1. Introduction

The project, carried out as part of the course "SD701 - Big Data Mining", is presented in this report. Its objective is to answer a problematic by conducting a study on a big data set and using the methods and tools presented and studied as part of this course.

Our group, composed of three students, decided to analyse data on an everyday life subject which speaks to us : music. Music is the 4th art, and is defined, in the Oxford dictionnary, as the ability of combining sounds according to rules, which vary according to place and time. If we focus on the last 70 years, music has gone from being sparse entertainment to being ubiquitous. Music has developed at a very high speed and this democratisation has led to the appearance of new genres at very short intervals : blues, rock, funk, disco, pop, rap, R&B... and many other in less than 70 years. We have asked ourselves : what has evolved or changed in music over the last decades ? And, what are the main differences between nowadays music and the one produced in the 50's or 60's ?

If you are initiated to music, you know that the analysis of a piece has 3 parts : the composer and the context of composition of the musical track, its technical characteristics (tempo, energy, acousticness, among others...) and the lyrics. The context being a difficult variable to quantify or categorise, the demonstration is divided in two stages : the study of musical attributes over the years and a textual analysis of the lyrics. As with all volumetric data study projects, the first part will discuss data collection and pre-processing.

2. Data collection

In order to answer our problematic, 3 sources of data are needed :

- A list of tracks from approximately 1960 to 2020, representative of the kind of music listened by people at that time and with a large enough volume of music per time period to be able to draw conclusions
- The lyrics of those musical tracks to conduct an textual analysis
- The musical attributes (tempo, key, loudness...) of those tracks to compare their technical characteristics
  - acousticness : [0–1] Confidence measure of whether the track is acoustic.
  - danceability : [0–1] Describes how suitable a track is for dancing based on musical attributes including tempo, rhythm, stability, beat strength, and overall regularity.
  - energy : [0–1] Perceptual measure of intensity and activity. Energetic tracks feel fast, loud, and noisy (e.g. death metal : high energy, Bach prelude : low energy).
  - instrumentalness : [0–1] Predicts whether a track contains no vocals (values above 0.5 represent instrumental tracks whereas rap songs would have a score close to 0).
  - liveness : [0–1] Detects the presence of an audience in the recording.
  - loudness : [-60–0 dB] The average volume across an entire track.
  - speechiness : [0–1] Detects the presence of spoken words in a track (values above 0.66 describe tracks that are probably made entirely of spoken words, 0.33–0.66 describe tracks that may contain both music and speech, and values below 0.33 most likely represent music and other non-speech-like tracks).
  - valence : [0–1] Describes the musical positiveness conveyed by a track. Tracks with high valence sound more positive (e.g. happy, cheerful, euphoric), while tracks with low valence sound more negative (e.g. sad, depressed, angry).
  - tempo : [0–300 BPM] The speed or pace of a given piece, as derived from the estimated average beat duration.

3 data sources appeared to be consistent with our specifications and problematic : the US Billboard Hot 100 from 1958 to 2021, combined with an extraction of musical attributes from Spotify music API and texts from a lyrics database.

2.1. The US Billboard Hot 100

Originally, the billboard was a magazine whose first issue was planned in 1894 and dedicated to carnivals. Since 1958, the Billboard Hot 100 publishes every week the top 100 tracks based on single sales and how often they are broadcast on the radio in the US. Nowadays, it also takes into account the tracks downloaded from Itunes, Spotify or viewed on Youtube. The weekly ranking are available on the official US Billboard website. However, the collection of these data has not been performed by web-scrapping since the job had already been done and a .csv file was available on the Internet : Data.world

	url	WeekID	Week Position	Song	Performer	SongID	Instance	Previous Week Position	Peak Position	Weeks on Chart
955	http :	1/28/2017	1	Shape Of You	Ed Sheeran	Shape Of..	1	NaN	1	1
1020	http :	1/28/2017	6	Castle On..	Ed Sheeran	Castle On The..	1	NaN	6	1

TABLE 1 – US Billboard Hot 100 data frame

The .csv is data frame that contains 327,895 rows, which correspond approximately to the number of weeks between the 8 of February 1958 and the 29 of May 2021 (3,303 weeks) multiplied by 100. These 327,895 rows represent a total of 29,389 different songs, sung by 10,061 different artists. This data frame provides interesting attributes about song popularity that will be useful for our study such as : "Peak Position", "Weeks on Chart". Moreover, beyond these additional attributes, it is a song list representative of the musical tastes over the years and sufficiently large to rely on for musical attributes and lyrics collection and analysis.

