Spotify top songs teardown

Amit Nautiyal, Olga Zharikova, Aurelia Kusumastuti, Frederic Neitzel

June 2020

Abstract

Playlists are becoming one of the most renowned ways to listen to music and understanding it can be a intricate part. The following work presents a classification model to identify successful songs by their distinct features based on top chart playlists from 2017 to 2019 from Spotify. In our music survey we perform in depth EDA, feature correlations, detailed distribution analysis and PCA to hinge on popularity score. We perform numerous algorithms to substantiate our hypothesis like xgb and tSNE methodology and then we assent it with spotify's 1921-2020 dataset's popularity matrix. Our assessment reveals the main components to engender a successful top-song. We consummate by affirming how features like "energy", "liveness", "tempo", "valence" and "danceability" can be a benchmark to arbitrate a top song and also be able to collocate top and not the top-songs.
Keywords: PCA, XGB, tSNE

1 Introduction

The music industry makes billions of dollars in sales every year[1] and has reformed its market model commercially several times in recent decades. Digitalisation and technical improvements enabled mass production to satisfy the demand for individual consumption. In order to satisfy individual needs it is essential for companies and artists to understand the listening habits and behavior of their consumers [1].

Record charts have been a stable measure to reflect consumers conduct. Hence one of the big business challenges is to identify characteristics of top songs, recognize changing trends and create predictions upon it. Interiano *et al.*[3] investigated trends in songs popularity by analysing more than 500.000 songs released between 1985 and 2015. He discovered a "[..] downward trend in 'happiness' and 'brightness', [..] " and an "[..] upward trend in 'sadness'.". Especially "[..] successful songs exhibit their own distinct dynamics [..]" being "[..] 'happier', more 'party-like', less 'relaxed' [..]". The result shows that a songs characteristics also becomes the object of uncovering psychological relationships, their perceptual aspects and computational models of music cognition rather than a purely economic subject[2].

Like the aforementioned study our research objective is to determine the correlations of a successful song with its musical features. Therefore we focus on 250 songs of chart lists from the years 2017 to 2019 using data based from the Spotify[2] streaming platform.

[1] https://www.forbes.com/sites/hughmcintyre/2019/04/02/the-global-music-industry-hits-19-billion-in-sales-in-2018-jumping-by-almost-10/a28c5a318a94

[2] `https://support.spotify.com/us/using_spotify/getting_started/what-is-spotify/`

2 Method

2.1 Dataset description

We are working with datasets which were collected from the Spotify Web API[3] The data is provided from Kaggle[4] while we also downloaded datasets with the Spotipy[5] library (We provide this data-mining script):

- Top Tracks of 2017 - 100 tracks [5]
- Top Tracks of 2018 - 100 tracks [6]
- Top Tracks of 2019 - 50 tracks
- Dataset 1921-2020 - 160k+ tracks [8]

We choose the Spotify API due to its possibility to get audio features for a track[6], which are standardized feature categories (highlighted in green in Table 1) to get meta information about a track.

Table column	Description	Value
id	Spotify ID of the track	String
name	track name	String
artists	artist names of the track	String
popularity	number of plays and how recent they are	0 - 100
duration_ms	duration of the track	ms
tempo	overall estimated tempo of a track	bpm
energy	represents a perceptual measure of intensity and activity	0.0 - 1.0
danceability	describes how suitable a track is for dancing	0.0 - 1.0
loudness	overall/average loudness of a track	0.0 - 1.0
liveness	detects the presence of an audience in the recording	0.0 - 1.0
acousticness	a confidence measure wether the track is acoustic	0.0 - 1.0
speechiness	detects the presence of spoken words in the track	0.0 - 1.0
instrumentalness	predicts whether a track contains no vocals	0.0 - 1.0
key	estimated overall key of the track	-1 - 8
valence	describing the musical positiveness conveyed by a track	0.0 - 1.0
mode	indicates the modality (major or minor) of a track	0/1
time signature	estimates how many beats in each bar	beat per bar

Table 1: SHORT DESCRIPTION OF DATASET STRUCTURE AND MUSIC FEATURES

2.2 Data Preparation

To start in-depth data modeling and evaluation we need to perform initial calculations to maintain a data symmetry across the data sets (Table 1). This means we start with operations like checking for null values in the data set and non-float to float data type conversion. Then we perform sequential data analysis across columns as described below in 2.2.1, which establishs the final foundation for our EDA.

