'interprétation des résultats suivants.

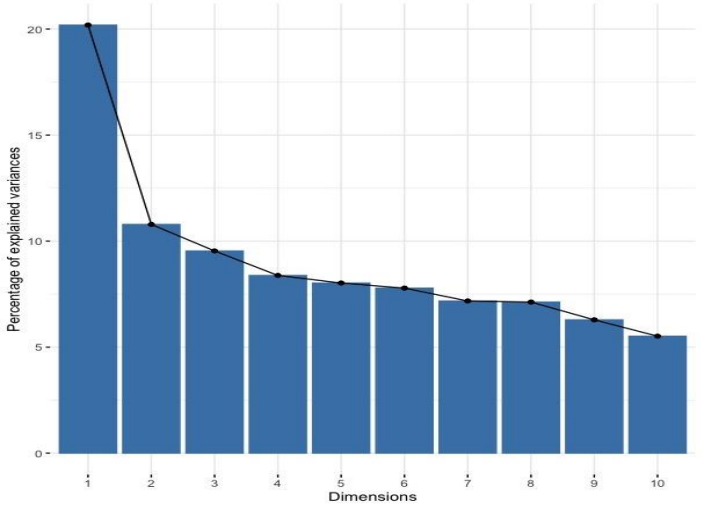


Figure 4 : Scree plot des données normalisées de Spotify

En ce qui concerne le cercle de corrélation (voir Figure 5), on remarque que la flèche de l'attribut "popularity" est assez loin du cercle donc il n'est pas bien représenté soit qu'il a une faible corrélation avec les composantes principales.
En analysant la relation entre "popularity" et les autres attributs, on remarque qu'il est bien corrélé avec les attributs "energy", "loudness", "tempo" et "key" car leurs flèches vont plus ou moins dans le même sens que celle de "popularity" sur le cercle. En effet, les musiques énergiques et dynamiques ont tendance à attirer davantage l'attention et peuvent être perçues comme plus populaires. Pour l'attribut "key", certaines tonalités sont associées à des émotions spécifiques, et les préférences des auditeurs pour certaines d'entre elles peuvent influencer la popularité de la chanson. Enfin, des tempos plus rapides peuvent rendre une chanson plus entraînante et adaptée à la danse et ainsi influencer sa popularité. Aussi, on peut

voir que les flèches des attributs “popularity” et “danceability” sont orthogonales ce qui indique que ces deux variables sont indépendantes entre elles ce qui est plutôt surprenant.

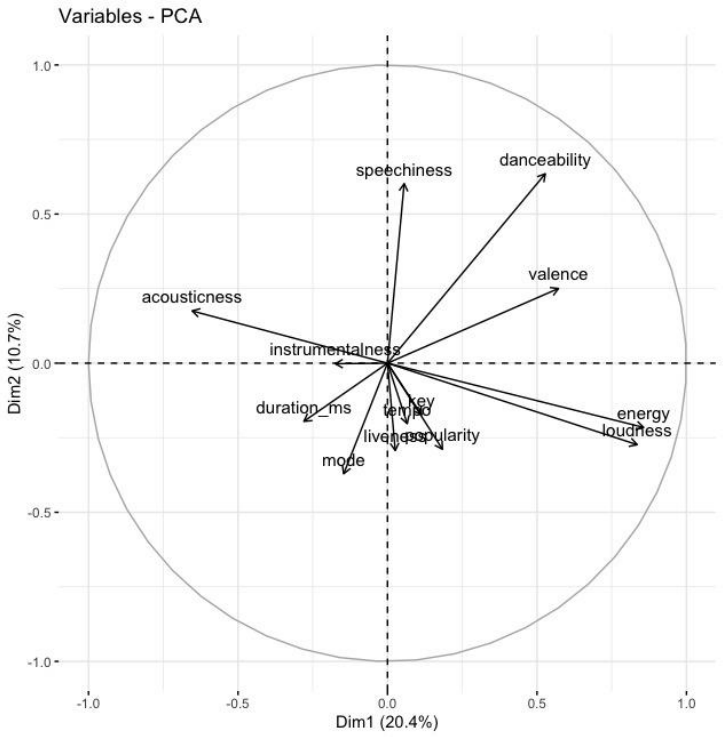


Figure 5 : Cercle de corrélation de l’ACP (données Spotify)

2.1.3. Analyse géographique données Spotify

Dans cette dernière section, l’analyse se porte sur les caractéristiques musicales médianes du Top 50 de tous les pays par continent grâce à la réalisation d’un graphique en radar une nouvelle fois (voir Figure 6).

Les observations révèlent que l'Amérique du Sud est le continent où la musique à caractère dansant (attribut “danceability”) est prédominante avec une médiane supérieure à 0,75 (suivi de peu par l'Amérique du Nord et par l’Afrique). Ainsi, le cliché porté sur l’Amérique de Sud qui écouterait le plus de musique à caractère dansante est justifié ici.

La perception de la positivité ou de la négativité d'une musique mesurée par l’attribut “valence” est également explorée. Ainsi, l'Amérique du Sud se distingue avec la médiane la plus élevée (0.586), suggérant une préférence pour des caractéristiques musicales perçues comme positives. A l’inverse, l'Asie affiche la médiane la plus basse (0.479), indiquant une tendance vers des caractéristiques perçues comme moins positives. Pour les attributs tels que "liveness", "instrumentalness" et "speechiness", les valeurs sont sensiblement similaires d'un continent à l'autre.

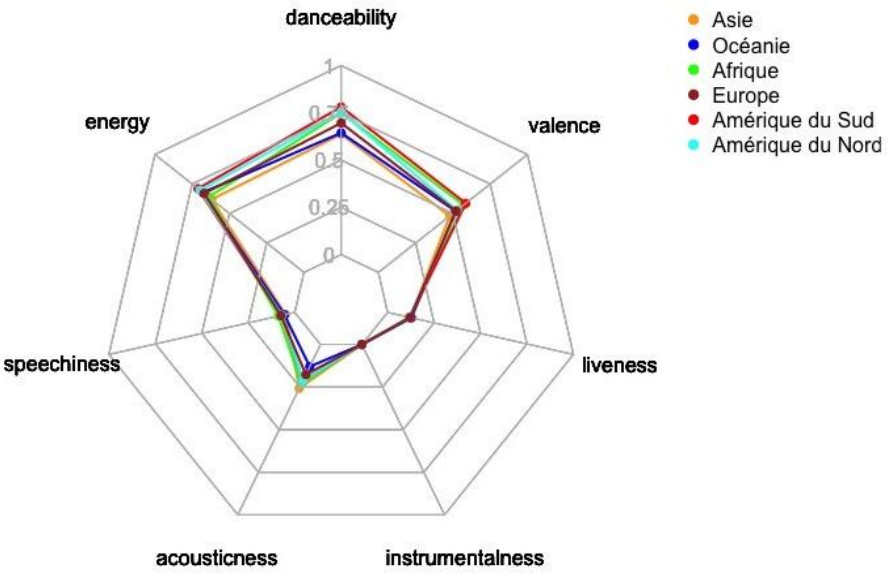


Figure 6 : Graphique en radar du profil médian des chansons du Top 50 de tous les pays selon le continent d'appartenance

En conclusion, il demeure complexe d'identifier un ou quelques attributs spécifiques au champ musical qui singularisent une chanson et la rendent populaire de manière universelle. L'étude des caractéristiques musicales, à travers diverses analyses telles qu'une analyse globale des données, l'ACP et les comparaisons à l'échelle continentale, suggère que la popularité musicale résulte d'une combinaison complexe de facteurs musicales tels que les attributs "loudness", "danceability", "energy", "valence" ainsi que "key" et "tempo. Cependant, d'autres facteurs peuvent jouer un rôle significatif notamment le facteur social qui sera étudié dans la suite de cette étude.
Néanmoins, grâce aux différents graphiques en radar réalisés, il est tout de même possible d'identifier les tendances musicales générales qui contribuent à la popularité d'une chanson à l'échelle nationale, continentale voire mondiale mais il est important de noter que ces caractéristiques peuvent considérablement varier en fonction du genre musical.