2.2. Spotify API : musical attributes

One of the many benefits of the advent of streaming platforms, such as Spotify, is the tagging of millions of music tracks. In fact, Spotify offers a very useful API for downloading information about all the songs available on its platform. As per the Billboard Hot 100, the data collection of musical attributes of the Billboard list of tracks had already been performed and a .xlsx file is available on the Internet : Data.world

	Song	Performer	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo
834	Afire Love	Ed Sheeran	0.552	0.637	5.0	-6.568	1.0	0.0445	0.4640	0.000016	0.1360	0.333	97.970
870	Afterglow	Ed Sheeran	0.641	0.324	11.0	-5.851	1.0	0.0299	0.6980	0.000000	0.3280	0.273	110.184
2271	Barcelona	Ed Sheeran	0.747	0.760	1.0	-4.294	1.0	0.1870	0.4480	0.000000	0.1530	0.682	99.975
3891	Castle On The Hill	Ed Sheeran	0.461	0.834	2.0	-4.868	1.0	0.0989	0.0232	0.000011	0.1400	0.471	135.007
5500	Dive	Ed Sheeran	0.761	0.386	4.0	-6.158	1.0	0.0399	0.3550	0.000000	0.0953	0.526	134.943
6113	Don't	Ed Sheeran	0.798	0.675	6.0	-5.041	1.0	0.0442	0.0912	0.000000	0.0894	0.842	101.956
6617	Eraser	Ed Sheeran	0.640	0.812	8.0	-5.647	0.0	0.0834	0.0860	0.000000	0.0509	0.914	86.013

TABLE 2 – Spotify musical attributes data frame of US Billboard Hot 100 tracks

The data frame provides 20 new features that characterise each song such as : tempo, loudness, energy, acousticness, among others... As expected it contains 29,386 rows, corresponding to within 3 songs of the number of unique songs in the Billboard Hot 100 data frame.

As our study focuses on songs over the years, a merger is performed between these two data frames in order to obtain a data frame with song characteristics and add some calculated popularity features to it : average week position in the Billboard , the week the song got in and out of the ranking and the count of weeks on chart. The final data frame owns 25 features characterising each song of the US Billboard. The next step is to process this data.

2.3. Lyrics

To be more complete in our analysis, it was important to also look at the lyrics of the songs, and therefore fetch them from the Internet. The GENIUS database offers song lyrics and has a RESTFUL API to retrieve them and a Python client for the API exists on github. After generation of a token on the website it is possible to retrieve all the lyrics of most of the songs : out of 29,389 songs, 22,883 lyrics where recovered, a little over 75%. Instead storing them directly as a text, a word count was applied, resulting in a data frame similar to the one discussed previously, however with an added column containing an ordered count of the words.

3. Data Preprocessing

Our work organisation has led us to consider separately the lyrics and musical attributes analysis. Therefore, in the rest of the report, these two aspects will be addressed independently, but despite this, the two assesments will sometimes refer to each other.

3.1. Musical attributes

The objective is to convert this raw data, resulting from the merging of two tables found on the internet, into a data frame that it is possible to exploit for our analysis. The cleaning work is divided in 4 parts : columns cleaning, Nan values handling, types conversion and music genres mapping.

3.1.1. Columns cleaning

Our final musical attributes data set is composed of 27 columns. They are not all relevant in the frame of this data mining work and therefore two of them have been dropped : the *spotify_track_id* and *spotify_track_preview_url*.

Then all columns have been renamed.

3.1.2. Nan values handling

The first step when handling Nan value is to identify them in the data frame. Some famous *seaborn* plot allow a good visualisation of their location and occurrence in the data frame.