[3] https://developer.spotify.com/documentation/web-api/reference/tracks/

[4] https://www.kaggle.com/

[5] https://github.com/plamere/spotipy

[6] https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/

2.2.1 Sequential Columns Analysis

Individual evaluation across columns is essential to understand our features and to make necessary assumptions before proceeding with EDA.

So we describe each feature to understand their significant properties in sense of 'mean', 'standard deviation' and 'distributions'. To proceed we sort the songs according to our initially assumed important features (Table 1). On the basis of beats-per-minute we subdivided the feature 'tempo' into multiple categories (Table 2). This abstraction gives us easier insight and overview of the music tempo preferences.

Tempo	Range
very slow	$(0 \leq tempo \leq 65)$
slow	$(66 \leq tempo \leq 76)$
moderate	$(77 \leq tempo \leq 108)$
fast	$(109 \leq tempo \leq 168)$
very fast	$(tempo > 168)$

Table 2 TEMPO CATEGORIES

To get a more comprehensive result we analyse the songs not only on the basis of important feature, instead we also shift our approach to understand the top artists relation to the important features and then note down the observations for EDA . This means we take a look on the features of artists with notable frequency in the top charts separate for each year to be able to interprete their importance for the overall weight in the feature set.

2.3 Exploratory Data Analysis (EDA)

Our approach is to analyze data sets to summarize their main characteristics with the help of visual methods mentioned in 2.3.1 to get an understanding of our data and its quality. We examine each dataset individually to find potential patterns in the characteristics of the songs. Not only we analyse the overall chart list, but also the detailed songs. Afterwards we compare and summarize the datasets as well as discovering trends between different years.

2.3.1 EDA Methods

In our initial data analysis we use **correlation matrices** (Fig. 1) to understand the basic relationships between the song features for their respective year.

For strongly correlated characteristics we plot regression based **join-plots** (Fig. 2 and 3). Here we can observe the correlation e.g. between "enrgy"-"loudness" and "tempo"-"danceability" respectively. Therefore, join-plots give us a very close correlation matrix, highlighting the relationship between combinations of two features (chosen from correlation matrix).

Our observations from **join-plots** are the basis for getting the top features which we plot with **distplots**. Figure 4 displays a combination of statistical representations of numerical data (i.e. our chosen features from join-plot analysis).

To validate our assumptions including our popularity index (as described in section 2.4) and to see the distribution of important features we perform **Principal Component Analysis (PCA)** (Fig. 5) which we visualize with 3D modelling for detecting the density for our hypothetical top features and to recognize potential outliers.

Further, as we previously mentioned for 1920 - 2019 dataset, for groups of artists we look on the overall dataset over all years while groups are created depending on their contribution of song occurences in the top charts. Successful groups with same

amount of songs are compared with less successful groups in their distribution to find distinctive features (Ref. section 3.2 and 3.3).

2.3.2 Creating a popularity column to calculate scores

Since we plot the distribution of the characteristics among the songs and note the most important features, we eventually create our own binary popularity column based on the observed distribution of the data in our assumed top features. With the computed feature (**popularity index**) we subcategories the song as top or not top song. Then we create correlation graphs between the features and the popularity index.

2.3.3 Validating model

We use XGBoosting as standard feature ranking algorithm for feature engineering to validate our hypthesis about the feature importance.

For the overall results we create a RandomForestClassifier machine learning model to just validate our accuracy score, taking the popularity as a function variable.

Afterwards, we perform tSNE for each dataset, which means we project top and not-top songs into the tSNE space and see how they cluster. On the bases of our popularity index we plot two tSNE clusters for popularity 1 (top songs) and for popularity 0 (not-top songs).

As well we perform the same tSNE on 2000 data points from 2017-2019 on the "Spotify 1921-2020" dataset as described before.

3 Results

3.1 Analysis of feature characteristics in 2017-2019

We plot correlation matrices for the characteristics and we found that loudness and energy of a song are very strongly correlated to each other. Other strongly correlated features are 'loudness and valence', 'loudeness and speechness', 'danceability and valence' and in last 'valence and loudness'.

