Rapport statistique :

Prédire les *certifications* de singles en France

Création d'un modèle capable de prédire si un single sera ou non certifié en France

OPENCLASSROOMS

2022

Anissa Mansour

Table des matières

INTRODUCTION

Ils enchaînent les milliers d'albums vendus, remplissent des salles de concerts en quelques jours, atteignent des millions d'abonnés sur les réseaux sociaux numériques, génèrent des chiffres d'affaires parfois exorbitants ; les artistes musicaux ont su créer un véritable business prolifique, leur permettant de vivre confortablement de leur passion.

Aujourd'hui, grâce au streaming, la musique n'a jamais été aussi facile d'accès et, fait alors partie du quotidien de tous, que cela consciemment ou inconsciemment. L'équivalent de 368 chansons d'une durée de 3 minutes est écouté chaque semaine dans le monde soit plus de 18h d'écoute par auditeur sur une semaine[1].

L'arrivée d'Internet a entrainé une remise en question des institutions. En effet, les nouveaux entrants ont créé un déséquilibre qui constitue un terrain propice à l'observation du changement institutionnel. Internet a laissé place à de nouveaux standards auquel il a fallu s'adapter. Les business models de l'industrie musicale doivent s'adapter aux innovations numériques mais doivent aussi prendre en compte le besoin des consommateurs. Les nouveaux modèles de streaming, comme YouTube ou Spotify, sont les meilleures réponses au piratage illégal des œuvres phonographiques. L'année 2016 marque une année charnière pour l'industrie musicale, le streaming musical sur YouTube, Deezer, Apple Music, Spotify, Dailymotion ou encore Tidal, a ainsi donné un nouveau souffle à l'industrie. Il est désormais facile de diffuser sa musique afin qu'elle soit disponible sur toutes les plateformes. Par conséquent, ces dernières années, les maisons de disques ont conclus de nombreux accords ces services de streaming, qui offrent aux fans l'accès et l'autonomie nécessaires pour écouter les artistes et les musiques qu'ils aiment.

Avec un chiffre d'affaires total de 21,6 milliards de dollars, l'industrie de la musique continue sa croissance non négligeable rendue possible par le streaming, qui représente désormais 62,1% des revenus mondiaux totaux de la musiques enregistrée grâce à ses 443 millions d'utilisateurs possédant un compte d'abonnement payant. Nous assistons à cette croissance dans le monde entier, les maisons de disques continuent de développer leurs activités dans de nouvelles zones géographiques pour rendre la musique accessible au plus grand nombre de fans, où qu'ils soient, ce qui prouve que notre industrie est aujourd'hui plus qu'elle ne l'a jamais été, connectée au monde entier[1].

Dans un contexte de massification de la musique, l'attribution de certifications n'a alors cessé de croître ces dernières années. Devenues un réel instrument marketing et argument commercial pour l'industrie phonographique, les certifications permettent de souligner les ventes des artistes. Les associations représentant l'industrie du disque ont mis en place un cadre officiel pour la remise de ces récompenses. La remise d'une certification est conditionnée à un seuil de ventes propre à chaque pays. La certification porte sur les ventes nettes de disques entre la maison de disques, les grossistes et distributeurs d'autre part. Le SNEP, depuis 1973, est l'institution chargée de vérifier et de remettre ces récompenses (or, platine ou diamant pour les singles). Les seuils de certification évaluent en fonction du marché du disque. En effet, lors de la crise du disque, *à la suite de la démocratisation d'Internet*, ces seuils ont été revus à la baisse. A partir de 2016, c'est la méthode de calcul qui a changé, ainsi, il faut, à présent, prendre en compte les écoutes en streaming. Par la suite, des révélations de tricherie dans le streaming ont

[1] Marché mondial 2020 – Publication du « Global music report » de l'IFPI

conduit à de nouvelles règles. En effet, seules les écoutes en streaming issues de la consommation premium (abonnements payants) sont prises en compte pour le calcul des meilleures ventes et des certifications.

Mais finalement, qu'est-ce qui explique qu'un single obtienne une certification d'or, de platine ou de diamant ?

Pour répondre à cette problématique, nous nous restreindront aux certifications attribuées en France. Nous avons identifié trois pistes de recherches résultant de cette problématique qui nous aiderons à traiter plus précisément le sujet :

- Quel est l'état des lieux des données fournies par le SNEP ?
- Qu'est-ce qui différencie un single certifié d'un single non certifié ?
- Quelles sont les variables déterminantes, et donc, nous permettent de créer un modèle permettant de prédire la certification d'un single ?

Ainsi, la première partie constituera une analyse des certifications SNEP remises depuis 1994. Ces données ont été récupérées directement sur leur site Internet sous forme de fichier PDF. Cette analyse nous permet alors d'avoir plus d'informations sur les variables suivantes : genres, artistes, labels, année et dates (dates de sortie et dates de constat de la certification).
La deuxième partie fera l'objet d'une analyse croisée entre les singles certifiées, ou non, en fonction de différentes variables, fournies par Spotify, liées à la structure de la musique. Nous avons donc créé un jeu de données constituées des chansons certifiées par le SNEP, mais également, de chansons non certifiées.

Enfin, la troisième partie, nous conduira à déterminer les variables déterminantes dans la certification d'un single.

Nous construirons plusieurs modèles qui seront testés afin de déterminer le plus fiable pour prédire la supposée certification d'un single en France.

I. Avant-propos : disques d'or, de platine ou de diamant – quelles différences ?

Depuis 1973, le Syndicat National de l'Édition Phonographique (SNEP)[2] s'est chargé de décerner les disques de certification en France. Le disque de certification est une récompense remise à un artiste pour souligner qu'un album ou un single a franchis certains seuils de ventes. Ces différents seuils de ventes sont déclinés en disques d'or, de platine et de diamant. Une fois qu'un seuil a été franchis, le SNEP effectue une vérification qui donnera le droit aux labels de demander à des entreprises spécialisées de fabriquer un objet représentant un disque d'or, de platine ou de diamant. Véritables symboles de réussite populaire et commerciale pour les artistes et les maisons de disques, ces objets sont réalisés en plusieurs exemplaires et distribués à l'artiste, les producteurs, les proches, la maison de disques voire à certains fans.

Les règles et seuils de ventes de singles en France n'ont de cesse d'évoluer en fonction du marché du disque. Avant 2016, les certifications étaient décernées au cas par cas, sous la demande des maisons de disques label et sous constatation d'huissiers, en s'appuyant sur les ventes hors taxes des singles. Dorénavant, les constatations se font automatiquement chaque semaine et les certifications sont générées en fonction des seuils atteints. A partir de cette même date, chaque téléchargement légal de single a été transformé en Équivalent Stream, sur la base de 1 téléchargement = 150 streams. Cette somme est ensuite ajoutée aux volumes de vente déjà connus pour le titre concerné.

Règles et seuils de vente singles

Création ou suppression des certifications
Evolution des seuils de certification

	SINGLES	OR	PLATINE	DIAMANT
Janvier 1973	Création des disques d'or	500 000		
Mai 1980	Création du single de platine	500 000	1 000 000	
Juillet 1985	Création du single d'argent	500 000	1 000 000	
Novembre 1988	Baisse des seuils des singles	400 000	800 000	
Mars 1991	Baisse des seuils des singles	250 000	500 000	
Janvier 1997	Création du single de diamant	250 000	500 000	750 000
Mai 2005	Baisse des seuils des singles	200 000	300 000	500 000
Juillet 2009	Baisse des seuils des singles	150 000	250 000	400 000
Janvier 2013	Baisse des seuils des singles	75 000	150 000	250 000
Janvier 2016	Automatisation des certifications (Physique + telech + streaming) En équivalent streaming	10 millions eq-stream	20 millions eq-stream	35 millions eq-stream
Avril 2018	Prise en compte des seules écoutes payantes augmentation des seuils En équivalent streaming	**15 millions** eq-stream	**30 millions** eq-stream	**50 millions** eq-stream