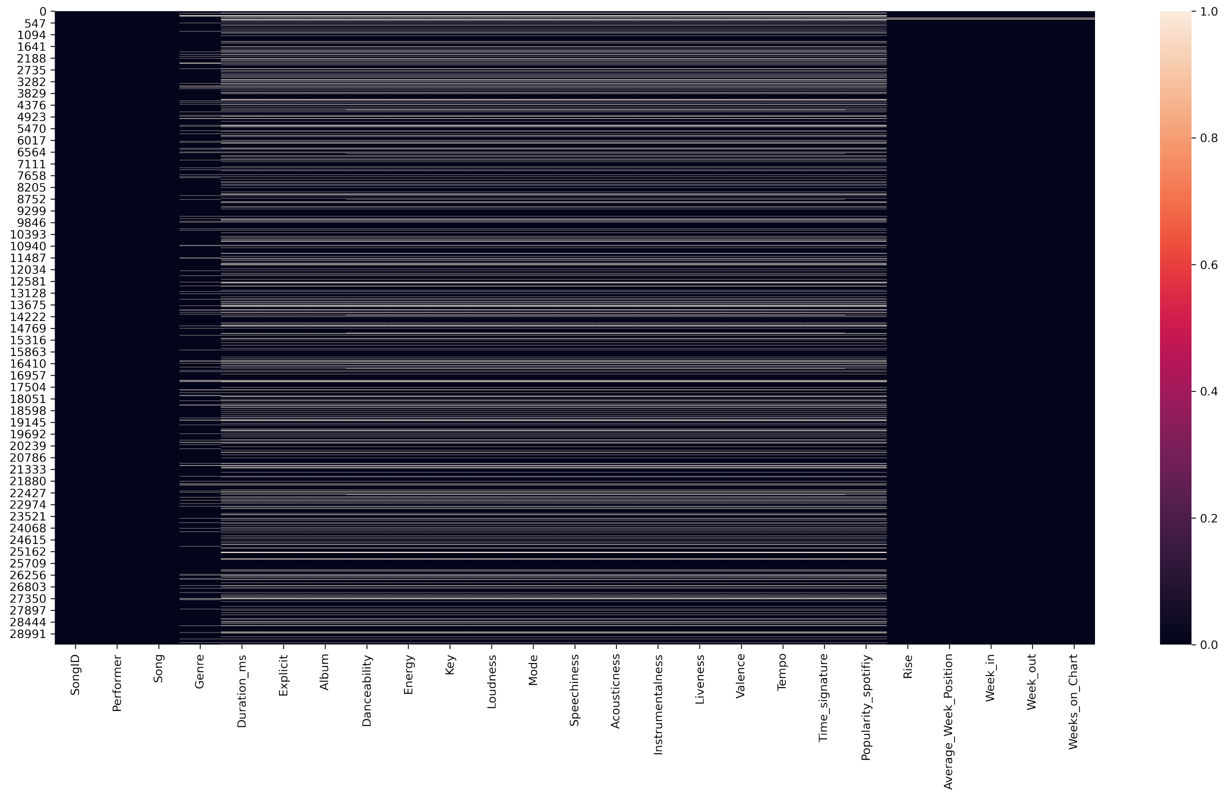

FIGURE 1 – Null values matrix

Thanks to this graph it is possible to identify the *patterns* of Nan values. The missing values correspond to musical attributes collected through Spotify music API. As these attributes are central to our analysis, the concerned rows (5,169 of them) have been removed. Then rows with Nan values in columns *Genre* and *Album* have also been removed, bringing the number of rows to 22,883. This loss of information was deemed to have little impact : indeed, our study focuses on the temporal aspect. Looking more closely at the impact of these rows deletions, it was observed that the average number of songs per year decreased significantly, from 460 songs per year to 354, but the variance also decreased, from 129 to 111. Less data but a more equitable distribution is an acceptable consequence of this pre-processing phase.

3.1.3. Types conversion

As the data frame was already a .csv, the only type conversion consisted in converting string to date format.

3.1.4. Music genres mapping

Our analysis relies on the evolution of music over the decades and therefore music attributes are expected to vary according to time and thus genres. The data frame as it is propose, in the column *Genre*, a list of genres for each music [3].

	Genres
0	adult standards, brill building pop, easy listening, mellow gold
1	rock-and-roll, space age pop, surf music
2	dance pop, pop, post-teen pop
3	pop, post-teen pop
4	album rock, bubblegum pop, country rock, folk rock, mellow gold, new wave pop, soft rock, yacht rock
5	country, country dawn, nashville sound
6	funk, motown, neo soul, new jack swing, quiet storm, r&b, soul, urban contemporary

TABLE 3 – Raw genres lists

As it can be seen in the table above, the possibilities for referencing the genres of a song on Spotify seem limitless. These lists are not very accurate as they sometime include too many genres : music n°3, for example, has 8 genres, and they can be very different from each other or far too detailed. Indeed the data frame contains 1042 different genres. In order to reduce their number, a mapping work has been performed. First of all, the genres represented more than 200 times in the entire data frame have been identified [4].

	Genres	Count_genre
0	mellow gold	3690
1	soft rock	3514
2	adult standards	3213
3	rock	3048
4	dance pop	2927
5	pop	2898
6	brill building pop	2872
7	soul	2551
8	motown	2481
9	pop rap	2421

TABLE 4 – Extract of genres count

Then internet researches have made it possible to rename the genres by assigning a more general category. For example : *Motown*, defined in the Oxford Dictionnary as a type of African American soul music has been replaced with *Soul* genre.
Thanks to this work, the number of genres has been drastically reduced from 1042 to 56.

3.2. Lyrics and API management

When connecting two different tables via an API, getting a perfect match wasn't ever expected, some songs where not loaded because they where not contained in the lyrics database and other simply failed for unknown reasons. Fetching all the songs took approximately 4 hours due to this error handling and the ability to restart at the point just before the failure had to be managed. Therefore two safety precautions where put in place :

- Saving data regularly, this involves saving the dataframe every 30 songs lyrics download to a local pickle file and starting with this file when the code is first run
- Adding a status *Lyric_status* to all song determining how the lyrics count went. The status issue has been handled using try-catch clauses :
  - *done* when the count went smoothly
  - *Failed_not_found* when the lyrics could not be found
  - *Failed_unknown* when an unknown error occurred, a sort of "catch all" clause

Once all the songs where installed and counted, it was noticed that a string *EmbedShare URLCopyEmbedCopy* was attached to all song lyrics, and was **??** therefore removed from all counts.