2.2. Données FMA ("tracks")

2.2.1. Analyse brute des données FMA

Maintenant, si on passe au jeu de données FMA, l'attribut correspondant le mieux à la popularité d'une chanson est le nombre d'écoutant de cette chanson soit "listens". Ainsi, les études statistiques faites plus loin sont relatives à cet attribut.
Tout d'abord, si on procède à une analyse globale des données provenant de FMA, on remarque plusieurs types de distributions des attributs (voir Annexe 6).
D'abord, les attributs tels que "favorites", "listens" et "duration" ont une distribution aplatie vers la droite soit vers de faibles valeurs car l'écart entre le maximum et la médiane de l'attribut est égal à plus de 300 fois (10 fois pour "duration") la valeur de cette dernière (également le cas avec la moyenne ou même leur troisième quartile) d'après l'Annexe 6. Cela signifie alors qu'il y a peu de chansons très écoutées ou mises en favori. On peut l'expliquer

par le caractère libre qu'est la plateforme FMA. En effet, les artistes amateurs peuvent publier librement et gratuitement leur morceau sur FMA contrairement sur Spotify.
Aussi, il y a le type de distribution précédent qui se retrouve dans des attributs comme "speechiness" (médiane = 0,04) et "liveness" (médiane = 0,12) qui correspondent respectivement à la proportion de parole dans le morceau et la probabilité que la chanson ait été enregistrée "en live". Il y a alors très peu de morceaux qui ont davantage de paroles que d'instrumentales et qui sont en live. Pour l'attribut "speechiness", on l'explique assez simplement car une musique est justement caractérisée par la présence de musique. Concernant l'attribut "liveness", la majorité des chansons sont produites en studio et peu en "live" alors.
Puis, il y a des distributions qui sont uniformément réparties comme l'attribut "energy" avec une médiane égale à 0,55 soit il y a autant de chansons qui sont moins énergiques que plus énergiques.
Enfin, concernant le tempo médian des chansons de FMA, il est de 120 BPM ce qui est un tempo plutôt moyen des chansons contemporaines.
Maintenant si on s'intéresse aux corrélations entre les attributs (voir Annexe 7), on observe que globalement les attributs ont peu de corrélations entre elles. La corrélation la plus forte dans le jeu de données est de 0,83 entre "listens" et "favorites". En effet, plus une chanson est écoutée plus elle a de chance d'être mise en favori et inversement. Sinon, aucun attribut ne porte une corrélation pertinente avec les attributs "listens" ou "favorites". Ainsi, le jeu de données est très hétérogène et globalement assez indépendant.
Cependant, il est possible de mettre en exergue d'autres liens d'attributs. Par exemple, la corrélation entre "danceabilility" et "valence" est à 0,43. Cela se justifie par le fait que plus il est possible de danser sur une chanson, plus cette dernière dégage de la positivité. Aussi, plus il est possible de danser sur une chanson, moins cette dernière a un temps long (-0,12 entre "danceability" et "duration") et plus elle a de paroles (0,17 entre "danceability" et "speechiness").
Ensuite, un graphique en radar du profil médian des chansons provenant du jeu de données FMA a été créé (voir Figure 7). On peut observer une prédominance de l'attribut "instumentalness" avec une médiane supérieure à 0,75, les chansons ont alors tendance à ne pas avoir beaucoup de paroles. En effet, les chansons amateurs sont en général dépourvues de paroles car ce sont souvent des débuts de créations.

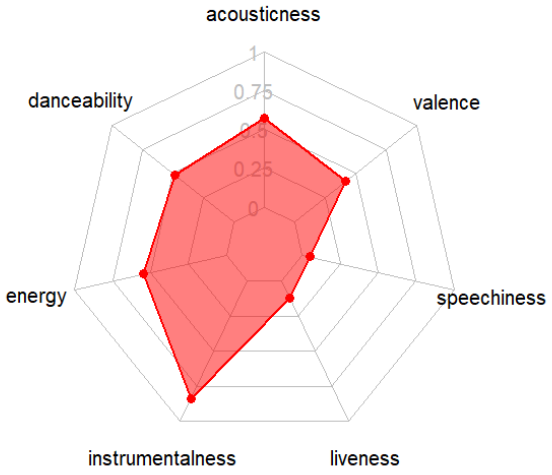


Figure 7 : Radarchart du profil médian des chansons provenant de FMA

Pour la suite, le genre musical de la chanson et son caractère explicite ou non vont être analysés. Pour cela, les deux attributs nécessaires sont “genres_top” et “tracks_explicit”. Pour ce dernier, on pose l’hypothèse forte que si la chanson ne contient pas de valeur alors la chanson est considérée comme non-explicite.

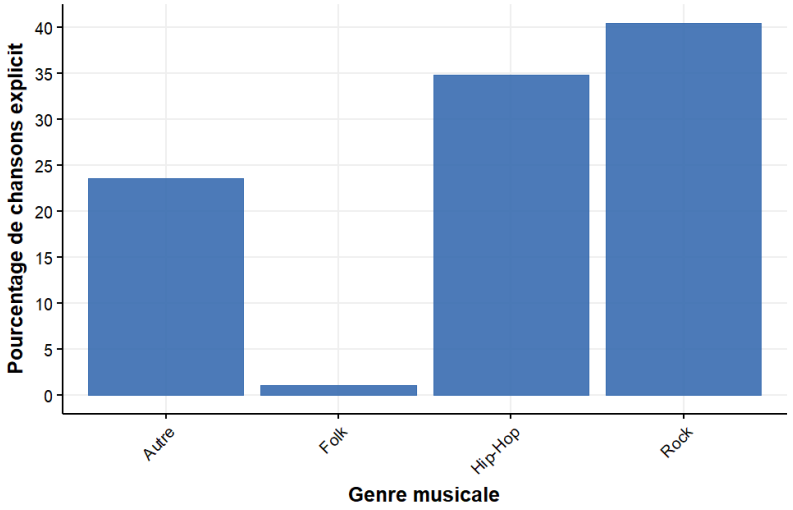


Figure 8 : Diagramme en barre du genre musicale en fonction du pourcentage de chansons considérées comme “explicite”

D’après la Figure 8, on remarque alors que ce sont les genres “Rock” et “Hip-Hop”, qui sont en tête, représentent respectivement 40% et 35% des chansons explicites du dataset. Ces résultats ne sont pas surprenants car historiquement, ces deux genres ont pour caractéristique principale de dénoncer et critiquer la société en général. Concernant le hip-hop, il ne faut pas négliger la violence et notamment sexuelle dans les paroles de ce genre musical. Pour aller plus loin et se rapprocher d’autant plus de la problématique, il serait intéressant d’étudier le nombre d’écoutes en fonction du genre musical.

2.2.2. ACP données FMA

Tout d'abord, avec le Scree Plot (voir Figure 9), on peut remarquer que les 10 dimensions capturent une quantité d'inertie de nos données équivalente entre elles. En effet, elles captent de 13% pour la dimension 1 à 5 % environ pour la dimension 10. Cela signifie que les composantes ne résument pas bien l'information du dataset en peu de dimensions, soit les variables ont des tendances très hétérogènes entre elles.
Pour la suite, le graphe des individus et le cercle de corrélation ne prendront que les deux premières dimensions qui capturent seulement 25,5 % de l'inertie totale. Les résultats suivants sont donc à prendre avec précaution mais peuvent tout de même apporter un éclairage supplémentaire aux relations entre les attributs des données de FMA.

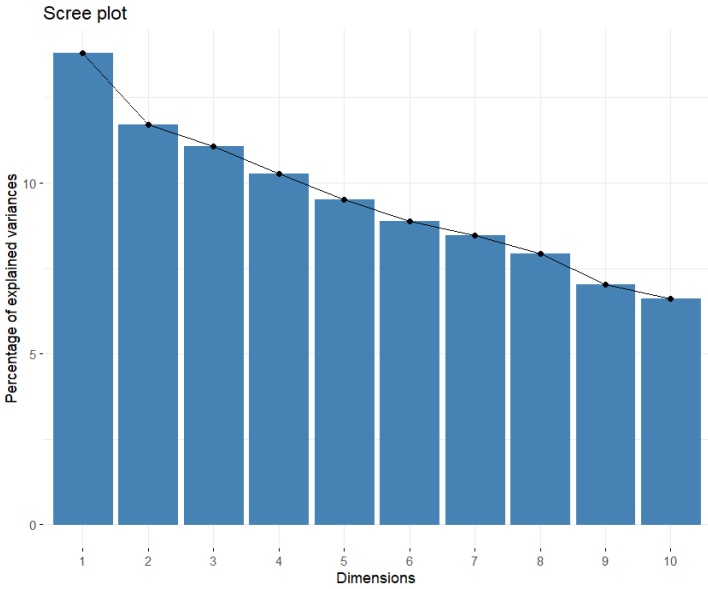


Figure 9 : Scree plot des données normalisées de FMA

Ensuite, au niveau de la répartition des chansons dans le nouvel espace d'arrivée (voir Figure 10), on observe une agglomération de chansons vers les origines des axes ce qui signifie que ce nouvel espace n'arrive pas à résumer les relations entre les individus comme déjà dit précédemment.