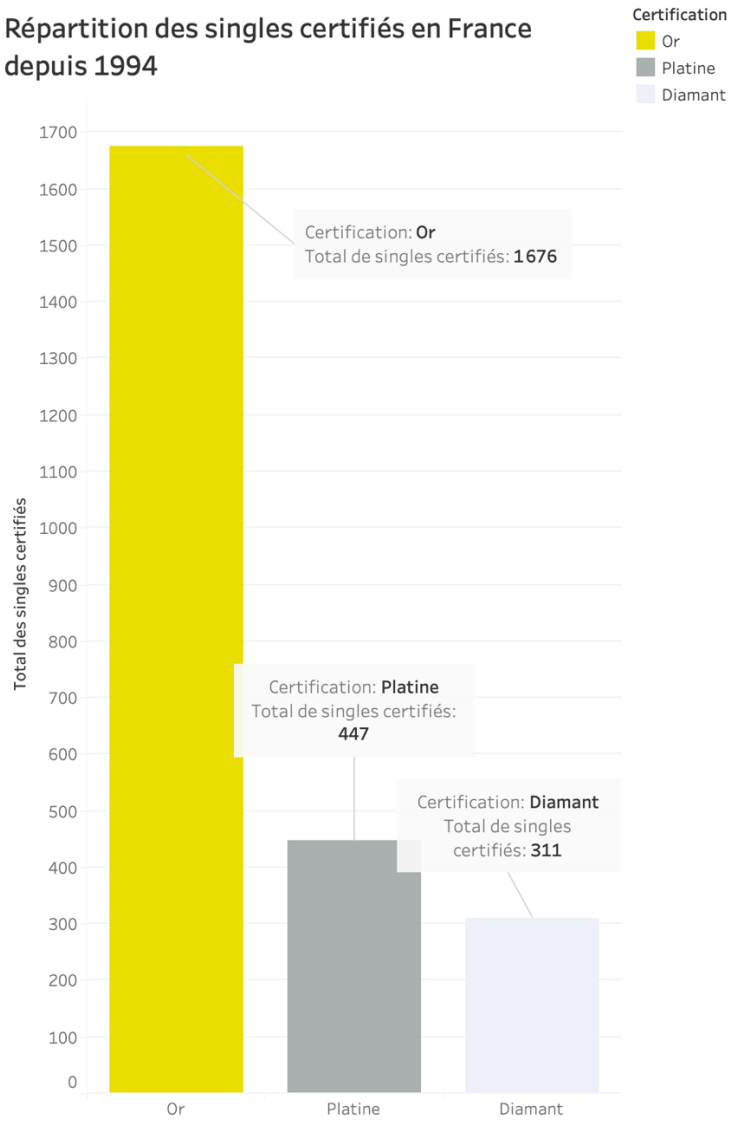
Source : SNEP – Règles et seuils de vente singles

[2] Le Syndicat national de l'édition phonographique est une association interprofessionnelle qui défend les intérêts de l'industrie française du disque phonographique depuis 1922.

Les règles ont de nouveau changé depuis 2018. Ainsi, pour qu'un single obtienne un disque d'or, il devra comptabiliser 15 millions d'Équivalent Stream. Pour un disque de platine, le seuil est fixé à 30 millions et enfin pour un disque de diamant, à 50 millions d'Équivalent Stream.

II. Certifications de singles en France

Le site du SNEP nous permet de récupérer les données concernant les certifications de singles attribuées depuis 1994. 2434 certifications ont ainsi été remises en France, soit 1676 disques d'or, 447 disques de platine et 311 disques de diamant.

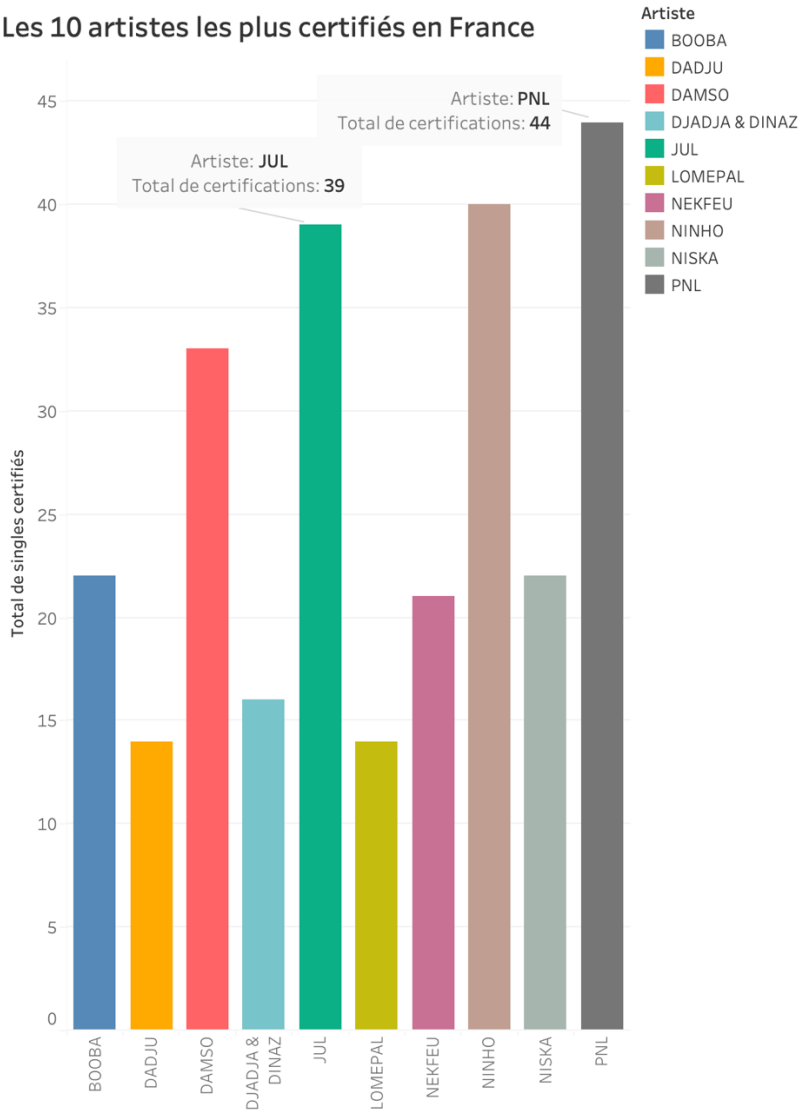


Graphique – Répartition des singles certifiés en France

De plus, les données disponibles nous ont permis d'analyser les artistes, les labels et également les genres musicaux les plus certifiés. Les graphiques ont été générés grâce à la plateforme Tableau.