4. Analytical processing and algorithms

The data collection and pre-processing are fundamental parts of a data mining project as they are the foundation on which the following analysis are based : on the one hand the analysis of musical attributes and on the other the lyrics.

4.1. Musical attributes analysis

The musical attributes review is divided in two parts. The first, consists in an exploration of the features. The second, is an unsupervised clustering analysis. They both serve the same goal : identify whether songs of the same periods of time have similarities or not.

4.1.1. Features exploration

Features exploration is an elementary task in a data mining study. It can be used to highlight outliers or confirm a preconceived notion about a variable. Of the 25 features, 4 were selected and investigated [2].

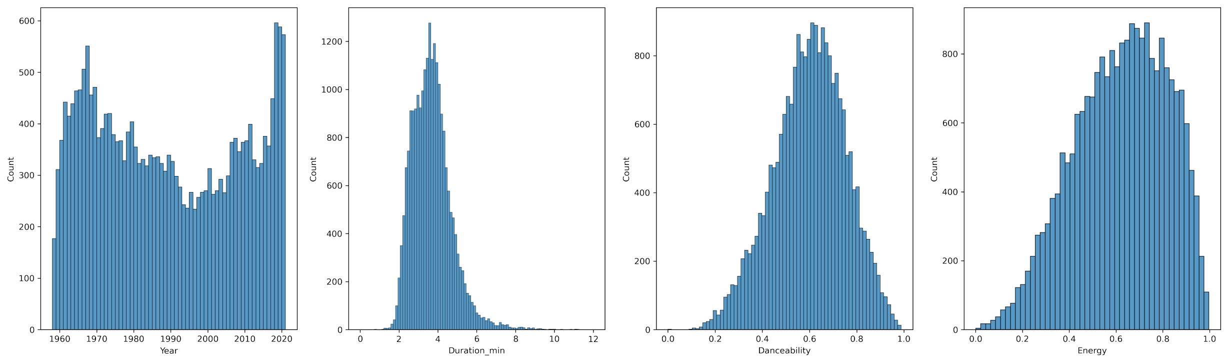

Figure 2 – 4 Features distribution

Several conclusions can be drawn from these graphs :

- *Year*, every year, from 1958 to 2021, is represented with at least 200 songs (full-years) and is therefore considered illustrative of the music consumed at that time.
- Duration, danceability or energy features confirm the idea that one had of it :
  - *Duration*, most of the song last between 3 and 6 minutes, and it is not a surprise since it is one of the fundamental principles of modern music.
  - *Danceability*, is a combination of tempo, rhythm stability, beat strength. The parameter is 1 on the music is danceable and 0 otherwise. As an example, *Around The World* from the *Daft Punk*

as a danceability of 0.956 while *When We Were Young* from *Adele* danceability is equal to 0.381. The spectrum is well covered, however the intuition that music is mostly upbeat and danceable is confirmed by this histogram.
  - *Energy* : energetic tracks feel fast, loud, and noisy. The parameter is 1 on the music is energic and 0 otherwise. For example, *Cyanide* from *Metallica* has an energy of 0.993 while *Nothing else matters* from the same group as an energy of 0.364. It explains why the variance of the energy variable is higher than the danceability one. However, this distribution was predictable if you are a fan of music from the 1950s to today.

Secondly, the average behaviour of the variables over the years is studied. It might highlight some key changes in the composition of music [figure 3].

The graphs confirm our assumptions about the evolution of music over the last 70 years.

- Danceability : as expected, songs have become more and more danceable over the years until the 90s and the advent of electronic music.
- Speechiness : detects the presence of spoken words in a track. The arrival of styles such as rap and hip-hop, where the lyrics have a predominant place, explains the important growth of this index over the last two decades. *Drake, Kid Cudi, Kendrick Lamar* are the most represented artists on this theme.
- Loudness : characterises the quality of a sound. The growth of this indicator is explained by the improvement of music recording equipment.
- Acousticness : measures the level of acousticness or synthetic means used for song production. The democratisation of samplers and synthesizers is clearly illustrated by this graph.

This last analysis of the behaviour over the years already provides some elements of response as to the possibility of characterising music according to the period, and that, relying solely on technical criteria.

Finally, to conclude this data exploration, the genre is studied. The idea is to calculate the dominant musical genre per year and ensure that the mapping has been consistent and that the data set is also representative of reality in this respect [table 5]. The results are very interesting since the table really highlights the chronology of musical genres in recent decades : *Brill Building Pop* in the 60's, *Blues* in the 70's, *Rock* in the 80's, *Hip-hop* in the 90's and until 2010, *Pop* and *Rap* from 2010 to today.