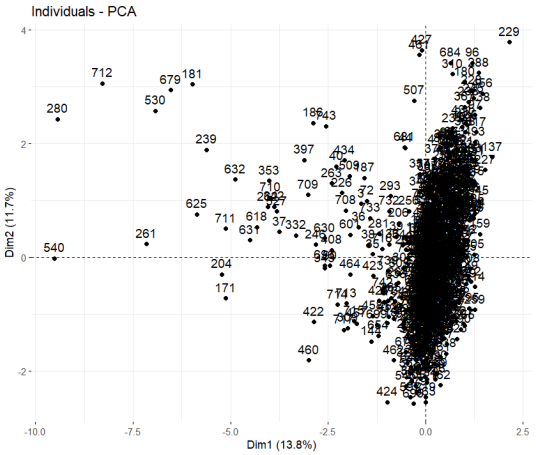


Figure 10 : Graphe des individus de l'ACP (données FMA)

Enfin, le cercle de corrélation donne les liens entre les attributs du jeu de données FMA (voir Figure 11). Avant de passer à l'analyse des liens entre les attributs, on remarque que les flèches de certains attributs sont petites. Cela signifie que ces derniers ne sont pas idéalement représentés sur l'espace d'arrivée.
Premièrement, on peut voir que les attributs "listens" et "favorites" sont corrélés car leurs flèches vont dans le même sens, ce qui a déjà été observé auparavant. Ensuite, deux groupes peuvent être distingués : il y a d'un côté "listen" et "favorites" et de l'autre toutes les autres caractéristiques musicales. C'est comme si toutes ces dernières étaient corrélées mais indépendamment du nombre d'écoutes ou de mises en favori de la chanson. On peut alors émettre l'hypothèse que les tendances d'écoute des auditeurs de la plateforme FMA n'ont pas de liens avec les caractéristiques musicales des chansons, soit qu'il y a comme une écoute aléatoire des auditeurs de FMA ou que la plateforme regorge d'énormes disparités de chansons du point de vue de leurs caractéristiques musicales qu'il est difficile d'établir une tendance générale des auditeurs de l'hébergeur FMA.

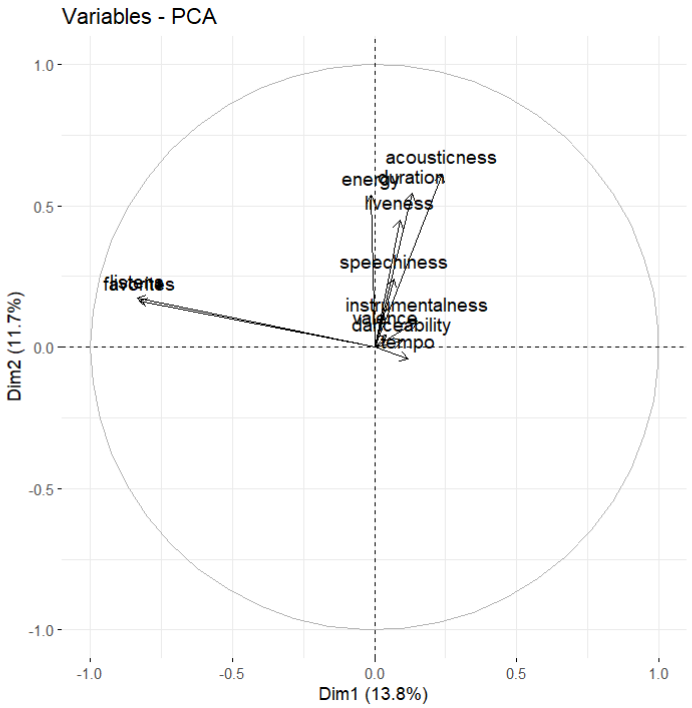


Figure 11 : Cercle de corrélation de l'ACP (données FMA)

2.2.3. Analyse géographique données FMA

La dernière partie de l'analyse des données FMA consiste à regarder si le profil médian des chansons FMA change en fonction du continent (voir Figure 12).
Contrairement aux données Spotify, le continent Américain ne sera pas divisé en Amérique du Sud et du Nord par facilité d'étude. Il serait pertinent de refaire la même analyse mais avec ces deux parties du monde séparées.
Aussi, il y a la catégorie "autre" qui représente les chansons qui n'avaient pas de pays

associés. Cette catégorie sera considérée dans le cadre de cette analyse comme un continent, il y aura alors au total 6 continents.
Tout d'abord, on peut remarquer qu'il n'y a pas beaucoup de différences de profil entre les continents. Par exemple, l'attribut "danceability" est égal à 0,5 pour les 6 continents; l'attribut "valence" est égal à 0,47 environ et l'attribut "liveness" est égal à 0,20.
Cependant, il est possible de noter des différences pour l'attribut "acousticness". En effet, le continent africain a un "acousticness" égal à 0,12 environ alors que l'Océanie a un "acousticness" égal à 0,75. On peut l'expliquer par des facteurs culturels ou plus probablement par le manque de représentativité des données. En effet, en Afrique il y a seulement trois chansons du jeu de données qui proviennent de ce continent là et cinquante pour l'Océanie. Cela peut s'expliquer par l'origine américaine de la plateforme FMA. Les continents les plus représentatifs et dont les résultats sont les plus fiables sont ceux de l'Amérique avec 4650 chansons, de l'Europe avec 2415 chansons. Quant à la catégorie "Autre", elle comporte 5774 chansons ce qui montre que le jeu de donnée à un niveau d'incomplétude élevé.

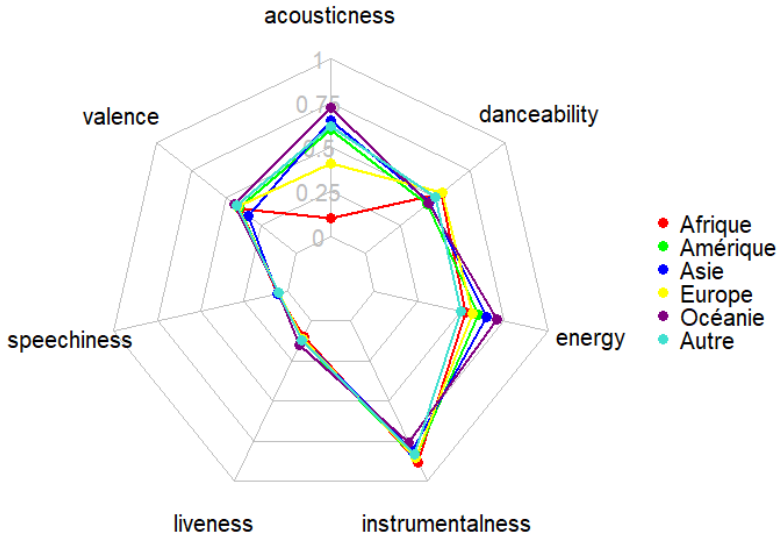


Figure 12 : Graphique en radar du profil médian des chansons selon le continent d'origine de l'artiste

2.3. Quelles différences entre musique commerciale et amateur ?

Il est à présent intéressant de comparer les caractéristiques musicales de Spotify et de FMA. Pour cela, deux graphiques en radar ont été réalisés (voir Figure 13) où les valeurs des attributs correspondent à la médiane de ces derniers.

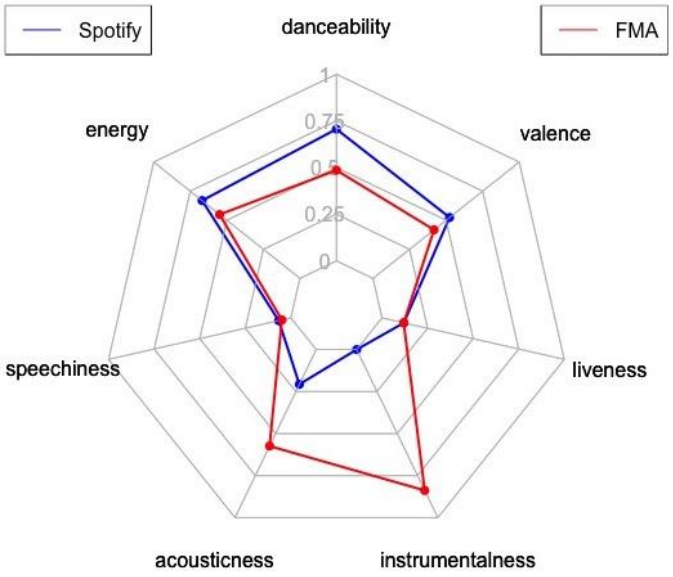


Figure 13 : Graphiques en radar comparant le profil médian des données Spotify et FMA

Les caractéristiques partagées entre les musiques de FMA et Spotify, telles que "speechiness" et "liveness", peuvent s'expliquer respectivement par la préférence générale des auditeurs pour des enregistrements en studio en raison de la qualité sonore qu'ils offrent et de la nécessité d'une composante instrumentale dans une chanson.
On observe que les chansons populaires de Spotify se distinguent par leur énergie (“energy”), leur positivité (“valence”) et leur capacité à inciter à la danse (“danceability”). Ces différences peuvent être dues à l'utilisation plus fréquente d'instruments électroniques et à une production numérique plus prononcée dans les musiques de Spotify (“acousticness”).
Et enfin, les chansons sur Spotify se caractérisent par une présence significative de paroles, par opposition aux compositions de FMA qui tendent davantage vers des musiques purement instrumentales (“instrumentalness”).
Ces conclusions sont cependant à prendre avec précaution car les données Spotify sont biaisées par le choix des chansons mise en avant sur la plateforme.