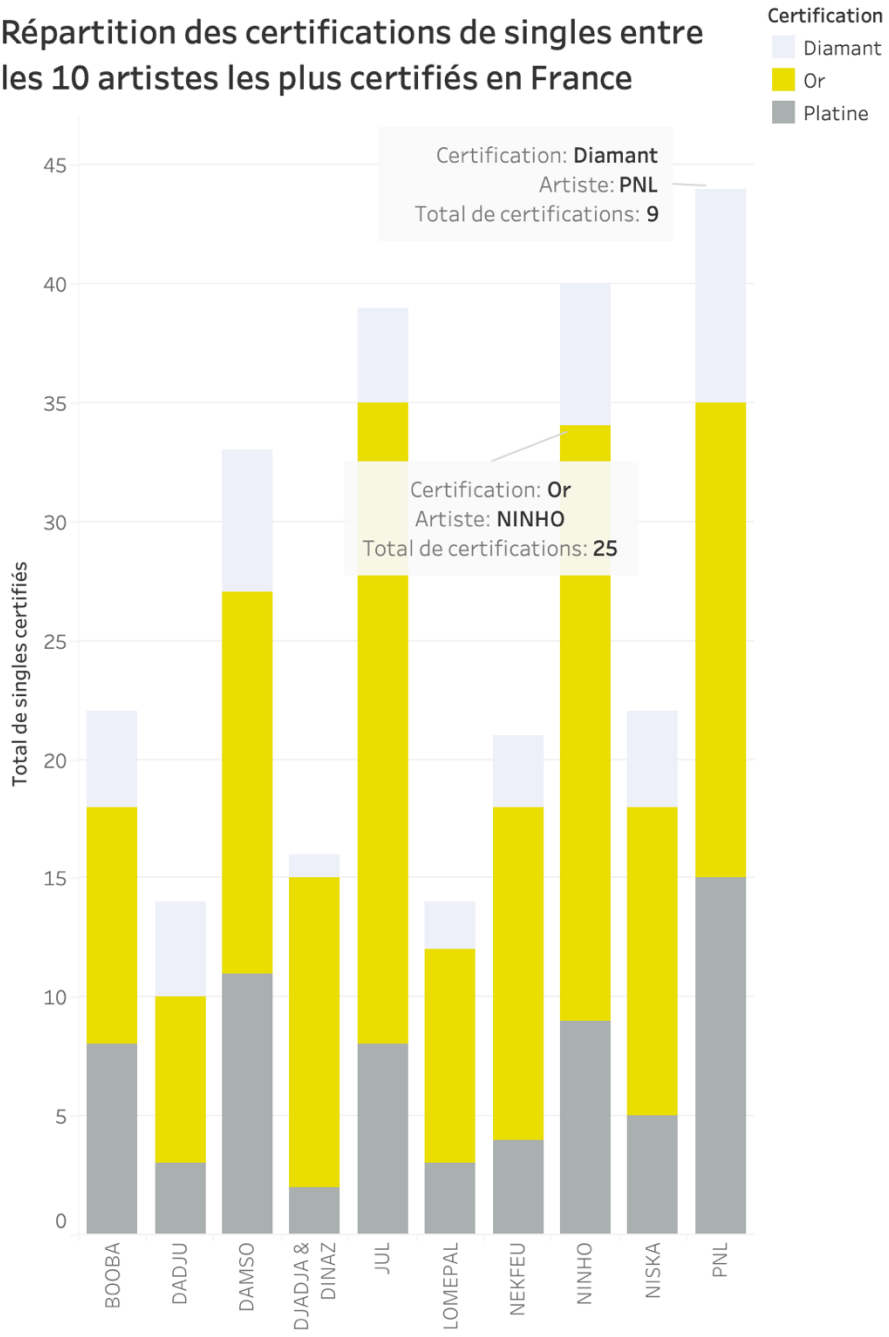
1. Les 10 artistes les plus certifiés

Au cours de leur carrière, les artistes peuvent obtenir plusieurs certifications. Elles deviennent donc un gage de notoriété et crée une curiosité de la part du public, mais également, des professionnels. La certification est souvent considérée comme un accomplissement personnel pour l'artiste, récompensant ainsi son travail et valorisant le poids de sa communauté.



Graphique – TOP 10 des artistes les plus certifiés en France

L'analyse de notre base de données montre que les cinq artistes les plus certifiés sont Ninho, PNL, JUL, Damso et Dadju, *quatre rappeurs et, un chanteur de francoton*[3]. A noter que ces artistes ont une carrière courte. En effet, ils ont moins de dix ans de carrière chacun. PNL et JUL sont des artistes dits « indépendants », ils n'ont pas l'appui de grandes maisons de disques, ce qui est donc un exploit contrairement aux autres qui sont sous contrats avec de grandes maisons de disques tels que Warner et Universal Music.



Graphique – Répartition des certifications entre les artistes les plus certifiés

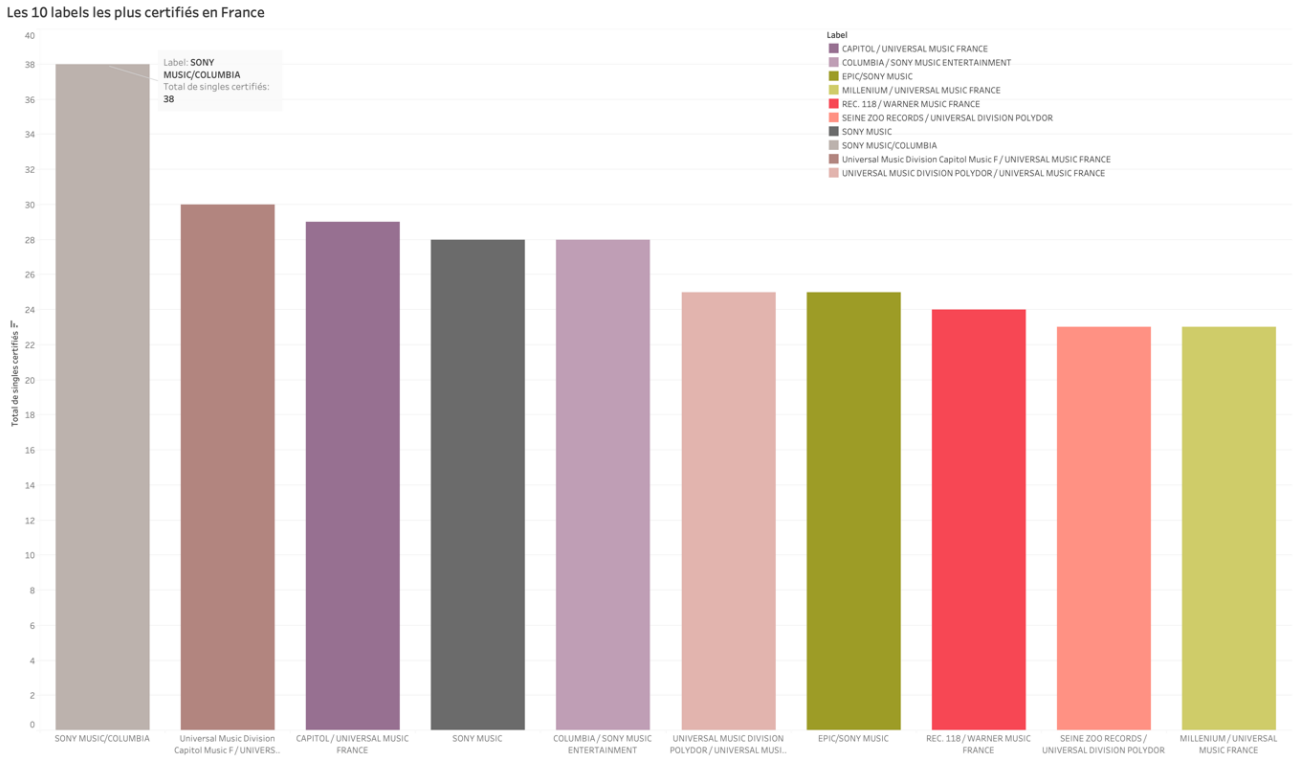
[3] Genre musical issu des musiques dites « urbaines », dérivées du R'n'B.

On constate ici que l'artiste ayant le plus de disques d'or est PNL avec 44 certifications, soit 20 disques d'or, 15 de platine et 9 de diamant. Le groupe est suivi par Ninho avec 40 certifications puis de JUL qui en comptabilise 40.

Il est important de souligner qu'un nombre important de certifications ne signifie pas forcément un grand nombre de ventes de singles. En effet, lorsque l'on se concentre sur le nombre d'Équivalent Stream, le top artistes est différent. Ainsi, PNL reste en tête et est par conséquent le plus gros vendeur de singles en France avec 780 000 000 d'Équivalent Stream. Cependant, on remarque l'arrivée de Niska (650 000 000) en deuxième position, dépassant Ninho, (480 000 000) suivi de près par le rappeur Orelsan (400 000 000), puis les artistes internationaux Ariana Grande (400 000 000) et Drake (360 000 000), la chanteuse belge Angèle (360 000 000) et, enfin, les rappeurs JUL (315 000 000), Damso (270 000 000) (pourtant dans le top artistes des artistes les plus certifiés), et Naza.

On notera que les Équivalent Stream de ces dix plus gros vendeurs en France représentent 27% des Équivalents Stream totaux.