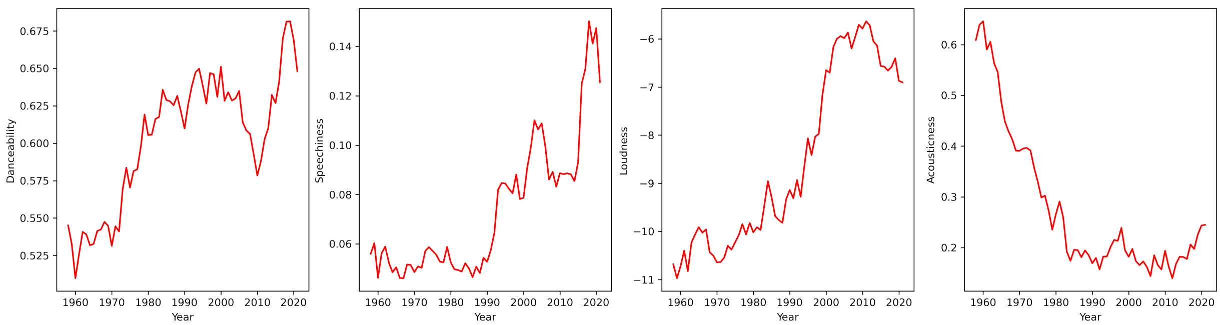

Figure 3 – 4 Features behaviour over the years

4.1.2. Clustering

K-means clustering is an unsupervised method that aims to partitions the data set into a pre-defined number of clusters based on the distance of each points to the different cluster's centroids. The idea behind the execution of a clustering algorithm on the data set, if we vulgarise it, is to observe if the songs composed over a same period are *close* to each other.

In order to have the best result for this model, some pre-processing has to be done on the data.

Year	Genre	Year	Genre	Year	Genre	Year	Genre	Year	Genre
1958	Rock	1968	Blues	1978	Soft Rock	1988	Soft Rock	1998	Hip hop
1959	Rock	1969	Blues	1979	Soft Rock	1989	Pop Rock	1999	Hip hop
1960	Brill Building Pop	1970	Blues	1980	Soft Rock	1990	Soft Rock	2000	Hip hop
1961	Blues	1971	Blues	1981	Soft Rock	1991	Rock	2001	Hip hop
1962	Brill Building Pop	1972	Blues	1982	Soft Rock	1992	Hip hop	2002	Hip hop
1963	Brill Building Pop	1973	Blues	1983	Soft Rock	1993	Hip hop	2003	Hip hop
1964	Brill Building Pop	1974	Blues	1984	Soft Rock	1994	Hip hop	2004	Hip hop
1965	Brill Building Pop	1975	Blues	1985	Soft Rock	1995	Hip hop	2005	Hip hop
1966	Brill Building Pop	1976	Soft Rock	1986	Soft Rock	1996	Hip hop	2006	Hip hop
1967	Brill Building Pop	1977	Soft Rock	1987	Soft Rock	1997	Hip hop	2007	Hip hop

Year	Genre	Year	Genre
2008	Hip hop	2018	Rap
2009	Dance Pop	2019	Rap
2010	Pop	2020	Rap
2011	Pop	2021	Pop
2012	Pop		
2013	Pop		
2014	Pop		
2015	Pop		
2016	Hip hop		
2017	Hip hop		

Table 5 – Most represented genre per year

K-means only support numerical data. The features chosen for the clustering are the following : *'Duration_ms', 'Explicit', 'Danceability', 'Energy', 'Key', 'Loudness', 'Mode', 'Speechiness', 'Acousticness', 'Instrumentalness', 'Liveness', 'Valence', 'Tempo', 'Time_signature'*. These features are attributes of a music and have, if taken one by one, little connection with an era or a style of music.

Another step with large data set is to reduce the number of variable using the PCA method. Principal Composant Analysis is a dimensionality-reduction method used to transform large data set of features into smaller one containing most of the initial information. It is necessary to scale the data before using PCA. Both «MinMaxScaler » and «StandardScaler» from the Scikit Learn library have been tested with the best result (regarding the elbow method) for the «MinMaxScaler ».

	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Explicability %	30.6	47.9	62.2	76.2	83.8	88.2	91.3	93.8	96.1	97.7	98.5	99.2	99.7	100.0

As shown in the previous figure and table, the first nine components will be kept for the rest of the analysis. These principal components synthesise more than 95% of the initial features. In other words by only keeping these 9 components we are able to keep 95% of the information contained in the initial data set. This change in dimension will improves the performance of the algorithm applied to the data.

After this pre-processing steps, the new data set is ready for the K-means clustering. The most appropriate value for « k » has been found using the Elbow method and the « inertia score » for the evaluation score [figure 5].