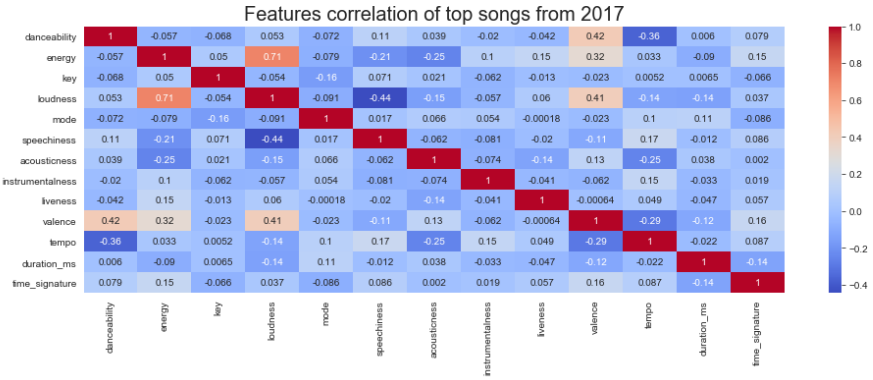


Figure 1: CORRELATION MATRIX FOR 2017

From the correlation matrix (Fig.1), joinplots are created from four of the most correlated pair: namely 'loudness and energy', 'tempo and danceability', 'acousticness and energy' and 'valence and energy'.

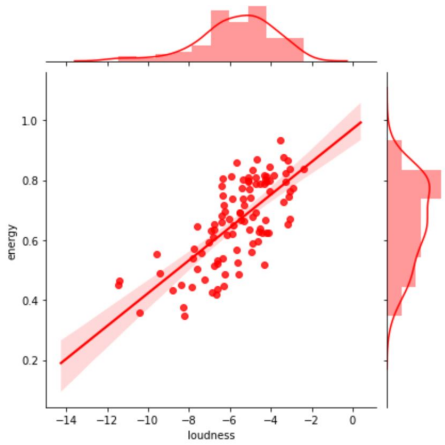


Figure 2: ENERGY-LOUDNESS DISTRIBUTION

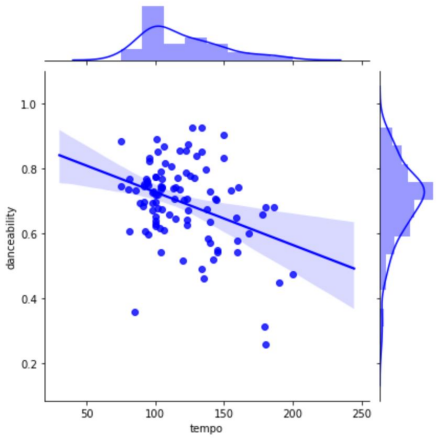


Figure 3: DANCEABILITY-TEMPO DISTRIBUTION

The distribution between loudness and danceability (Fig. 2) we can see the clustered representation where points go along the regression line. Whereas by the distribution between danceability and tempo points (Fig. 3) are spread centrally. So that more songs are by a tempo between 70 and 180 and by danceability between 0.5 and 0.9.

Using distribution plot, we can comprehend the distribution of the most important characteristics (danceability, energy, loudness, acousticness, valence, and tempo) throughout the songs.

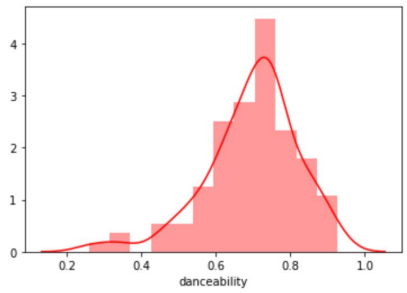


Figure 4: DISTRIBUTION PLOT FOR DANCEABILITY

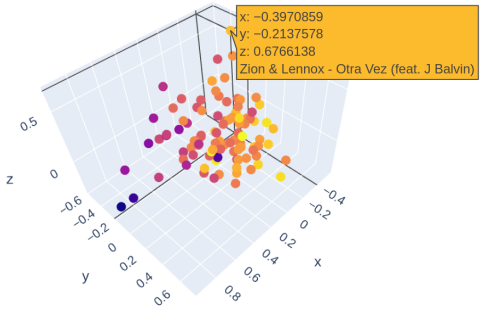


Figure 5: TOP SONGS 2017 PCA DATA VISUALIZED WITH PLOTLY

Based on the above-mentioned characteristics, the distribution of the songs are visualized for the features like loudness, energy, tempo, danceability, acousticness using PCA. We can observe from the figure how densely the data points are distributed, also there are a few outliers to be noted. Why these outliers made it as top songs will be discussed in the *Discussion* part.