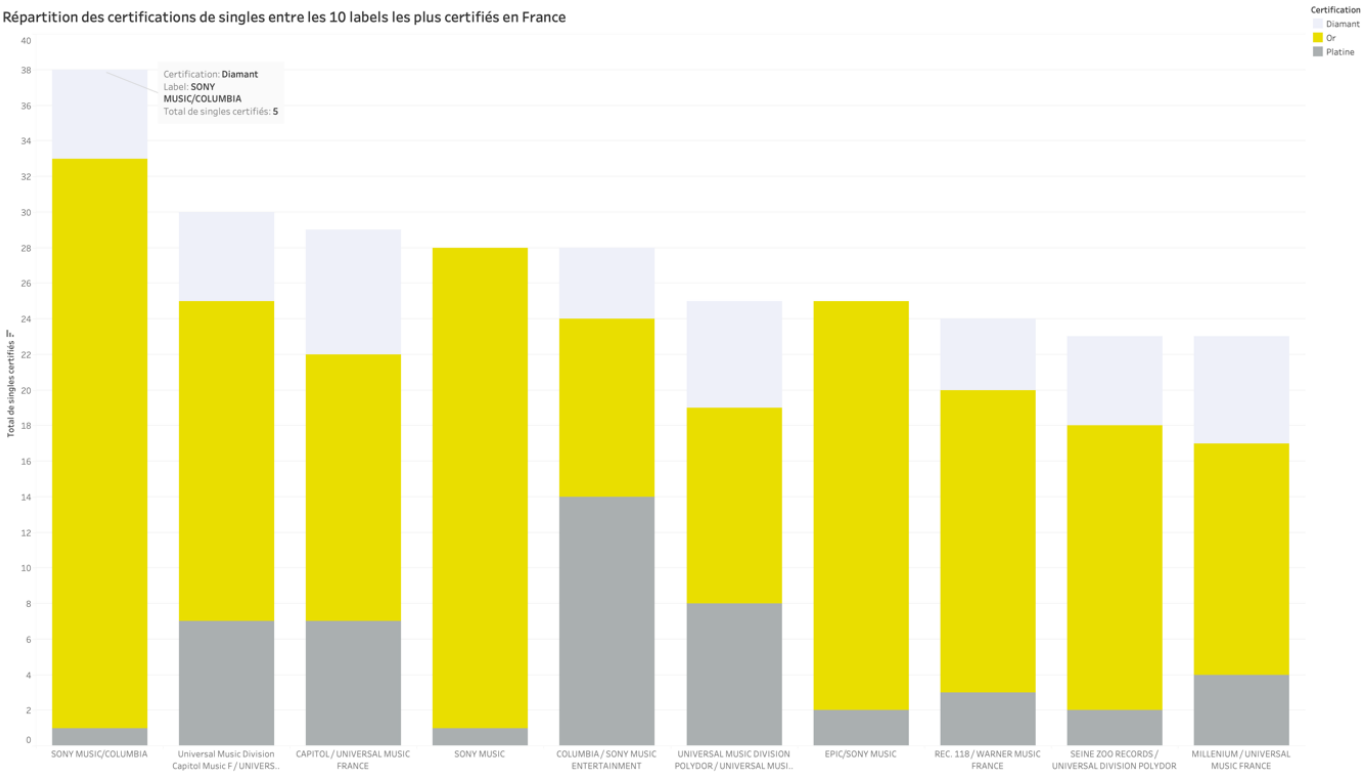
2. Les 10 labels les plus certifiés



Graphique – Les 10 labels les plus certifiés en France

Les dix labels les plus certifiés appartiennent majoritairement à des maisons de disques, ce qui montre l'influence qu'elles ont sur l'industrie du disque.

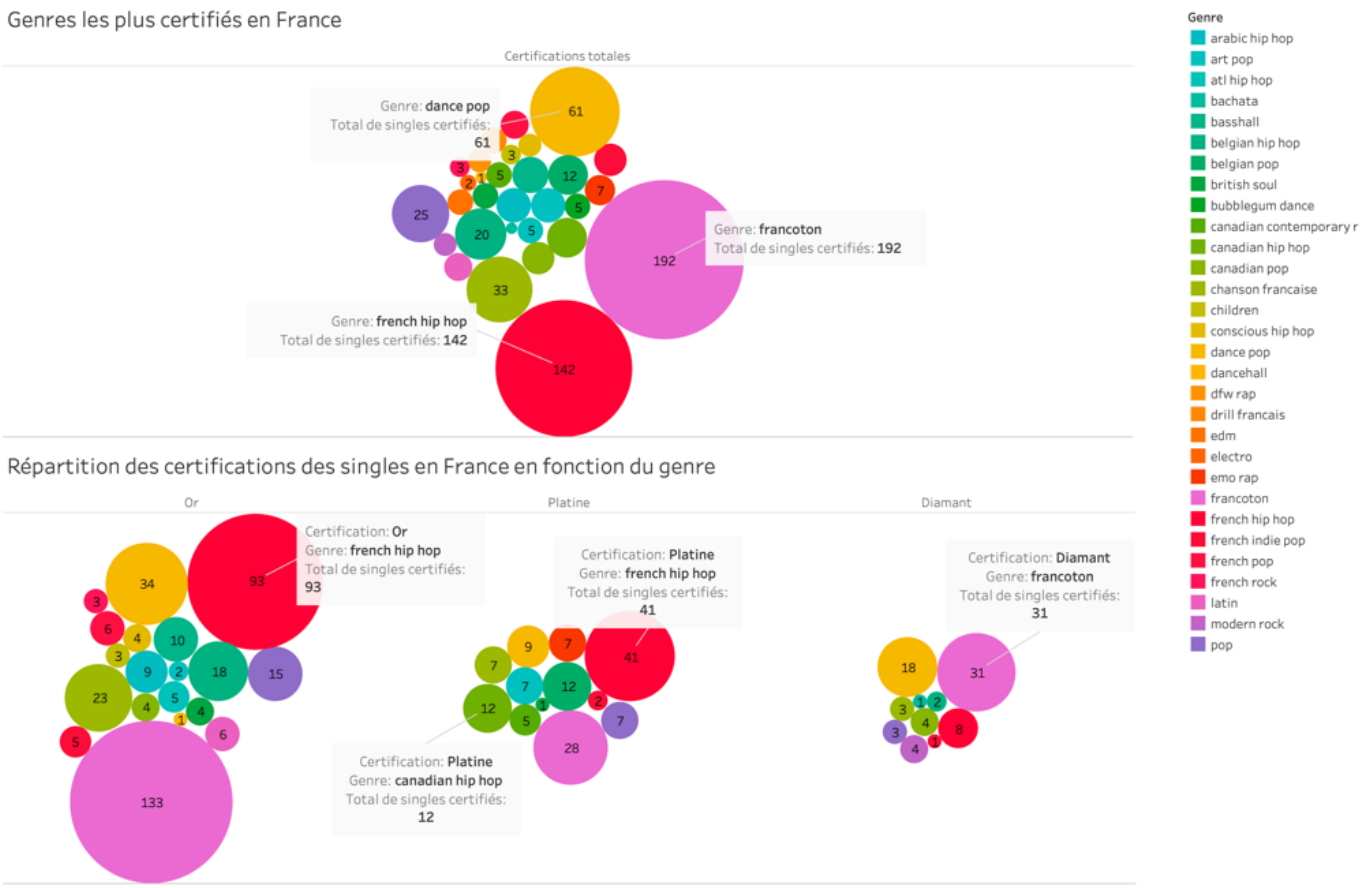
On constate que le label ayant le plus de singles certifiés compte moins de singles certifiés (38) que PNL (44). Cela s'explique par le fait qu'entre 1994 et aujourd'hui, il y a eu une grande diversité de labels, créés ou fermés mais également des changements de contrats des artistes.



Graphique – Répartition des certifications de singles entre les 10 labels les plus certifiés en France

Le label Columbia détenu par Sony Music est celui qui enregistre le plus de certifications de singles avec 48 certifications. Il détient également le plus grand nombre de disques d'or avec 32 singles d'or. De plus, la maison de disques Sony Music apparaît trois fois dans le top 10, faisant ainsi d'elle, la maison de disques la plus certifiée. Elle est suivie d'Universal Music et ses labels puis de Warner. Ces trois maisons de disques sont les plus importantes en France, mais aussi, dans le monde.

3. Les 30 genres musicaux les plus certifiés



Graphique – Les 30 genres musicaux les plus certifiés en France

Spotify a mis en place des catégories très précises pour les genres musicaux. Notre base de données comptabilise 137 genres musicaux différents, chiffre important mettant alors en avant la diversité musicale en France. Parmi ces genres, la chanson française, le french hip hop, la pop, le francoton ou la dance pop. En comparant ces données avec nos analyses précédentes, on peut mettre en évidence que les artistes les plus certifiés sont issus du hip-hop français et du francoton.

III. Constitution de la base de données

Afin de comprendre ce qui différencie un single certifié d'un single non certifié, nous devons constituer une base de données. Les caractéristiques de cette base de données seront étudiées dans cette partie.