Even if the « Elbow » is not really obvious, the choice was made to take a number of cluster equal to 4, giving us the following distribution in each clusters [figure 6]

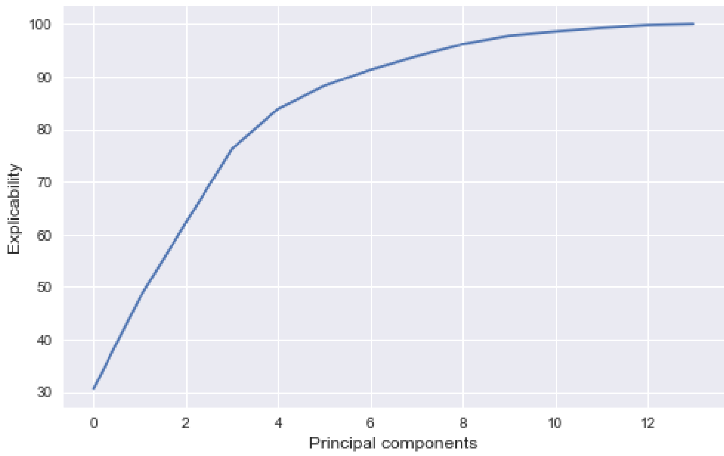

FIGURE 4 – Explicability of principal companants

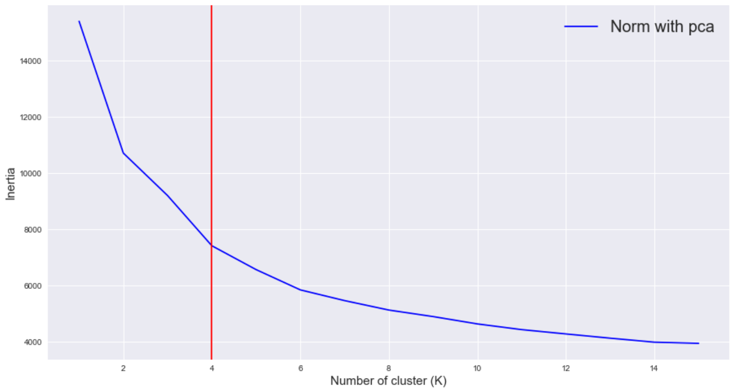

FIGURE 5 – Elbow curve for multiple K

For the moment, nothing much can be said about each cluster. The rest of the analysis will aim to characterise these cluster and their spread.

The graph [7] answers the objective of characterising the Billboard songs over the years. Indeed, following the evolution of the clusters would perhaps shine a light a their correlated evolution to time and finally highlight a link between musical attributes and period.

	Cluster 0	Cluster 1	Cluster 2	Cluster 3
Genre n° 1	Brill Building Pop	Rock	Soul	Hip hop
Genre n° 2	Rock	Soft Rock	Pop	Rap
Genre n° 3	Blues	Pop	Rock	Trap

TABLE 6 – Top 3 of represented music genres in clusters

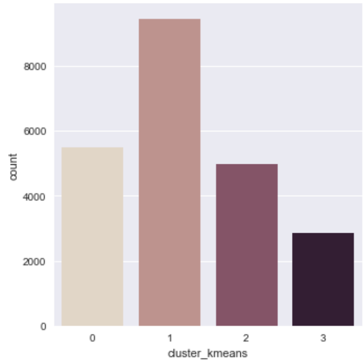

FIGURE 6 – Clusters count

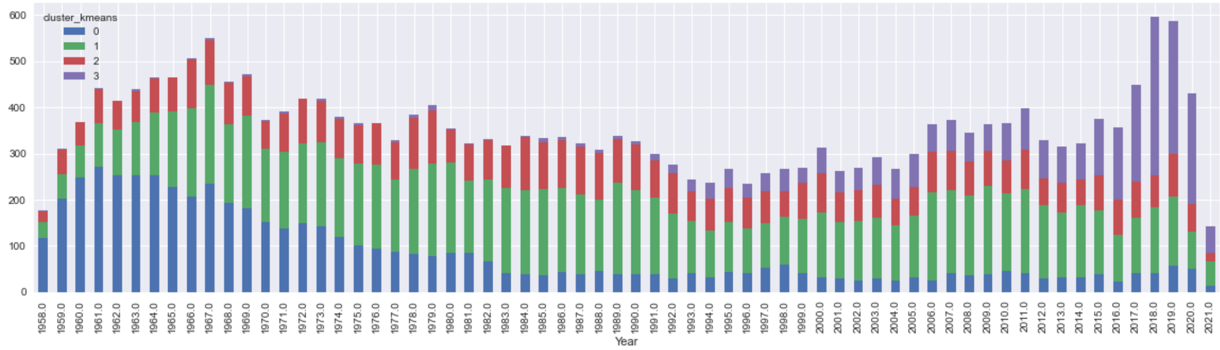

FIGURE 7 – Cluster evolution over years

Something interesting is shown in the previous graph [7]. On the one hand, since the 60's, cluster 1 and 2 shares a real constancy over time, while cluster 3 seems to appear when cluster 0 begins to disappear. By linking the cluster to the genre that is most represented within it [table 6] some interesting statements can be done. As discussed earlier, a link can be drawn between clusters and genres. It is clear that a cluster 3 genre's (*Hip hop*, *Rap* and *Trap*) have gained popularity in the 90s. Whereas the most important genre in the 60's : *Brill Building Pop*, *Rock* and Blues from cluster 0 are slowly losing popularity among the top 100 listened songs.