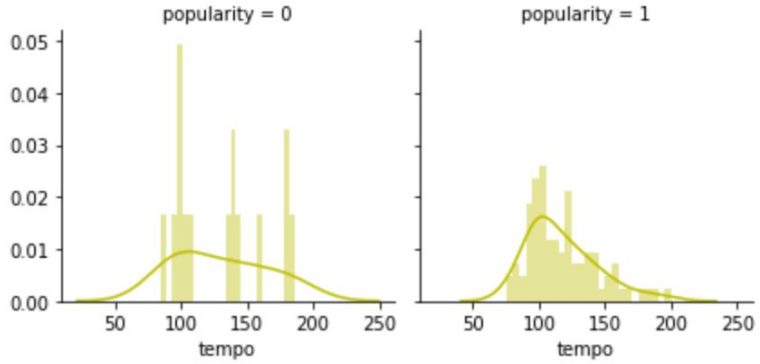


Figure 6: DISTRIBUTION OF TEMPO BETWEEN POPULAR AND NON-POPULAR SONGS

Again, using distribution plot, the important characteristics are compared against the newly created popularity feature. For popularity = 0, the distribution is rather scattered, but for popularity = 1 category we have the data around mean and not scattered as much.

To validate our accuracy, we created a RandomForestClassifier ML model, taking the popularity as a function variable. We compare our ranking with a standard feature ranking algorithm from XGBoost, and we can see that our hypothesis is accurate.

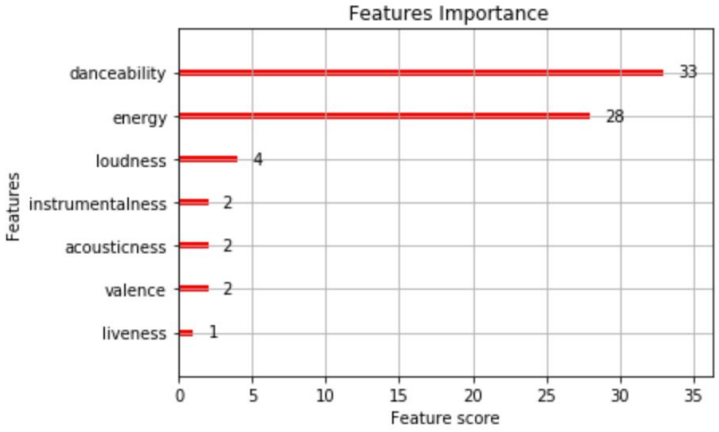


Figure 7: FEATURE IMPORTANCE CALCULATED BY XGBOOST

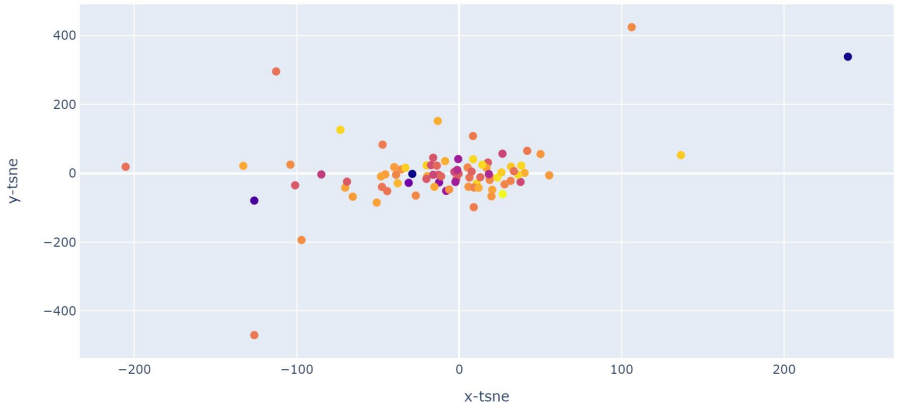


Figure 8: DISTRIBUTION OF POPULAR SONGS ON tSNE SPACE

We then performed tSNE based on the important features. Then we project the 'top' and 'non-top' songs (popularity = 1 and popularity = 0) onto the tSNE space and see how they cluster. The top songs are densely distributed, while the non-top songs are scattered.