III. L'influence de la position géographique sur la popularité des chansons

3.1. Analyse de l’impact de la localisation sur la popularité des chansons

Dans cette partie, l'exploration de l'aspect spatial des données musicales de Spotify vise à évaluer l'influence potentielle du facteur géographique sur la popularité des chansons. L'analyse se concentre sur la chanson top 1 mondial de la période du jeu de données Spotify qui est “Monaco” de l’artiste puertoricain Bad Bunny en examinant ses variations de popularité d'un pays à l'autre.
Pour comprendre ces différences, une catégorisation des pays en deux groupes a été envisagée : les pays du Nord (pays développés) et les pays du Sud (pays en voie de développement), en utilisant l'Indice de Développement Humain (IDH) comme indicateur

pour classer les pays du jeu de données. Cet indicateur va de 0 à 1. Si le pays a un IDH proche de 1, cela signifie que ce pays est développé. Sinon, le pays est considéré comme en voie de développement.
Cependant, il faut savoir que les 70 pays classés dans le jeu de données de Spotify ont un IDH minimum égal à 0,65 donc il a été décidé que les pays dits “du Sud” seraient des pays “en voie de développement”. Ainsi, l’étude est biaisée par le choix des pays et notamment par le fait qu’il n’y a pas de pays considérés comme pauvres.
Une régression linéaire a été réalisée pour étudier la relation entre la popularité quotidienne d'une chanson (représentée par "daily_rank") et l'IDH (représenté par "HDI_value") du pays où cette chanson est la plus populaire.

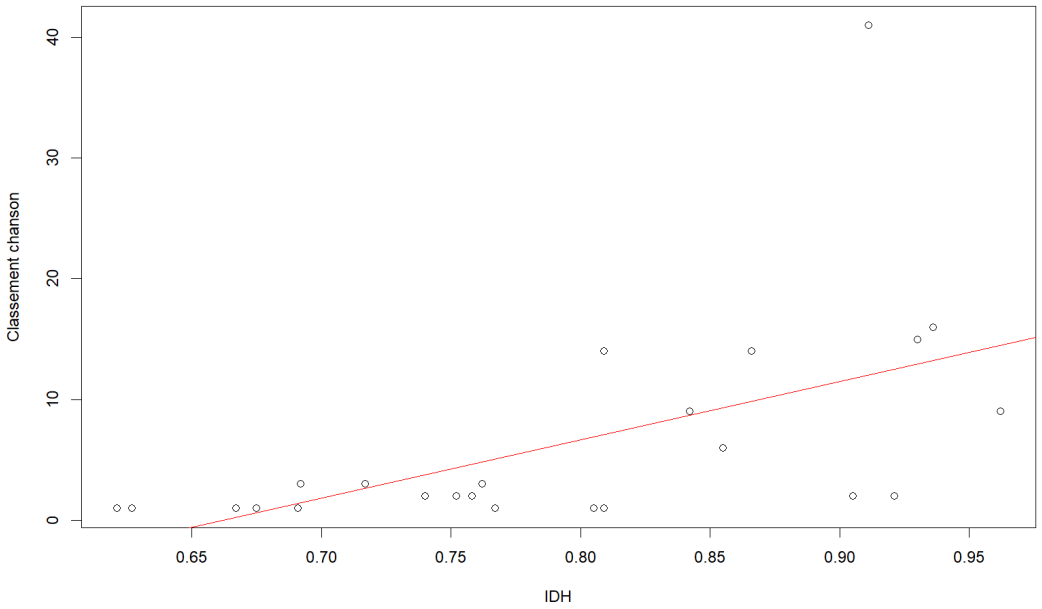


Figure 14 : Relation entre le développement d’un pays et la popularité des chansons

On remarque d’abord que la p-value associée à l'IDH est de $5,1 \times 10^{-3}$, cela montre que les résultats obtenus pour la régression linéaire sont fiables. Puis, le coefficient de détermination (R^2) est environ égal à 0.31, ce qui signifie qu’environ 31% de la variance de la popularité de la chanson peut être expliquée linéairement par l'IDH.
Selon cette analyse, il semble y avoir une relation positive entre l'IDH et la popularité de la chanson étudiée. En effet, une augmentation de l'IDH est associée à une augmentation du classement (soit moins de popularité) de la chanson, selon le modèle de régression linéaire. Cependant, il faut être vigilant vis-à-vis des résultats car le coefficient de détermination linéaire n’est pas très élevé, il faut donc rester critique des résultats données. Pour ce faire, une carte des résidus par pays de l’attribut “popularity” a été réalisée (voir Annexe 8). Cette carte montre que les résidus sont plutôt élevés et que la relation entre les 2 variables n’est pas nécessairement linéaire ici.

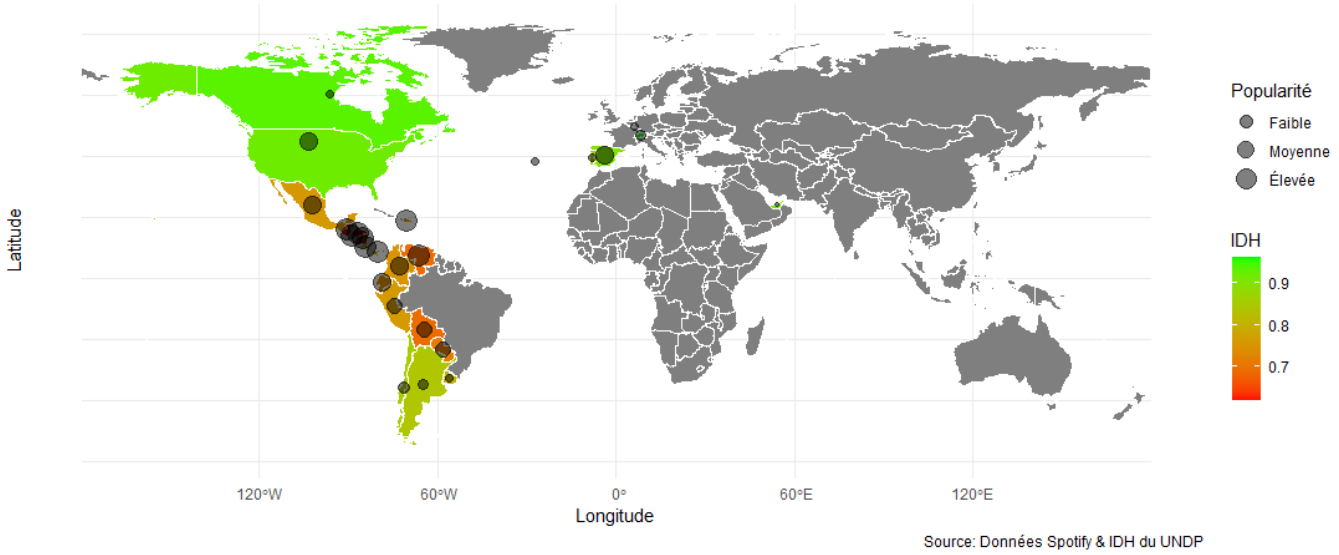


Figure 15 : Carte du classement des pays en fonction de leur IDH et de la popularité de la chanson Top 1 dans ces pays

Dans la figure ci-dessus (voir Figure 15), la distribution géographique des pays développés, distingués par leur IDH élevé et représentés en vert, contraste avec celle des pays en voie de développement caractérisés par un IDH plus faible et représentés par des couleurs tendant vers le rouge. Les cercles présents sur la carte indiquent l'indice de popularité des chansons, leur taille étant proportionnelle à leur niveau de popularité.
L'observation importante de cette carte est la variation de la popularité de la chanson "Monaco" de Bad Bunny d'un pays à un autre, alors même qu'il s'agit d'un succès mondial (Top 1). Cette popularité tend à être généralement plus élevée dans les pays du Sud que dans ceux du Nord. On peut l'expliquer par le genre musical et la langue parlée par l'artiste qui sont respectivement le reggaeton et l'espagnol ce qui explique leur popularité en Amérique du Sud globalement et en Espagne par exemple. Cela serait alors intéressant à étudier par la suite.
Cependant, il est crucial de souligner que la corrélation observée ne constitue pas nécessairement une relation de cause à effet et linéaire entre les deux variables. D'autres facteurs, autres que la localisation spatiale et non pris en compte dans notre modèle, pourraient également influencer la popularité des chansons tel que le choix de la plateforme.