Another way to look at these clusters is to investigate them at the features level. This could help to understand musical attributes and people tastes evolution over time.

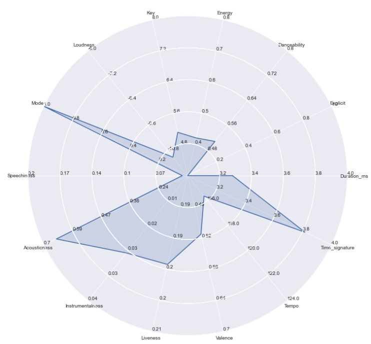

FIGURE 8 – Features characteristics of cluster 0

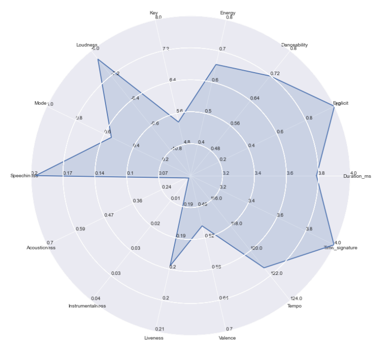

FIGURE 9 – Features characteristics of cluster 3

As shown in the two radar plots [8] and [9] , the clusters 0 and 3 are really different from each other. They show a real evolution in the composition of music. Cluster 3, that represents *Hip hop and Rap* has an average level of *Speechiness* higher than the one of cluster 0 (*Rock and Blues*). It is also possible to observe that *Accousticness* and *Instrumentalness* are no longer relevant in music. They gave way for characteristics like *Energy*, *Speechiness* and *Tempo*.

This musical attributes analysis confirms an intuition : music from the same period are linked to each others by musical attributes : it is indeed what characterises music genres. So, what has evolved in music since the 60's ? Well... a lot. However, could we say the same regarding the lyrics of the songs ?

4.2. Lyrics analysis

FIGURE 10 – Wordcloud of most common words found in songs

The first thing that could hopefully shed some light on the evolution of music lyrics throughout the years was the top words per year, this would highlight the words used by artist that made into the top chart. Maybe indicating a change in trends through the years. This was done in a two step strategy, first identify the top 5 words per song, and then weighing the top songs per year, in other words,

1. Select all songs corresponding to a year
2. Take top 5 words from each song
3. Add the top 5 songs to the year count, with a weight (for example add the first song five times and the second four times).

A random sample of the result is shown in table [7], the full data is in the appendix [**??**] and a word cloud of the most prevalent words are shown in figure 11.

	0	1	2	3	4	5	6	7
1968	love :271	im :173	baby :146	oh :128	dont :121	know :103	yeah :93	like :81
1969	love :296	oh :183	baby :183	im :175	dont :161	know :102	got :91	verse :90
1980	love :381	im :192	dont :151	oh :119	like :113	know :95	baby :76	youre :65
1982	love :322	dont :133	im :108	want :85	baby :74	got :68	know :66	one :65
1985	love :222	im :172	dont :126	oh :93	know :82	got :80	baby :70	youre :70
1987	love :230	im :125	dont :123	oh :119	baby :93	know :71	night :58	youre :55
1993	love :175	im :160	dont :92	like :78	baby :77	know :69	girl :49	come :47
1998	love :183	im :162	know :118	dont :108	baby :104	like :98	wanna :67	youre :59
2002	im :217	love :127	dont :111	know :105	like :78	one :70	got :64	yeah :57
2004	im :203	love :134	like :116	dont :94	know :87	yeah :74	get :53	one :45
2010	im :305	like :163	love :144	oh :115	dont :111	baby :94	yeah :91	get :64

TABLE 7 – Wordcount

It was quite evident by simply analysing the data that the same words come up very regularly, the words *love*, *im* & *dont* being present in every single year, it would seem there is very little evolution in the words used throughout the years. This being counter-intuitive, multiple attempts were made to detect an evolution over the years.