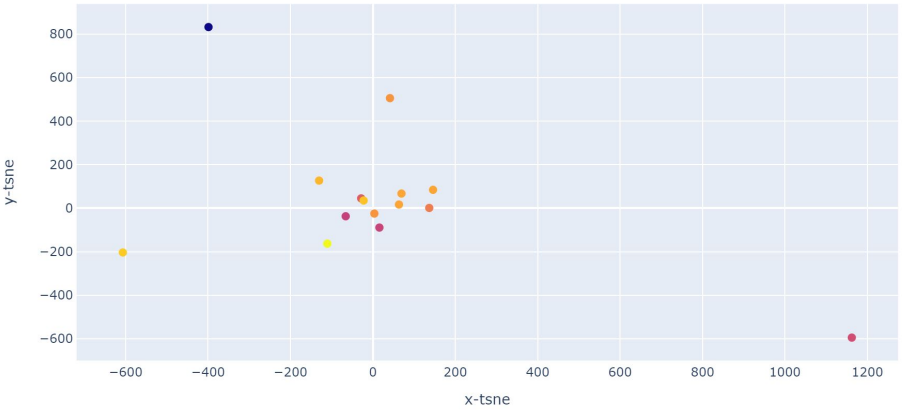


Figure 9: DISTRIBUTION OF NON-POPULAR SONGS ON tSNE SPACE

3.2 Comparison of data from 2017, 2018 and 2019

We compare all datasets with each other to see potential patterns in characteristics of the songs as well as discovering trends between different years based on our strongly correlated features (Fig. 1 and Appendix A) as mentioned in 3.1. These are energy and loudness, tempo and speechiness, and valence and energy.

The distribution between loudness and energy (Fig. 10) as well as by energy and valence (Fig. 11) is almost linear and all datasets show similar results. Most songs are by energy between 0.5 and 0.9 and by loudness between -8 and -3 and have valence values between 0.2 and 0.8.

Results we got from the distribution between speechiness and tempo (Fig. 12) were more valuable.

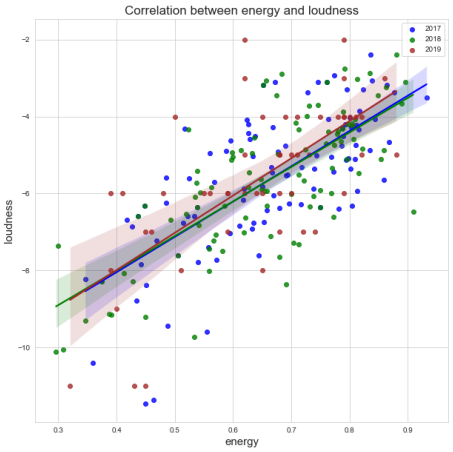


Figure 10: DISTRIBUTION OF ENERGY AND LOUDNESS

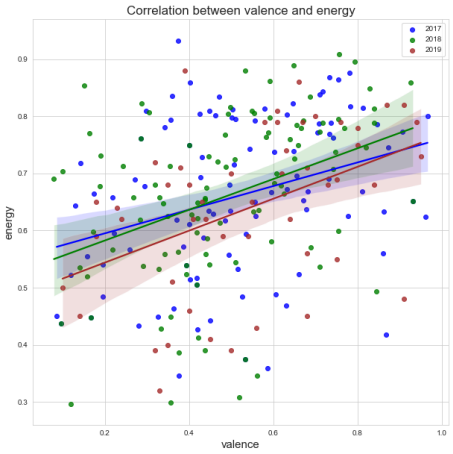


Figure 11: DISTRIBUTION OF ENERGY AND VALENCE

The distribution of songs by a correlation between speechiness and tempo has a typical cluster representation. Most songs are by a tempo between 80 and 130 and by speechiness between 0 and 0.15. Data from years 2017 and 2018 have similar values in both features quite contrary to the data from 2019 where more songs have a distribution with a high number of songs that have speechiness over 0.2 by tempo over 140 bits per minute.

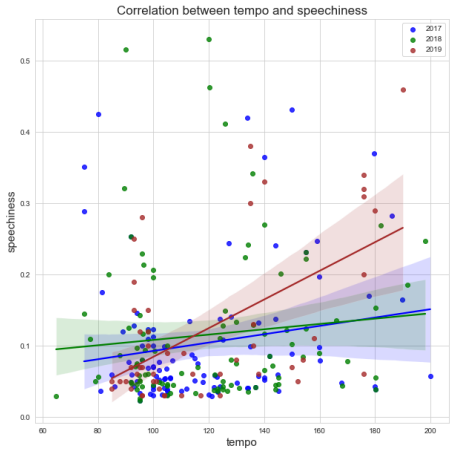


Figure 12: DISTRIBUTION OF SPEECHINESS AND TEMPO

Since the dataset from 2019 contains only 50 songs in comparison to 100 songs from each dataset of 2017 and 2018 our goal is to concentrate not on the number of songs by value but the distribution curve and associations between datasets.

The distribution of danceability (Fig. 13) in all datasets have a Gaussian distribution curve whose distribution converges to a normal distribution as the number of samples increases. This shows us some patterns that most songs have danceability values between 0.5 and 0.9. Similar is also distribution by duration. Most songs from all datasets have a duration between 170000 and 250000 milliseconds which by translation to minutes are between 2.83 and 4.17 minutes.