1. Présentation des données récupérées

La constitution de la base de données a été permise grâce au téléchargement des certifications singles et albums du SNEP. Afin de rendre possible le comparatif entre les singles certifiés ou non, l'application web Exportify a été utilisée pour télécharger les singles variés non certifiés présents dans les différentes playlists de Spotify.

L'application web nous a permis de récupérer les caractéristiques audio, utiles pour notre analyse. Pour les singles certifiés, nous avons collecté les caractéristiques audios fournis par Spotify grâce à la librairie Spotipy. Ainsi, nous avons 36 variables pour notre base de données. L'analyse a été réalisée sous le langage Python grâce à ses librairies. Les graphiques générés eux, ont été réalisés sur Tableau.

2. Présentation des variables disponibles pour notre analyse

Comme évoqué précédemment, notre base de données contient 36 variables. Elles peuvent être divisées en 3 groupes :

- Les données concernant l'artiste et ses singles
- Les données concernant les certifications
- Les données concernant les caractéristiques audios des singles

a. Variables concernant l'artiste et ses singles[4]

Nous comptons ici 14 variables permettant d'obtenir des informations sur les artistes et leurs singles :

`artist_name` : le nom de l'artiste ; **`artist_popularity`** : popularité de l'artiste calculée à partir de la popularité de toutes ses chansons (la valeur est comprise entre 0 et 100, *100 étant le plus populaire*) ; **`followers`** : nombre de followers sur Spotify ; **`track_name`** : le nom de la chanson ; **`track_popularity`** : popularité de la chanson (la valeur attribuée est comprise entre 0 et 100, *100 étant le plus populaire*) ; **genre** : le genre musical de la chanson ; **`date_release`** : la date complète de la sortie du single ; **`day_release`** : le jour de la sortie du single ; **`month_release`** : le mois de la sortie du single ; **`year_release`** : l'année de sortie du single ; **`weekday_release`** : le jour de la semaine de sortie du single ; **`album_name`** : le nom de l'album ; **labels** : le label sous lequel la chanson est sortie ; **`explicit_content`** : contient des paroles grossières ou non.

[4] https://developer.spotify.com/documentation/web-api/reference/#/operations/get-an-artist

b. Variables concernant les certifications

Ces variables nous permettent des indications selon que le single est certifié ou non. On compte 7 variables :

is_certified : si la chanson est certifiée oui ou non ; **certif_single** : le seuil de ventes atteint par la chanson ; **certif_album** : le seuil de ventes atteint par l'album ; **nb_jours_certif_album** : nombre de jours qu'il a fallu avant la constatation de la certification de l'album ; **nb_jours_certif_singles** : nombre de jours qu'il a fallu avant la constatation de la certification du single ; **eq_streams_singles** : le nombre de streamings du single ; **eq_ventes_albums** : le nombre de ventes du single.

c. Variables concernant les caractéristiques audios[5]

La librairie Spotipy nous permet de récupérer les caractéristiques audios des chansons. Ces données sont établies par la plateforme de streaming Spotify. 15 variables nous permettent d'expliquer ces caractéristiques :

acousticness : accorde un nombre entre 0 et 1 (*1 correspondant à une chanson acoustique c'est-à-dire qu'il n'y a pas d'instruments électroniques modernes dans la composition de la musique* (exemple, guitare seule)) ; **danceability** : décrit la chanson comme étant dansante ou non en attribuant entre 0 et 1 (*1 est attribuée à une chanson très dansante*) ; **duration_min** : durée de la chanson en minutes ; **energy** : correspond à l'intensité et l'activité de la musique (une chanson énergique est ressentie comme rapide et bruyante comme les chansons appartenant au genre du death métal, 1 correspond donc à une chanson très énergique) ; **instrumentalness** : plus la valeur attribuée est proche de 1 moins la chanson contient de « vocal content » et donc de chants ; **speechiness** : détecte la présence de mots dans la chanson (les valeurs entre 0,33 et 0,66 montrent que la chanson contient aussi bien de la musique que des paroles) ; **key** : correspond à la clé de la chanson et donc la tonalité de la chanson ; **liveness** : détecte s'il y a du public dans l'enregistrement (par conséquent si c'est un enregistrement qui a eu lieu à un concert par exemple. Des valeurs proches de 0,8 montrent que la chanson est en live) ; **loudness** : intensité sonore, exprimée en décibels, est une valeur numérique représentant le volume sonore perçu par l'être humain (les valeurs sont généralement comprises entre -60 et 0 dB ; **mode** : indique si la chanson est en majeur (1) ou mineur (0), **tempo** : exprimé en BPM, correspond à la fréquence de pulsation permettant de construire les valeurs rythmiques ; **time_signature** : correspond à la signature temporelle. Cela diffère en fonction du genre musicale (*la valeur 4 est celle la plus commune*) ; **valence** : décrit la positivité de la chanson. Une chanson avec une valeur élevée correspond donc à une chanson positive tandis qu'une chanson avec une valence faible est négative (triste, dépressive, énervée) ; **end_of_fade_in** : exprimé en secondes,

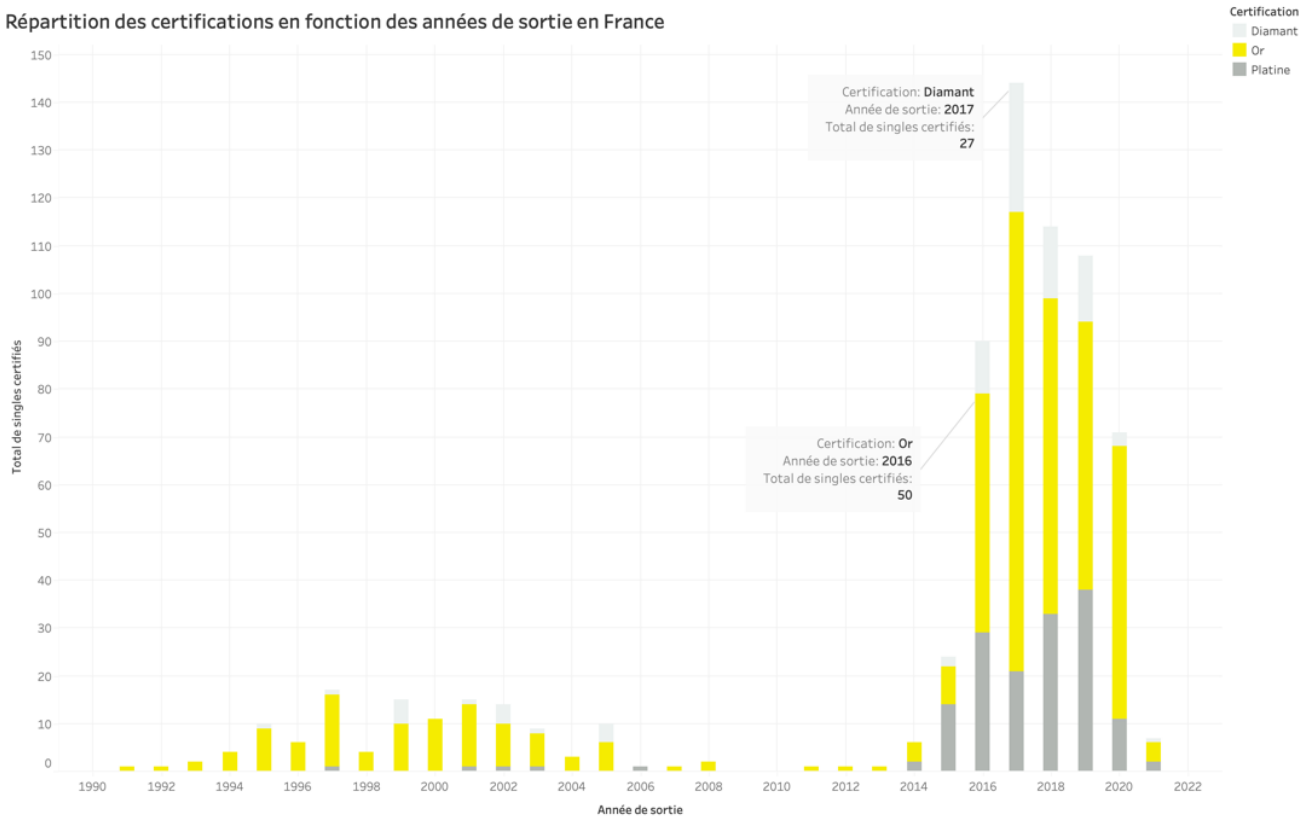
[5] https://developer.spotify.com/documentation/web-api/reference/#/operations/get-several-audio-features

correspond à la fin du moment où le son augmente ; **`start_of_fade_out`** : exprimé en seconds correspond au début du moment où le son diminue.