The first attempt involved looking at each year's top words divided by the total count of top words,(a preview of which is shown in table 8 this would essentially remove all the top words from the ranking and only show words that a very prevalent for the the year in question. The resulting data however was very erratical, words that where only present in a single year would be over-emphasised making it impossible to show any evolution throughout the years.

	love	im	dont	like	know	oh	yeah	baby	got
count	13105	12907	8258	6937	6010	5968	5942	5853	4582

TABLE 8 – Top words all years mixed

To counteract this effect a filter would be placed removing all the lower count words from the results, In other words take the top X songs from a certain year and divide it by the total amount of times it comes up all years mixed. This would give correct results however the resulting words where still very similar year on year. The final attempt at trying to show trends was to perform a TF-IDF on the words and look at the evolution, this result in similar conclusions to the previous attempts.

Even thought the top words where fairly similar it was still possible to look at small variation in the ranking and evolution, Figure 11 show the evolution of a couple of the most common words.

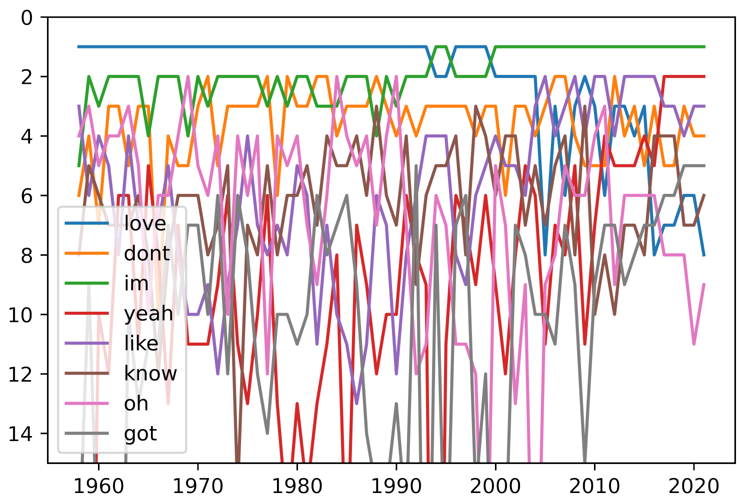

FIGURE 11 – Network graph showing years linked with the same words in their top count

Two notable conclusions can be drawn from this graph, the first is just how prevalent the word *love* is in lyrics, even thought in recent years the word has been losing a bit of its might. It has held the top rank from 1960 to 1995 and stays in the top 10 in the XIth century, the word cloud fig 11 emphasis this point further. The second point is the evolution of music in the XIth century, the words *yeah* & *got* gain prevalence, this is likely due to evolution of hip-hop in the era, discussed in the previous part.

The original ideas of looking at the lyrics was to try to perform a graph analysis, by linking years together. It was hoped doing so would show patterns. For the same reasons described above simply creating a network with a link joining two words would result in a graph with all nodes connecting to all other nodes, however by heavily tweaking the parameters it is possible to generate Figure 12. This graph was generated by selecting the 74 weighted top words for each year and creating a link between two years if they have a word in common.

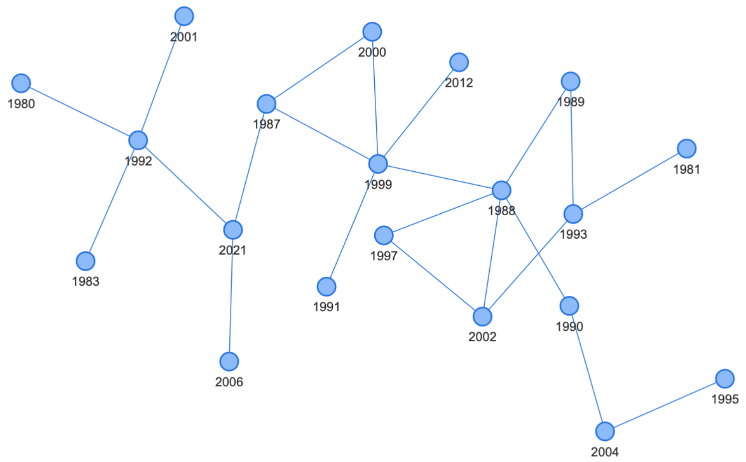

FIGURE 12 – Network graph showing years linked with the same words in their top count

5. Conclusion

The analysis described in this report allows us to answer the simple question that was asked at the beginning : what has evolved in music since the 60's ? If you ask around you what has changed they probably will be able to give you an answer, but they will not be able to argue. Today, with a data set of 20,000 pieces of music, it is possible to identify very interesting patterns that mark breaks between certain groups of songs. This is what the clustering analysis has shown by identifying *Rap* as a very different genre from *Rock*. But, on the other hand, the music has not evolved that much. Artists still talk about the same thing : their loves, and in a very simple vocabulary. *Love, like, dont, im, know, want, cant, kiss* are the words they use to express their feelings since 70 years... This leaves opportunities, if you wish to compose a hit song, just tell a love story with a danceable melody of about 3min40 and a 122 bpm tempo.