From the definition of the Spotify API about the liveness feature if the distribution

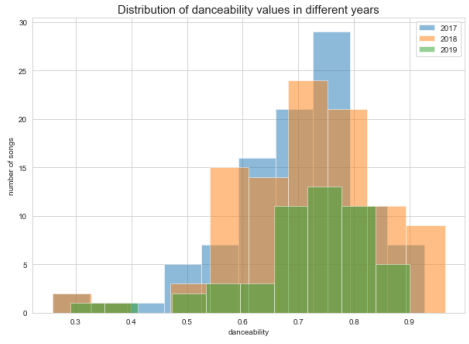


Figure 13: DISTRIBUTION OF DANCEABILITY

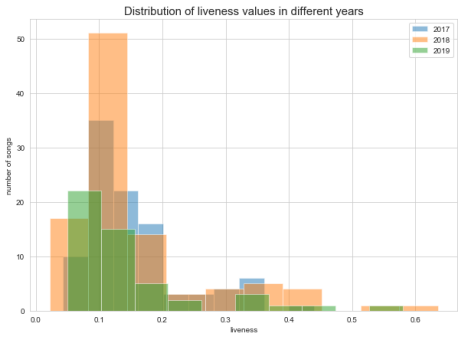


Figure 14: DISTRIBUTION OF LIVENESS

is bigger than 0.8 then it provides a strong likelihood that the track is live. Most of the top song has a low liveness value between 0.05 and 0.2. So it is difficult to recognize if the song was live.(Fig. 14)

3.3 Analysis of artists from datasets 2017-2019

Based on the representation of the Figure 15, we can see multiple loved artists in 2017 till 2019 who had more than a one song in a chart. Representation contains duplicates and artist with more than one song, but for our analysis, we concentrate on the number of songs at all, so we can also see which artists were in all top charts. Specification on songs which contains all datasets we will analyze belove.

Top Artists	
Songs	**Artists**
> 8	Ed Sheeran, Post Malone and Drake
5-6	Ariana Grande, Khalib, Marshmello, The Chainsmokers, and XXXTENTACION
4	The Weeknd, Shawn Mendes, Martin Garrix, Maroon 5, Kendrick Lamar, Imagine Dragons, Calvin Harris and Clean Bandit
3	ZAYN, Sam Smith, Nick Jam, Migos, Maluma, Luis Fonsi, Lauv, J Balvin, Dua Lipa, DJ Snake, DJ Khaled, Camila Cabello, Bruno Mars and Billie Eilish
2	All others singers

Figure 15: ARTISTS WITH MORE THAN ONE SONG OVER ALL DATASETS

From all 11 songs of "Ed Sheeran", only two songs were in top 2017 as well as in top 2018. It means that after all nine songs "Ed Sheeran" were top songs over three different years. One song from "Post Malone" was in top chats of 2017 and 2018 as well. Artist "Drake" had in comparison to "Ed Sheeran" and "Post Malone" each year from 2017 till 2019 new top songs.

At the heat map (Fig. 16) of features from group one, we already can notice some differences. Our earlier analysis showed almost all of the highly correlated features represented here, accept the correlation between speechiness and danceability. These in comparison to our previous heat maps has value by 0.35 which is almost four times higher than the average value where it was lying by ca. 0.08.

The second group contains 20 songs including five different artists which had about six top songs and group three contains 26 different top songs from eight artists who produced about four top songs.

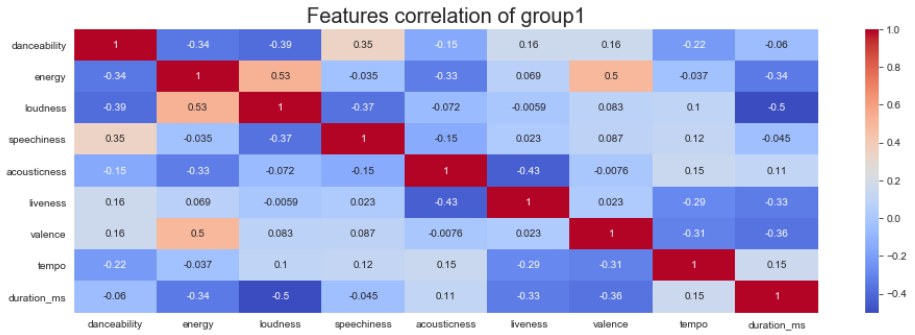


Figure 16: HEAT MAP OF FEATURES BY GROUP1

More songs from group 2 have low speechiness which is similar to the distribution of group 1. From the distribution of loudness (Fig. 17), we can see that most songs distributed around -6. Gaussian curve similar in that distributions shows in the representation of loudness in grop1 as well. Also feature like valence shows more grouped result in comparison to group 1. By distribution of energy, we can notice that songs of group 2 distributed almost equally where es distribution by group 1 showed songs with higher values between 0.45 and 0.65.