3.2. Étude spatiale des clichés et des croyances populaires

Les stéréotypes à l'échelle mondiale suggèrent par exemple que les habitants des pays du Sud ont tendance à apprécier davantage les chansons à la fois entraînantes et dansantes que ceux des pays du Nord. La musique dynamique, souvent associée à des rythmes tropicaux, africains ou latino, est souvent perçue comme étant plus populaire.
Au contraire, la croyance populaire est de penser que les habitants des pays du Nord ont un plus grand intérêt envers une musique moins axée sur la danse et davantage vers des chansons qui offrent une profondeur musicale, une exploration artistique ou bien même une

réflexion intellectuelle. Ainsi, les genres musicaux expérimentaux et/ou complexes peuvent être plus valorisés.
Dans cette section, l'analyse se concentre sur le lien entre l'attribut "danceability" d'une chanson soit sa capacité à faire danser et l'appartenance d'un pays à la catégorie pays du Nord ou pays du Sud.
Pour ce faire, la moyenne de l'attribut "danceability" du Top 50 par pays a été calculée. Ensuite, une analyse a été menée pour établir la relation entre l'Indice de Développement Humain (IDH) d'un pays et la "danceability" moyenne de ses chansons populaires.

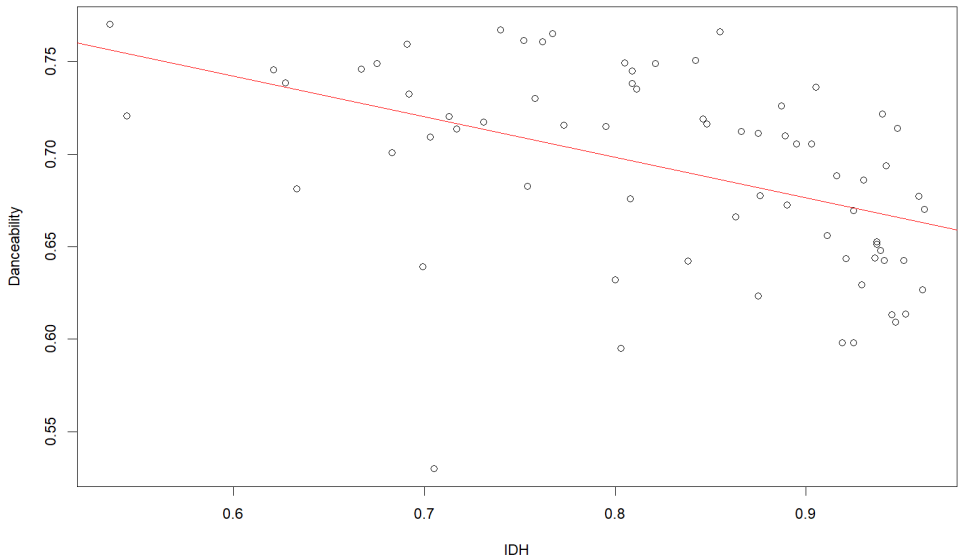


Figure 16 : Relation entre le développement d'un pays et l'attribut "danceability" des chansons

La p-value (5.24e-05) associée à la régression linéaire est très faible, indiquant une forte fiabilité statistique. Le R^2, quant à lui, est égal à environ 0.21. Cela signifie qu'environ 21% de la variance de l'attribut "danceability" moyenne des chansons peut être expliquée par l'IDH dans notre modèle. Ce coefficient est peu élevé, il faut donc rester critique une nouvelle fois sur les résultats donnés par la suite.
Dans ce modèle de régression linéaire, on contate qu'il y a une relation négative entre l'indice de développement humain et la moyenne de la "danceability" des chansons soit à mesure que l'IDH augmente, la "danceability" moyenne, elle, diminue.
Une représentation cartographique de cette relation est présentée dans la figure ci-dessous (voir Figure 17) où les pays développés présents dans le jeu de données Spotify se distinguent par leur IDH élevé et sont représentés en vert. A l'inverse, les pays en voie de développement sont caractérisés par un IDH plus faible et représentés par des couleurs tendant vers le rouge. Les cercles présents sur la carte indiquent, cette fois-ci, l'indice de "danceability" des chansons dont leur taille est proportionnelle à leur niveau de"danceability". On observe que les cercles de plus grande taille sont majoritairement situés dans les pays dits "du Sud" donc le cliché qui est d'associer les chansons les plus dansantes aux pays du Sud est considéré comme vérifié d'après la carte.

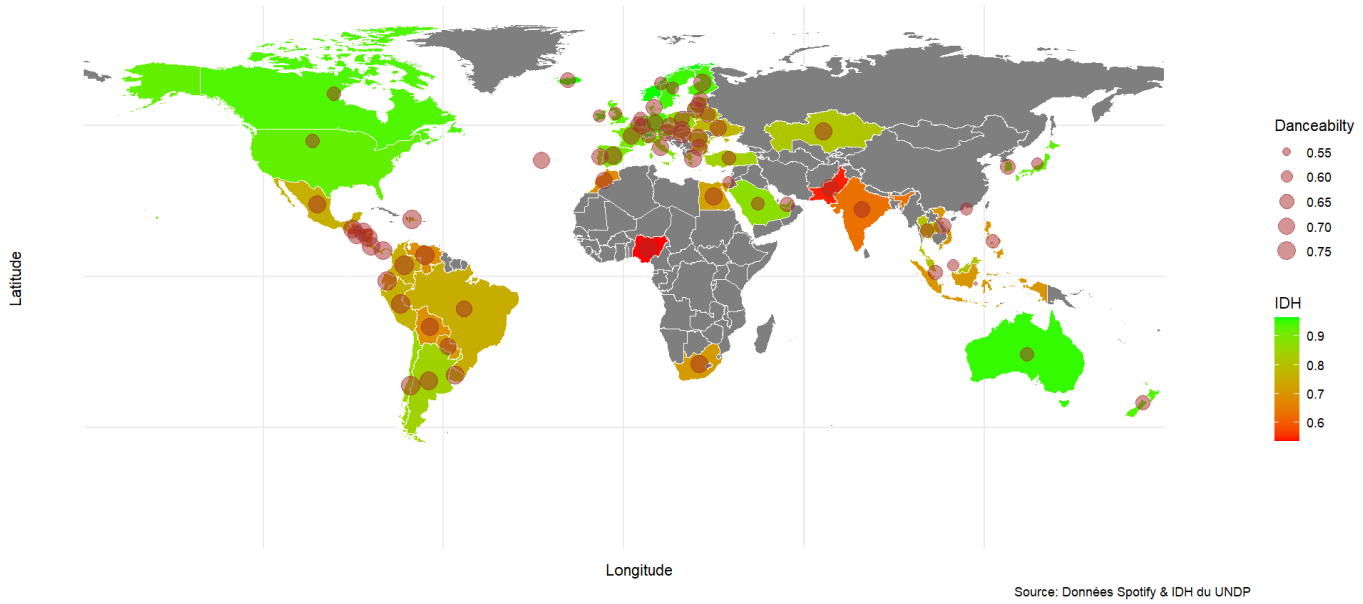


Figure 17 : Cartes du classement des pays du dataset Spotify en fonction de l'attribut "danceability" et de l'IDH

Enfin, la corrélation entre l'IDH et la "danceability" sur cette carte (voir Figure 17) est de -0.52. Cette dernière suggère donc une relation modérée et négative entre ces deux variables. Cela indique qu'à mesure que l'IDH d'un pays augmente, le caractère dansant des chansons tend à diminuer. On peut alors confirmer une nouvelle fois le cliché ici.
Une remarque est importante à faire sur l'étendue de l'attribut "danceability" ici qui est de 0,20 donc l'écart n'est pas conséquent entre les différentes catégories de l'attribut. Il serait alors intéressant, comme à l'étude précédente, d'étudier la relation entre "danceability" et des pays avec un IDH bien plus faible pour vérifier si le cliché persiste ou non.