Toutes ces variables nous ont permis de réaliser une analyse complète de notre base de données. Nous disposons d'informations sur les sorties des singles, les genres, et les caractéristiques des singles. Tout cela nous permis également d'effectuer des comparaisons entre les singles certifiés ou non.

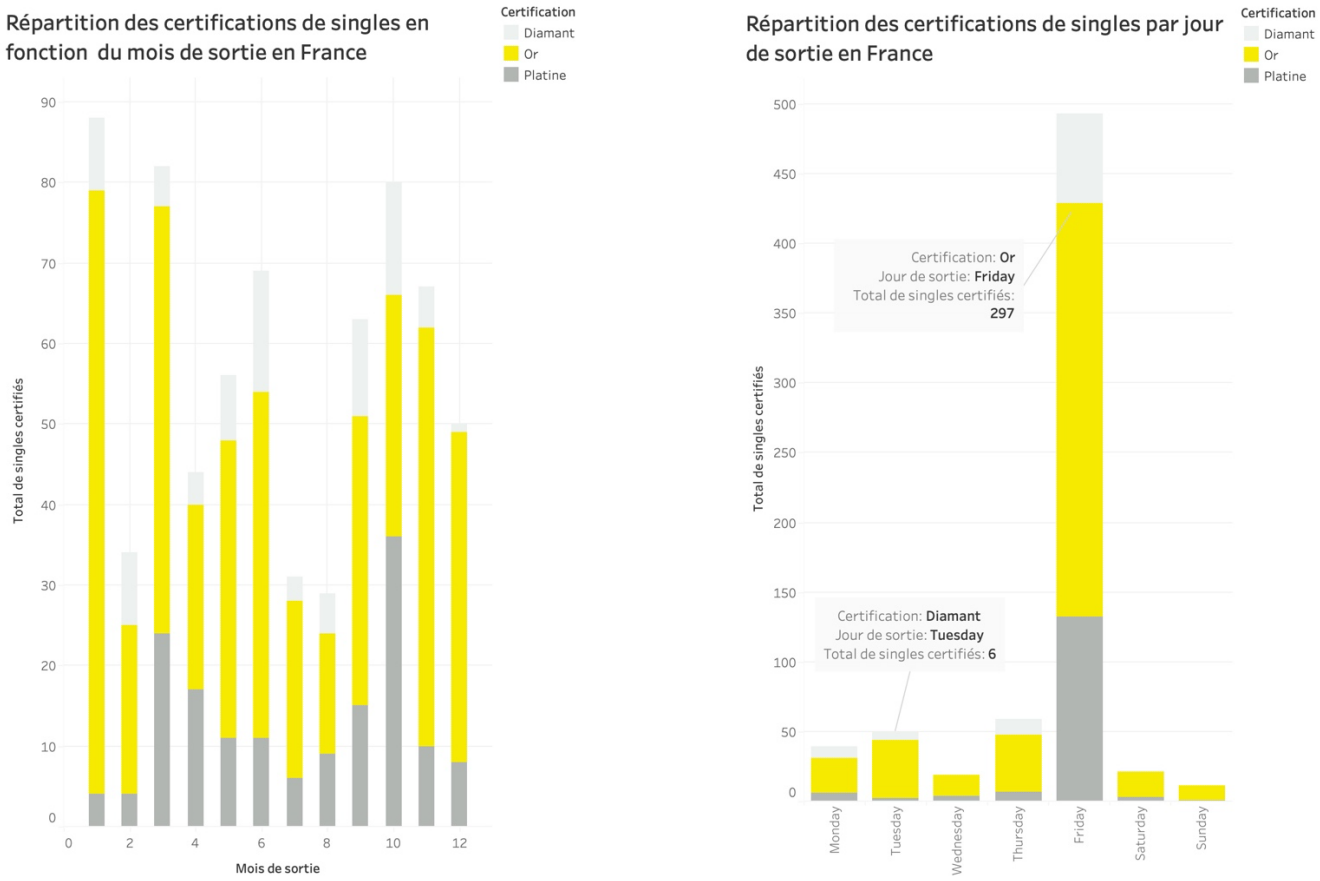
IV. Étude analytique

1. Analyse des dates de sortie des singles certifiés



Graphique – Répartition des certifications de singles en fonction des années de sortie en France

On constate une explosion des certifications sur l'année 2016. Cette année correspond à l'avènement de l'ère du streaming qui a été notamment bénéfique pour les artistes de hip-hop français, touchés plus particulièrement et de façon plus importante que les autres genres musicaux par le piratage illégal.



Graphiques – Répartition des certifications de singles en fonction du mois et du jours de sortie en France

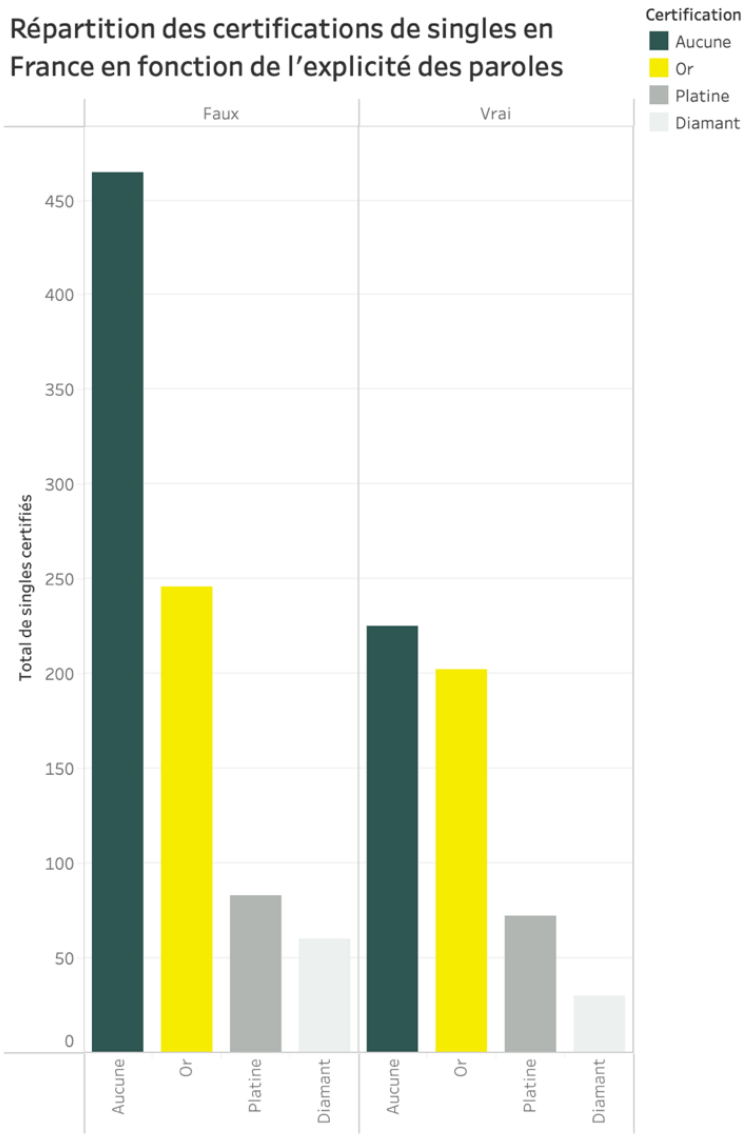
On constate que les singles sortis en janvier et octobre sont les plus certifiés. Et, plus précisément, les singles disponibles le vendredi sont ceux qui obtiennent plus facilement une certification. Cependant, il est important de souligner qu'un single sortira systématiquement le vendredi car, c'est le jour dédié aux sorties, au même titre que le mercredi pour les sorties cinéma. Il faut, par ailleurs, en moyenne moins d'un an (332 jours) pour obtenir un disque d'or, 614 jours pour un disque de platine et 802 jours pour un disque de diamant. Il n'est pas rare, surtout depuis l'arrivée des plateformes de streaming, que la durée d'obtention d'une certification de single est réduite. Un artiste peut donc obtenir ces certifications en une semaine.