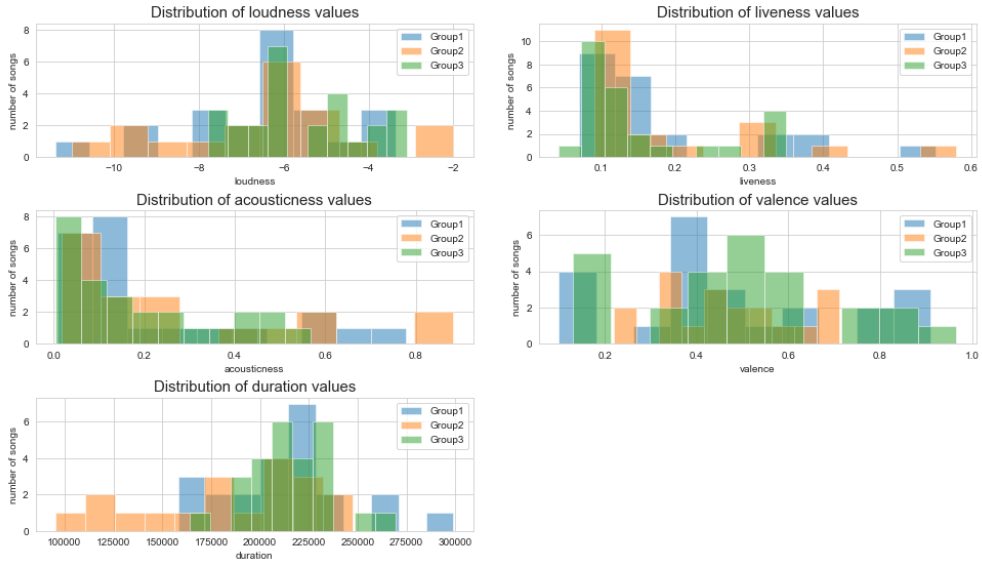


Figure 17: DISTRIBUTIONS BETWEEN GROUP 1,2, AND 3

In contrast to groups 1 and 2, group 3 has higher distribution values by danceability and energy. Impressive to see that values of loudness stayed regular between these first three groups. By the distribution of duration shows similarities to groups 1 and 2 as well.

The fourth group of artists with about three top songs over three different years. The group contains 29 different top songs produced by 14 different artists. This group includes additionally a higher number of songs than by groups one, two, or three.

Comparable to group 3, group 4 has higher distribution values by danceability and energy (Appendix B). Valence values distributed equally and loudness values are higher than by distributions in groups 1, and 2.

Our last group contains 22 top songs from 13 different artists. These artists produced two top songs in the top chart of 2017, 2018, and 2019.

The group has a small variation and points out that the most correlated features are speechiness and tempo. All other characteristics relationships have similarities to our previous analysis on a full dataset. (Appendix C)

Groups three, four, and five show associations again by the distribution of danceability and energy values. Also loudness in a range between -5 and -2 and grouping values by valence between 0.3 and 0.7. The duration of songs is in comparison to all other groups is by group 5 a little lower. (Appendix B, C)

3.4 1921-2020

Now, we have another data set(Spotify 1921-2020). And we perform tSNE on 2000 data points,to see if the validation from 2017-2019 datasets holds upright with this data set. And as you can see in figure 18 popularity 1(top songs) and Figure 19 popularity 0 (not-top songs), it can be clearly distinguished that the trend remains same. The top-songs have a compact clusterring due to its obvious nature whereas in not-top songs we have loosely coupled data points.