IV. Les limites et perspectives de l'étude

Avant de conclure, il est important d'exposer dans un premier temps, les limites de notre étude ainsi que dans un second temps, les futures perspectives si l'étude devait se poursuivre.
Premièrement, concernant les limites des données Spotify, il est nécessaire de rappeler le biais existant concernant la sélection et la promotion des chansons sur la plateforme. En effet, Spotify est doté d'un algorithme qui peut mettre en avant le titre souhaité. De plus, les chansons importées sur la plateforme le sont majoritairement faites par les labels musicales qui choisissent eux-mêmes quel morceau mettre en avant. Tout cela est plus propice à ce que la chanson soit populaire contrairement aux chansons de FMA qui ne font pas ou peu de communications.
Puis, concernant les limites des données FMA et plus précisément la source des données, il n'est pas décrit, que ce soit dans le répertoire github ou dans l'article scientifique de ce jeu de données, comment ces données ont été collectées. En effet, en consultant le site internet de l'hébergement de musique [3], aucune page web du site ne se réfère à la possibilité d'exporter

des jeux de données. Alors, cela contribue à prendre du recul et à avoir un aspect critique sur les résultats de nos données.
Aussi, lors de l'analyse du lien entre genre musical "genres" et l'attribut "track_explicit", ce dernier contient des cellules sans valeurs (différent de NA). Alors, cela pose un doute sur la raison de cette incomplétude. Est-ce parce que la chanson est considérée comme "non explicite" ou est-ce un oubli d'informations ?

Deuxièmement, si l'étude devait se poursuivre, il serait d'abord intéressant d'analyser d'autres attributs non-traités comme la durée des chansons ou bien d'essayer d'améliorer les résultats des ACP et notamment celui de FMA en essayant de savoir si la distribution de nos chansons dans l'espace d'arrivée relève d'une réelle distribution aléatoire ou non. Pour cela, on pourrait créer 100 chansons avec des caractéristiques musicales aléatoires et les représenter dans l'espace d'arrivée. Si la dispersion des points correspond à celle de notre distribution de base alors on pourrait conclure à une distribution plutôt aléatoire de nos données.
Ensuite, une étude comparative des caractéristiques musicales des pays en temps de guerre et en temps de paix pourrait offrir une perspective nouvelle sur l'évolution de la musique dans un contexte géopolitique. Cette analyse permettrait de saisir comment les dynamiques de conflit influencent potentiellement la création artistique et les tendances musicales. Cette exploration enrichirait la compréhension des interactions entre la musique, la géopolitique et les sociétés, offrant ainsi une perspective novatrice sur l'influence des contextes historiques sur les expressions culturelles.
Enfin, il serait judicieux d'explorer l'évolution temporelle des caractéristiques musicales. En effet, analyser comment les paramètres musicaux fluctuent au fil du temps permettrait de saisir les tendances émergentes, les changements culturels et les influences externes sur la création musicale. Cette approche chronologique fournirait des "insights" précieux sur l'évolution de la musique au sein des sociétés, enrichissant ainsi notre compréhension des dynamiques culturelles à travers les époques.

Conclusion

Pour conclure, une chanson peut devenir populaire selon différents facteurs. D'abord, il y a des facteurs musicaux comme la présence de parole ("instrumentalness") ou si la chanson a été produite en studio ("liveness"). On peut également citer l'énergie de la chanson ("energy"), sa capacité à faire danser ("danceability"), son tempo ("tempo") ainsi que sa durée ("duration") qui sont des facteurs importants pour rendre une chanson populaire.
Puis, le facteur géographique influe sur la popularité d'une chanson avec les études effectuées à l'échelle nationale et continentale ainsi que le facteur social que l'on a montré avec les études spatiales de l'IDH des pays du dataset de Spotify.
Enfin, le choix de la plateforme a aussi toute son importance car selon celle-ci, la chanson pourrait être mise plus ou moins en avant et cela joue alors sur sa popularité. Ainsi, les chansons hébergées par Spotify ont bien plus de chances d'être populaires que celles de FMA.

Sitographie

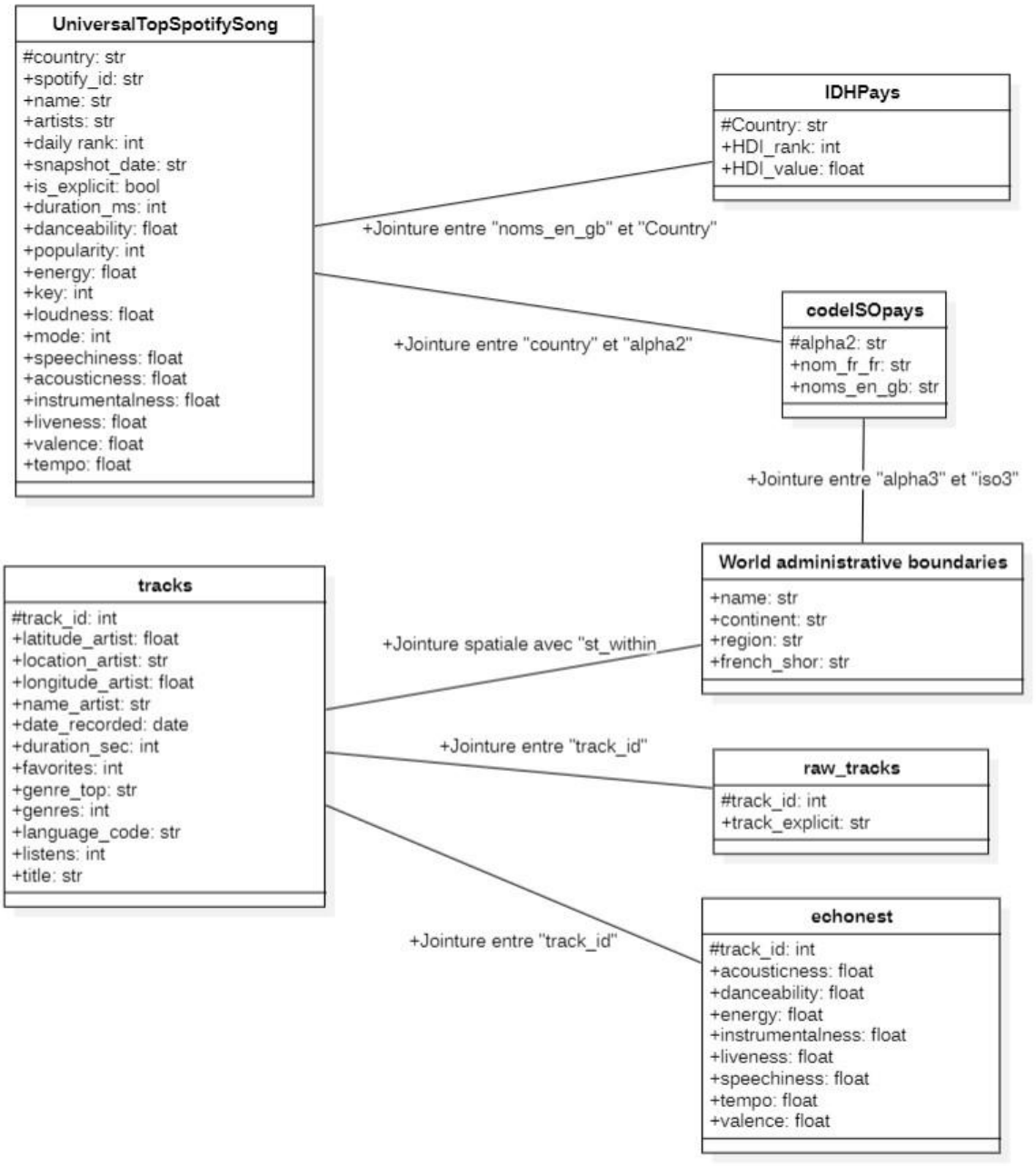
Numéro du lien dans le rapport	URL du site internet
[1]	https://www.kaggle.com/datasets/asaniczka/top-spotify-songs-in-73-countries-daily-updated
[2]	https://github.com/mdeff/fma/blob/master/LICENSE.txt
[3]	https://freemusicarchive.org/home
[4]	https://github.com/Hasdichaima/SpotifyTopSongsGeoStat
[5]	https://sql.sh/514-liste-pays-csv-xml#google_vignette
[6]	https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fhdr.undp.org%2Fsites%2Fdefault%2Ffiles%2F2021-22_HDR%2FHDR21-22_Statistical_Annex_HDI_Table.xlsx&wdOrigin=BROWSELINK
[7]	https://public.opendatasoft.com/explore/dataset/world-administrative-boundaries/table/
[8]	https://datavizcatalogue.com/FR/methodes/graphique_en_radar.html

Annexes

Annexe 1: Récapitulatif des différentes tables présentes dans le jeu de données FMA