On constate qu'un nombre notable de certifications est attribué un autre jour que le vendredi. Il serait alors intéressant pour les artistes de s'affranchir du vendredi afin de créer un engouement pour leur single qui serait peut-être passé inaperçu des auditeurs le vendredi, face aux grands nombres de sortie ce jour.

Cette tradition a été mise en place en 2015 afin de lutter plus efficacement contre le piratage. Malgré l'arrivée du streaming, celle-ci a perduré et, peut s'expliquer par deux facteurs : tout d'abord, le calcul de la première semaine de vente et, le positionnement sur les plateformes en ligne légales. Les classements prennent en compte les ventes du vendredi au jeudi, ce qui permet de bénéficier de comptabilisation de sept jours pleins. De plus, les diverses playlists mises en ligne par les plateformes de streaming sont actualisées le vendredi. Cela

permet alors aux artistes d'avoir une chance de placer leurs singles dans ces playlists. Ainsi, en sortant un single en fin de semaine, le potentiel dans les charts est maximisé[6].

2. L'explicité des paroles : un élément déterminant dans l'obtention d'une certification ?



Graphique – Répartition des certifications de singles en France en fonction des grossièretés des paroles

Notre graphique montre la répartition des certifications de singles selon l'explicité des paroles. On constate que la grossièreté des paroles n'influence pas la remise de certifications, il y a quasiment le même nombre de certifications de part et autre.

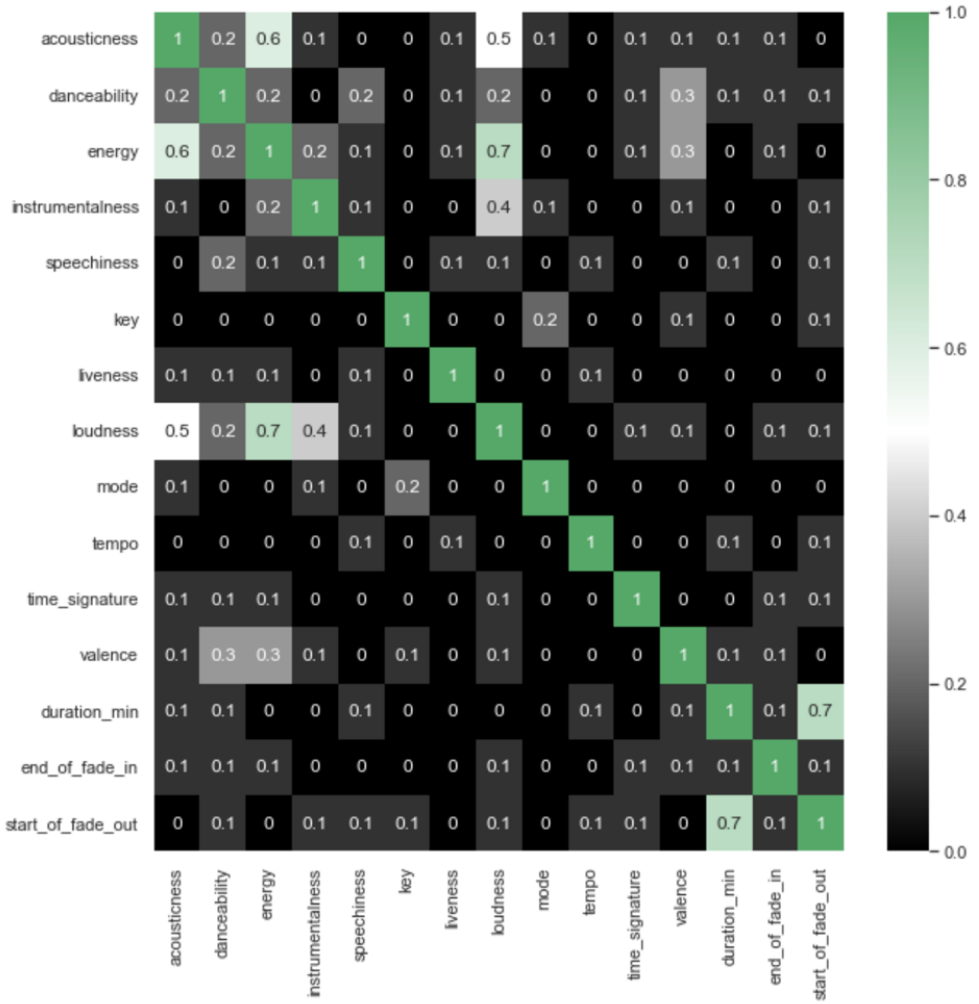
[6] *« Pourquoi les albums sortent le vendredi ? »* par Booska-p : https://www.booska-p.com/musique/rap-us/pourquoi-les-albums-sortent-le-vendredi/

3. Corrélations et variables déterminantes dans l'obtention d'une certification de single

Dans cette partie, nous allons découvrir et quantifier la relation entre les variables. Pour obtenir la dépendance entre deux variables, différentes méthodes sont appliquées en fonction du type des variables. Notre base de données contient des variables numériques et des variables dites catégorielles, nous utiliserons ainsi les coefficients de Pearson, l'ANOVA et les tests de Khi-Deux.

a. Corrélations entre les variables continues : coefficients de Pearson

Pour analyser les corrélations entre les variables exclusivement numériques (aussi appelées *variables continues*), nous utilisons le coefficient de Pearson. Une carte de chaleur permet graphiquement et, de façon plus efficace, de nous rendre compte des corrélations. Pour rappel, un coefficient égal à 0 signifie qu'il n'y a pas de corrélation. Un coefficient compris entre -1 correspond à une corrélation négative (lorsqu'une variable augmente l'autre diminue). Un coefficient proche de 1 correspond à une forte corrélation positive.



Graphique – Représentation graphiques des corrélations entre les variables numériques

Nous constatons que la variable **energy** est corrélée positivement avec la variable **acousticness**. Concrètement cela indique qu'une chanson acoustique est moins énergique, à contrario une chanson très énergique n'est pas acoustique. De plus, on constate que la variable **energy** est corrélée à la variable **loudness**. Plus une chanson est énergique plus l'intensité sonore augmente et donc, les décibels sont plus élevés, et inversement. On peut compléter cette analyse grâce à la variable **acousticness** qui est corrélée également à la variable **loudness**.

Par conséquent, plus le son est acoustique, plus l'intensité sonore est faible, et inversement.

La variable **loudness** est aussi corrélée à la variable **instrumentalness**, plus la chanson contient de chants plus l'intensité sonore est importante. Ainsi, on peut conclure grâce à ces observations qu'une chanson acoustique est moins énergique, a une intensité sonore moins élevée et contient moins de paroles.

Aussi, la variable **valence** est corrélée aux variables **danceability** et **energy**. Cela signifie que plus la chanson est positive, plus elle est dansante et plus énergique, et inversement.

Enfin les variables **start_of_fade_out** et **duration_min** sont fortement corrélées. Ce qui est normal car par définition, la variable **start_of_fade_out** correspond à la minute où le son de la chanson diminue.

b. Corrélations entre variables continues et catégorielles : ANOVA

Pour étudier les corrélations entre les variables continues et catégorielles, nous avons recours à l'analyse de la variance (ANOVA). Pour accepter les résultats de l'ANOVA, certaines conditions doivent être validées : la normalité des variables et l'homogénéité des variances. Ici, les conditions ne sont pas validées, nos variables ne suivent donc pas une distribution normale et leurs variances sont hétérogènes. Nous avons donc procédé à une alternative de l'ANOVA pour déterminer quelles variables sont significatives. Un test de Kruskal-Wallis a été réalisé.

Au sein de cette partie, nous chercherons à déterminer ce qui explique l'obtention d'une certification de single. Pour cela, notre variable de référence sera **is_certified_single**, elle sera testée avec nos variables continues pour connaître l'influence. Dans le test de Kruskal-Wallis, nous souhaitons accepter l'hypothèse nulle H0, qui affirme ici qu'il y a une corrélation entre les variables. Les p-valeurs des tests, c'est-à-dire les probabilités d'obtenir les résultats observés des tests, doivent être significatives et donc supérieures à notre seuil d'acceptation de 5%.

Ainsi, nous constatons que les variables **energy**, **speechiness**, **key**, **liveness** et **time_signature** sont corrélées à la variable **is_certified_single**. Par conséquent, ces variables ont un impact sur la certification d'un single.

c. Corrélations entre les variables catégorielles : Khi-Deux

Le test de Khi-Deux permet de déterminer les relations entre les variables catégorielles. Dans ce test, l'hypothèse nulle affirme que « les deux variables testées sont indépendantes ». Puisque nous souhaitons connaître les variables dépendantes, les p-valeurs de nos tests doivent

être inférieurs à notre seuil de 5%. De plus, le coefficient Phi nous permet de mesurer l'intensité de liaison entre deux variables.

Nous souhaiterons savoir dans cette partie ce qui explique l'obtention d'une certification de single. Pour cela, notre variable de référence sera is_certified_single, elle sera testée avec nos variables catégorielles pour connaître l'influence.

Ainsi, les variables **`genre`**, **`is_certified_album`** et **`label`** sont les variables très fortement corrélées à notre variable de référence. La *popularité de l'artiste* et le *nombre de followers* sur Spotify ont une forte corrélation avec la variable **`is_certified_single`**. Quant à la *popularité de la chanson*, elle est une corrélation modérée sur l'obtention d'une certification de single. Cependant, le fait qu'un single contienne ou non des contenus explicites n'influence pas la remise d'une certification. Cela confirme ce que nous avons observé précédemment.

d. A quoi ressemble un single certifié ?

Grâce à nos différentes observations, nous avons pu dresser le portrait type d'un single certifié. Ces observations ont été complétées par une analyse de la distribution des données des variables continues grâce à des *violinplots* comparant les données des singles certifiés et celles des singles non certifiés.

En conclusion, un single certifié est plus dansant, énergique, contient plus de chants et de positivité. Un singe certifié n'excède pas 7 minutes, *ce qui est le cas des singles non certifiés* ; en effet, il dure en moyenne moins de 4 minutes.

V. Prédire les certifications de singles en France

Pour ce projet, nous avons choisi d'estimer si une chanson sera certifiée ou non. Plusieurs algorithmes ont été testés, un choix final de modèle se portera sur celui qu'on estimera le plus performant. Notre jeu de données a été divisé en deux : 80% des données nous servirons à l'apprentissage et 20% nous servirons de test.

1. Comparaison entre les différents modèles de prédictions

Nos analyses précédentes nous ont permis de nous rendre compte de la significativité de nos variables. Notre modèle doit nous mettre de déterminer si un single sera potentiellement certifié ou non, en s'appuyant sur nos variables.

Il existe trois grands types d'algorithmes : les algorithmes *gradient descend-based*, les algorithmes *distance-based* et les algorithmes *tree-based*.

Les algorithmes *gradient descend-based*[7] permettent de trouver ou d'approcher un point stationnaire, point en lequel le gradient de la fonction à minimiser est nul. Le modèle que nous avons testé ici est la régression logistique qui permet d'expliquer le rapport entre une variable principale et des variables explicatives dans le but de prédire des valeurs. En d'autres termes, nous souhaitons savoir quelle est la probabilité qu'un single soit certifié en fonction des

[7] Définition de Wikipédia : https://fr.wikipedia.org/wiki/Algorithme_du_gradient

variables dont nous disposons notre base de données. Après avoir diviser notre jeu de données en deux, *une partie d'apprentissage et une partie de test*, nous avons obtenu comme résultats une accuracy[8] de 81% et une precision score de 81%, ce qui sont des résultats très satisfaisants. Les algorithmes *distance-based*, permettent eux de classifier les données en calculant la distance et la proximité entre les points. Nous avons testé ici deux modèles K Nearest Neighbor (kNN) et LinearSVC. Les résultats obtenus sont supérieurs à ceux de la régression logistique. On obtient 83% de accuracy score et 91 de precision score pour notre premier modèle ainsi que 81% de accuracy score et 85% de precision score pour le second.

Enfin, les algorithmes *tree-based* sont basés sur des arbres de décision. Ce sont les plus performants et les plus utilisés. Nous avons utilisé plusieurs modèles mais deux de ces modèles ont été les plus performants. Il s'agit de Random Forest et LightGBM.

2. Choix du modèle de prédiction

Le modèle Random Forest a obtenu les meilleurs résultats avec 85% de accuracy score et 91% de precision score. Ce modèle se base sur l'assemblage d'arbres de décision. Il est constitué d'un ensemble d'arbres de décision indépendants. La précision faite par ce modèle est la moyenne de tous les arbres[9].

C'est ce modèle que nous avons choisi ; 117 singles certifiés ont été correctement prédits (vrais positifs), 13 singles certifiés ont été prédits comme non certifiés (faux positifs). Concernant les singles non certifiés, 126 singles ont été correctement prédits (vrais négatifs) et 21 singles non certifiés ont été prédits comme certifiés (faux négatifs).

Nous pouvons également observer la pertinence de nos variables pour notre modèle, un graphique est alors généré. Ici, trois variables se détachent et sont les plus pertinentes : le nombre de followers, le genre musical et la popularité de l'artiste.

Le modèle créé ici, est un modèle universel. En effet, notre jeu de données regroupe des chansons de tous genres musicaux et de toutes époques confondues. En nous spécialisant sur un genre musical par exemple, nous pourrions améliorer un modèle plus performant.

VI. Conclusion

Lors de cette étude, nous avons effectué un état des lieux des certifications en France. Les observations ont montré que les trois genres les plus certifiés en France sont le francoton, le french hip-hop et la dance pop. Quant aux artistes les plus certifiés, ce sont Ninho, PNL et JUL des artistes issus du rap français. Le genre est ainsi devenu dominant en France, passant d'un genre confidentiel et craint, à un genre populaire parlant aux plus grands nombres mais surtout aux jeunes, cibles importantes du streaming. Il est donc commun de voir en tête des classements SNEP et de plateforme de streaming comme Spotify, des singles de rap français. PNL et Ninho sont les plus gros vendeurs de singles en France avec respectivement 780 000 000 Équivalent Stream et 480 000 000 Équivalent Stream. Suivi de Niska, qui compte 650 000 000 Équivalent Stream ; ces trois artistes détiennent alors le nombre de singles les plus vendus. Ces artistes les

[8] Définition : https://www.mathsisfun.com/accuracy-precision.html

[9] Explications : https://blog.ysance.com/algorithme-n2-comprendre-comment-fonctionne-un-random-forest-en-5-min

plus certifiés restent soutenus par de grandes maisons de disques, notamment Warner Music, ainsi qu'Universal Music, pour Ninho et Niska. Cependant, comme le montre notre analyse, des labels indépendants rivalisent avec les géants de la musique.

Dans la suite de notre étude analytique, nous avons constitué une base de données contenant des titres certifiés ou non. Ces titres sont décrits par 36 variables. Ainsi, le portrait d'un single certifié a été dressé : un single certifié est, plus dansant, énergique, contient plus de chants et de positivité. Il n'excède pas 7 minutes et dure en moyenne moins de 4 minutes.

Concernant les corrélations, le genre est la variable la plus corrélée à l'obtention d'une certification. En France, comme évoqué précédemment, trois genres se détachent. Un single a donc plus de chance d'être certifié s'il appartient au francoton, hip-hop français ou dance pop. Cependant, les autres variables telles que **`danceability`**, **`energy`**, **`liveness`** et **`speechiness`** exercent également une influence. Les singles certifiés sont plus énergiques. Un morceau énergique qui contient aussi bien des chants que des paroles, n'étant pas enregistré en live mais plutôt en studio d'enregistrement et, contenant moins de mots, a beaucoup plus de chance d'être certifié. Ainsi, notre analyse nous a permis de créer un modèle universel et fiable avec des métriques élevées permettant de déterminer, ou non, grâce à des variables données, si un titre a le potentiel d'une certification. Ce modèle constitue alors un support pouvant servir comme outil aux maisons de disque dans l'élaboration du budget marketing alloué au single. En effet, en identifiant précisément les caractéristiques du single grâce aux variables que nous avons déterminées, il serait alors plus aisé de prédire si le single mérite une mise en avant notable. Cependant, gardons à l'esprit, que le succès d'un single n'est pas uniquement explicable par les statistiques et, qu'un morceau qui respecte pourtant les caractéristiques mathématiques d'un single certifié, peut être mal accueilli par le public.