Nom de la table	Description	Nombre d'attributs
Features	Sept statistiques ont été calculées sur plusieurs groupes de chanson regroupées : la moyenne, l'écart-type, l'asymétrie, l'aplatissement, la médiane, le minimum et le maximum. Ces 518 caractéristiques précalculées sont réparties dans features.csv pour toutes les pistes.	Énormément
Tracks	Caractéristiques complètes des chansons	53
Raw_tracks	Informations associées aux chansons	39
Raw_genres	Caractéristiques des genres musicaux	5
Genres		5
Raw_albums	Caractéristiques des albums liés aux chansons	19
Raw_artists	Caractéristiques liés à l'artiste de la chanson associée	25
Echonest	Caractéristiques des chansons fournis par l'API de Spotify (anciennement appelé "Echonest") pour un sous-ensemble d'environ 13 000 chansons	244
Raw_echonest	Contient les mêmes informations que la table "Echonest"	250

Annexe 2 : Schéma avec les différents attributs sélectionnés dans chaque table et leurs jointures associées (les attributs débutant par "#" sont les clés de jointures de leur table)



Annexe 3 : Tableau des définitions des attributs sélectionnées

Nom table	Nom attribut	Définition
UniversalTopSpotifySong	country (str)	Code ISO du pays associé à la playlist du top 50 des chansons. Quand la valeur est “null”, la playlist correspond au classement global soit à 70 pays. Clé de jointure avec “codeISOpays.csv”
	spotify_id (str)	Identifiant de la chanson selon l’API de Spotify. Clé primaire de la table
	name (str)	Titre de la chanson
	artists (str)	Nom de/des artiste(s) de la chanson
	daily_rank (int)	Rang journalier de la chanson dans la liste des 50 meilleures (1 à 50).
	snapshot_date (str)	Date à laquelle la chanson a été collectée depuis l’API Spotify
	is_explicit (bool)	Indique si la chanson contient des paroles non conseillées aux jeunes (pornographie, sexuel, violence)
	duration_ms: int	Durée de la chanson en millisecondes
	danceability (float)	Mesure de la “possibilité de pouvoir danser” sur la chanson basée sur divers éléments musicaux comme le tempo, la stabilité rythmique, la force des percussions (beat strength) et la régularité de la chanson. - 0,0 → “least danceable” - 1,0 → “most danceable”
	popularity (int)	Mesure de la popularité actuelle de la chanson sur Spotify (de 0 à 100). il se peut que la popularité soit égale à 0. Dans ce cas, cela peut être dû au fait que la chanson vient de sortir.
	energy (float)	Mesure de l’énergie de la chanson soit son niveau d’intensité et d’activité. Les chansons considérées comme énergiques sont fortes, bruyantes et rapides. Exemple : - Death metal → proche de 1,0 - Bach prelude → proche 0,0
	key (int)	Clé de la hauteur musicale majoritaire de la chanson. Les clés vont de 0 à 11. Exemple : Si la clé est égale à 11, cela signifie que la chanson comporte globalement pas mal d'octaves “Si” soit d'aigus.
	loudness (float)	Mesure de l’intensité sonore de la chanson en dB
	mode (int)	Le mode le plus couramment utilisé est le mode majeur (1), qui crée une atmosphère lumineuse et joyeuse. Le mode mineur (0) est également largement utilisé et donne une sensation plus sombre ou mélancolique à une chanson.
	speechiness (float)	Mesure de la présence de parole dans la chanson.

		> 0,66 : la chanson comporte majoritairement des paroles Entre 0,33 et 0,66 : la chanson contient à la fois des paroles et de la musique. Exemple : Rap < 0,33 : la chanson contient majoritairement de la musique
	acousticness (float)	Mesure de la qualité acoustique de la chanson. - 0,0 : musique non acoustique - 1,0 : musique acoustique
	instrumentalness (float)	Représente le nombre de voix dans la chanson. Plus proche on est de 1, le plus probable la chanson ne contient pas de voix.
	liveness (float)	Décrit la probabilité que la chanson ait été enregistrée “en live”. Si > 0,8 : Grande probabilité que la chanson soit enregistré “en live”
	valence (float)	Mesure de la positivité d’une chanson. - 0,0 : musique à tendance négative comme de la tristesse, de la colère - 1,0 : musique à tendance positive comme l’euphorie, la joie, la gratitude
	tempo (float)	tempo (float) : Nombre de beats/min soit la vitesse de la chanson
tracks	track_id (int)	Identifiant de la chanson (clé de jointure)
	latitude_artist (float)	latitude géographique de l’artiste
	longitude_artist (float)	longitude géographique de l’artiste
	location_artist (str)	Lieu géographique de l’artiste
	name_artist (str)	Nom de l’artiste
	date_recorded (date)	Date d’enregistrement de la chanson
	duration_sec (int)	Durée de la chanson en secondes
	favorites (int)	Nombre de “mise en favoris” de la chanson sur FMA
	genre_top (str)	Genre prédominant de la chanson
	genres (int)	Code du ou des genres de la chanson
	language_code (str)	Langue utilisée dans la chanson
	listens (int)	Nombre d’écoutes de la chanson sur FMA
	title (str)	Titre de la chanson
codeISOpays	alpha2 (str)	Code ISO à deux lettres. Clé de jointure.
	nom_fr_fr (str)	Nom du pays en français associé au code ISO

	noms_en_gb (str)	Nom du pays en anglais associé au code ISO
IDHPays	HDI_rank (int)	Rang de l'IDH du pays
	HDI_value	Valeur de l'IDH du pays
raw_tracks	track_explicit (str)	idem que "is_explicit"
World administrative boundaries	name (str)	Nom du pays en anglais
	continent (str)	Nom du continent auquel appartient le pays 'name"
	region (str)	Nom de la région à laquelle appartient le pays "name"
	french_shor (str)	Nom du pays en français

Annexe 4 : Synthèse statistique de différents attributs (données Spotify)

```
  daily_rank      popularity      duration_ms      danceability        energy            key
Min.   : 1.00   Min.   :  0.00   Min.   :     0   Min.   :0.2220   Min.   :0.0242   Min.   : 0.000
1st Qu.:13.00   1st Qu.: 67.00   1st Qu.:162767   1st Qu.:0.5980   1st Qu.:0.5460   1st Qu.: 2.000
Median :25.00   Median : 83.00   Median :188108   Median :0.7060   Median :0.6690   Median : 6.000
Mean   :25.51   Mean   : 78.57   Mean   :194698   Mean   :0.6905   Mean   :0.6456   Mean   : 5.536
3rd Qu.:38.00   3rd Qu.: 90.00   3rd Qu.:220653   3rd Qu.:0.8000   3rd Qu.:0.7530   3rd Qu.: 9.000
Max.   :50.00   Max.   :100.00   Max.   :641941   Max.   :0.9740   Max.   :0.9970   Max.   :11.000
   loudness            mode         speechiness      acousticness       instrumentalness
Min.   :-22.497   Min.   :0.0000   Min.   :0.0232   Min.   :0.0000075   Min.   :0.0000000
1st Qu.: -8.027   1st Qu.:0.0000   1st Qu.:0.0426   1st Qu.:0.0856000   1st Qu.:0.0000000
Median : -6.210   Median :0.0000   Median :0.0665   Median :0.2060000   Median :0.0000017
Mean   : -6.634   Mean   :0.4888   Mean   :0.1107   Mean   :0.2902858   Mean   :0.0186684
3rd Qu.: -4.912   3rd Qu.:1.0000   3rd Qu.:0.1450   3rd Qu.:0.4550000   3rd Qu.:0.0001000
Max.   :  1.155   Max.   :1.0000   Max.   :0.7840   Max.   :0.9840000   Max.   :0.9680000
   liveness         valence           tempo
Min.   :0.0154   Min.   :0.0373   Min.   : 47.91
1st Qu.:0.0986   1st Qu.:0.3630   1st Qu.: 99.97
Median :0.1200   Median :0.5240   Median :120.03
Mean   :0.1723   Mean   :0.5293   Mean   :122.12
3rd Qu.:0.2110   3rd Qu.:0.7100   3rd Qu.:140.06
Max.   :0.9680   Max.   :0.9780   Max.   :217.97
```

Annexe 5 : Matrice de corrélation (données Spotify)

	daily_rank	popularity	duration_ms	danceability	energy
daily_rank	1	-0.122364755485604	0.0467697255543284	-0.0982575347148539	0.028804927287867
popularity	-0.122364755485604	1	0.0441360480161956	-0.0262012724524519	0.00935977321683261
duration_ms	0.0467697255543284	0.0441360480161956	1	-0.209378009420512	-0.0776352830060422
danceability	-0.0982575347148539	-0.0262012724524519	-0.209378009420512	1	0.230946565068783
energy	0.028804927287867	0.00935977321683261	-0.0776352830060422	0.230946565068783	1
key	-0.00372299235480126	-0.019405141559262	-0.0645671930685617	-0.008191420658772	0.0907056982032941
loudness	0.0184375497453264	0.145726984650217	-0.0470636602887995	0.227955385959086	0.761144789016324
mode	-0.0172051389845912	0.0726615199561078	0.0749311242248014	-0.158244822526376	-0.0501421493429224
speechiness	-0.035844892398883	-0.0719942062136391	0.00330354875334027	0.226775757611976	0.000429985932904362
acousticness	-0.0424182087298372	0.015696654497511	0.0485246596115926	-0.28879375565508	-0.580752561366198
instrumentalness	0.0373114249244014	-0.0365981188445219	-0.00795212867557364	-0.0665536919158066	0.00070404756563774
liveness	0.0223308521423097	-0.0268276674738046	-0.0343125800714008	-0.108714951391535	0.095695907909521
valence	-0.0286632759578403	-0.0260839955056314	-0.174683001627917	0.360380374814147	0.350113157891089
tempo	-0.00266508276431524	0.0193378565000236	-0.0275715097899238	-0.151859036665201	0.104558721040448

	key	loudness	mode	speechiness	acousticness
daily_rank	-0.00372299235480126	0.0184375497453264	-0.0172051389845912	-0.035844892398883	-0.0424182087298372
popularity	-0.019405141559262	0.145726984650217	0.0726615199561078	-0.0719942062136391	0.015696654497511
duration_ms	-0.0645671930685617	-0.0470636602887995	0.0749311242248014	0.00330354875334027	0.0485246596115926
danceability	-0.008191420658772	0.227955385959086	-0.158244822526376	0.226775757611976	-0.28879375565508
energy	0.0907056982032941	0.761144789016324	-0.0501421493429224	0.000429985932904362	-0.580752561366198
key	1	0.0382585307166239	-0.0591470944403095	-0.0354591053475637	0.000220349230002224
loudness	0.0382585307166239	1	-0.0273291345748011	-0.0701595268196633	-0.459207579284939
mode	-0.0591470944403095	-0.0273291345748011	1	-0.0370194425849432	-0.00657676233741669
speechiness	-0.0354591053475637	-0.0701595268196633	-0.0370194425849432	1	-0.0415000596574795
acousticness	0.000220349230002224	-0.459207579284939	-0.00657676233741669	-0.0415000596574795	1
instrumentalness	0.019963839424936	-0.118836900185992	-0.0109251144317833	-0.0275281375274027	0.00632310533662449
liveness	0.00811976773490637	0.0735818353167319	-0.0289469285318999	-0.0120028330021815	-0.0612518118743665
valence	0.104481370391261	0.306618034090634	-0.06077681616507	0.0146695453434902	-0.182510824155291
tempo	0.124870051649138	0.0543762683832283	-0.0486688839860611	0.0906854567566362	-0.021536794452138

	instrumentalness	liveness	valence	tempo
daily_rank	0.0373114249244014	0.0223308521423097	-0.0286632759578403	-0.00266508276431524
popularity	-0.0365981188445219	-0.0268276674738046	-0.0260839955056314	0.0193378565000236
duration_ms	-0.00795212867557364	-0.0343125800714008	-0.174683001627917	-0.0275715097899238
danceability	-0.0665536919158066	-0.108714951391535	0.360380374814147	-0.151859036665201
energy	0.00070404756563774	0.095695907909521	0.350113157891089	0.104558721040448
key	0.019963839424936	0.00811976773490637	0.104481370391261	0.124870051649138
loudness	-0.118836900185992	0.0735818353167319	0.306618034090634	0.0543762683832283
mode	-0.0109251144317833	-0.0289469285318999	-0.06077681616507	-0.0486688839860611
speechiness	-0.0275281375274027	-0.0120028330021815	0.0146695453434902	0.0906854567566362
acousticness	0.00632310533662449	-0.0612518118743665	-0.182510824155291	-0.021536794452138
instrumentalness	1	-0.019869260921769	-0.126297942926325	0.0276991824577696
liveness	-0.019869260921769	1	-0.0125360996912765	0.0755689458441625
valence	-0.126297942926325	-0.0125360996912765	1	0.0275246571116778
tempo	0.0276991824577696	0.0755689458441625	0.0275246571116778	1

Annexe 6 : Synthèse statistique de différents attributs (données FMA)

```
    duration      favorites          listens        acousticness        danceability
 Min.   :  18   Min.   :   0.00   Min.   :    12   Min.   :0.0000009   Min.   :0.05131
 1st Qu.: 156   1st Qu.:   0.00   1st Qu.:   299   1st Qu.:0.1037726   1st Qu.:0.34476
 Median : 214   Median :   1.00   Median :   694   Median :0.5739848   Median :0.48563
 Mean   : 249   Mean   :   4.31   Mean   :  2147   Mean   :0.5246876   Mean   :0.48729
 3rd Qu.: 282   3rd Qu.:   4.00   3rd Qu.:  1714   3rd Qu.:0.9207270   3rd Qu.:0.62909
 Max.   :3033   Max.   :1482.00   Max.   :543252   Max.   :0.9957965   Max.   :0.96864
     energy          instrumentalness    liveness       speechiness          tempo
 Min.   :0.0000202   Min.   :0.0000   Min.   :0.0253   Min.   :0.02232   Min.   : 12.75
 1st Qu.:0.3213004   1st Qu.:0.3235   1st Qu.:0.1014   1st Qu.:0.03693   1st Qu.: 95.97
 Median :0.5491128   Median :0.8381   Median :0.1190   Median :0.04902   Median :120.06
 Mean   :0.5375155   Mean   :0.6405   Mean   :0.1878   Mean   :0.09917   Mean   :123.08
 3rd Qu.:0.7762542   3rd Qu.:0.9182   3rd Qu.:0.2110   3rd Qu.:0.08545   3rd Qu.:145.32
 Max.   :0.9999637   Max.   :0.9980   Max.   :0.9803   Max.   :0.96618   Max.   :251.07
    valence
 Min.   :0.00001
 1st Qu.:0.19732
 Median :0.41774
 Mean   :0.43976
 3rd Qu.:0.66558
 Max.   :0.99999
```

Annexe 7 : Matrice de corrélation (données FMA)

```
                    duration    favorites      listens acousticness danceability       energy
duration          1.0000000000  0.003050044 -0.004538799  0.041737106  -0.12132169 -0.114296373
favorites         0.0030500440  1.000000000  0.835643840  0.008357059   0.05769243 -0.022431315
listens          -0.0045387989  0.835643840  1.000000000  0.004849266   0.05105670 -0.040069471
acousticness      0.0417371061  0.008357059  0.004849266  1.000000000  -0.18959884 -0.477273313
danceability     -0.1213216922  0.057692434  0.051056695 -0.189598842   1.00000000  0.045344583
energy           -0.1142963725 -0.022431315 -0.040069471 -0.477273313   0.04534458  1.000000000
instrumentalness  0.0534207062  0.038726575  0.022137451  0.110033185  -0.11803321 -0.002411791
liveness          0.0854020010 -0.012648942 -0.009207159  0.041319264  -0.14333924  0.045752380
speechiness      -0.0007307816 -0.030687341 -0.020449731  0.038784518   0.17131107 -0.008644883
tempo            -0.0523933058  0.002032443 -0.005549406 -0.110701178  -0.09435193  0.227324349
valence          -0.2040997558  0.024509048  0.017307157 -0.085436222   0.42851487  0.219384081
                 instrumentalness     liveness   speechiness        tempo     valence
duration             0.053420706  0.085402001 -0.0007307816 -0.052393306 -0.20409976
favorites            0.038726575 -0.012648942 -0.0306873410  0.002032443  0.02450905
listens              0.022137451 -0.009207159 -0.0204497311 -0.005549406  0.01730716
acousticness         0.110033185  0.041319264  0.0387845183 -0.110701178 -0.08543622
danceability        -0.118033208 -0.143339245  0.1713110720 -0.094351931  0.42851487
energy              -0.002411791  0.045752380 -0.0086448835  0.227324349  0.21938408
instrumentalness     1.000000000 -0.058593173 -0.2166893872  0.023003172 -0.14519970
liveness            -0.058593173  1.000000000  0.0731040742 -0.007566410 -0.01788594
speechiness         -0.216689387  0.073104074  1.0000000000  0.032187644  0.09479414
tempo                0.023003172 -0.007566410  0.0321876444  1.000000000  0.12991112
valence             -0.145199700 -0.017885943  0.0947941446  0.129911125  1.00000000
```

Annexe 8 : Carte des résidus du modèle de régression linéaire entre la popularité d'une chanson et l'IDH d'un pays