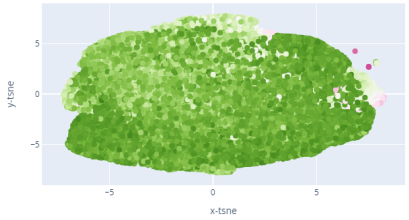


Figure 18: tSNE for popularity=1

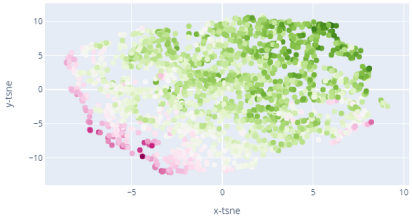


Figure 19: tSNE for popularity=0

4 Discussion

From all our analysis, we can summarize the following main characteristics ranges: The typical duration of top songs lies between 200000 and 235000 ms, which is between 3,33 and 3,92 minutes. Danceability values distributed between 0.55 and 0.8 similar to energy values which lies between 0.45 and 0.7. The top songs have small speechiness values between 0.02 and 0.15 with the loudness values between -7 and -5. Tempo of songs is between 90 and 110 and between 140 and 160 and valence values are between 0.35 and 0.55. All other features like acousticness and liveness were very low so that does not show us appropriate values to make any conclusions. Acousticness values are between 0.02 and 0.2 and by liveness betwenn 0.05 and 0.2.

We concluded that top songs score high on danceability, energy, loudness, and valence but low on acousticness, and have tempo in the range moderate or fast. This tells us that people listen to more upbeat songs than sad ones. These features are also shown to follow a certain trend. The distribution graphs show that more and more songs are in the 'desirable' interval (for example danceability between 0.55 and 0.8).

Musical characteristics are not the only deciding factor for the popularity of a song. In 2017, for example, two of the outlier points, *How Far I'll Go* by Alessia Cara and *Dusk 'til Dawn* by ZAYN, are actually film soundtracks, from the films *Moana*

and *Fifty Shades Freed* respectively. This is also the case for *Bohemian Rhapsody* by Queen in 2019, which is a soundtrack from a Queen biopic under the same name.

The artists themselves are also a contributing factor to a success of a song. As we can see in subsection 3.3, multiple artists show up in the charts in 2018 and 2019 who have been on the chart the years before. These artists are likely to have large dedicated followings, therefore their songs are guaranteed a certain amount of plays.

The music industry is worth a lot of money these days, and record companies are getting very competitive. Music is hardly about raw talent anymore, it's also about how to attract more listeners, thus generating more revenue. Therefore companies look at data from previous months or years to see what people like the most and try to emulate previous trends. This creates a supply-demand cycle, because more listeners are attracted to songs with certain characteristics, and then these characteristics are more and more emulated, and so on and so forth.

This doesn't necessarily mean that 'outsiders' cannot get into the charts. A few of the outliers in the years 2018 and 2019 are from the singer Billie Eilish. The figures suggest that she has a rather different approach in her songs than the mainstream, yet her songs managed to be in the top songs for 2018 and 2019. This has to do with a feature that wasn't factored in in this project: lyrics. Though her songs are very slow, sad, and not danceable, the lyrics speak to her listeners because of how cathartic they are and a lot of people can relate to them.

Possible next step in studying music trends could be using NLP to analyze the meaning of the lyrics. Spotify API also has the Audio Analysis endpoint to analyze rhythm, pitch, and timbre. These two methods can help us have deeper insight on a song's mood. We can also use the amount of followers of playlists on Spotify as an additional determining feature of song or artist popularity. Lastly, we can figure out the exact rankings of the songs (because the data we have is not sorted based on popularity) and possibly use weight in our analysis (more popular songs are weighted more than less popular ones), and comparing Spotify data with data from other sources, for example Billboard.

The way we consume music have always changed and will always be changing. Interiano *et al.* [3] mentioned an upward trend in sadness in 1985-2015, now happiness and loudness is in. Music, and the arts in general, reflects the spirit of the age, thus changes in listening behaviors reflect changes in society.

References

Quotations

[1] Zuboff, Shoshana, "AUGUST 9, 2011: SETTING THE STAGE FOR SURVEILLANCE CAPITALISM" in *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*, 1th ed., 2018, pp. 33-36

[2] Diana Deutsch, "Computational Models of Music Cognition" in *The Psychology of Music*, 2nd ed., 2013, pp. 359-361

[3] Interiano *et al.*. Musical trends and predictability of success in contemporary songs in and out of the top charts. [online] [Accessed on: 13.06.2020]. Available under: `https://royalsocietypublishing.org/doi/10.1098/rsos.171274`

Data source

[4] Spotify API. [online] [Accessed on: 25.05.2020]. Available under: `https://developer.spotify.com/documentation/web-api/reference/tracks/`

[5] Nadin Tamer. Top Spotify Tracks of 2017. [online] [Accessed on: 2.06.2020]. Available under: `https://www.kaggle.com/nadintamer/top-tracks-of-2017`

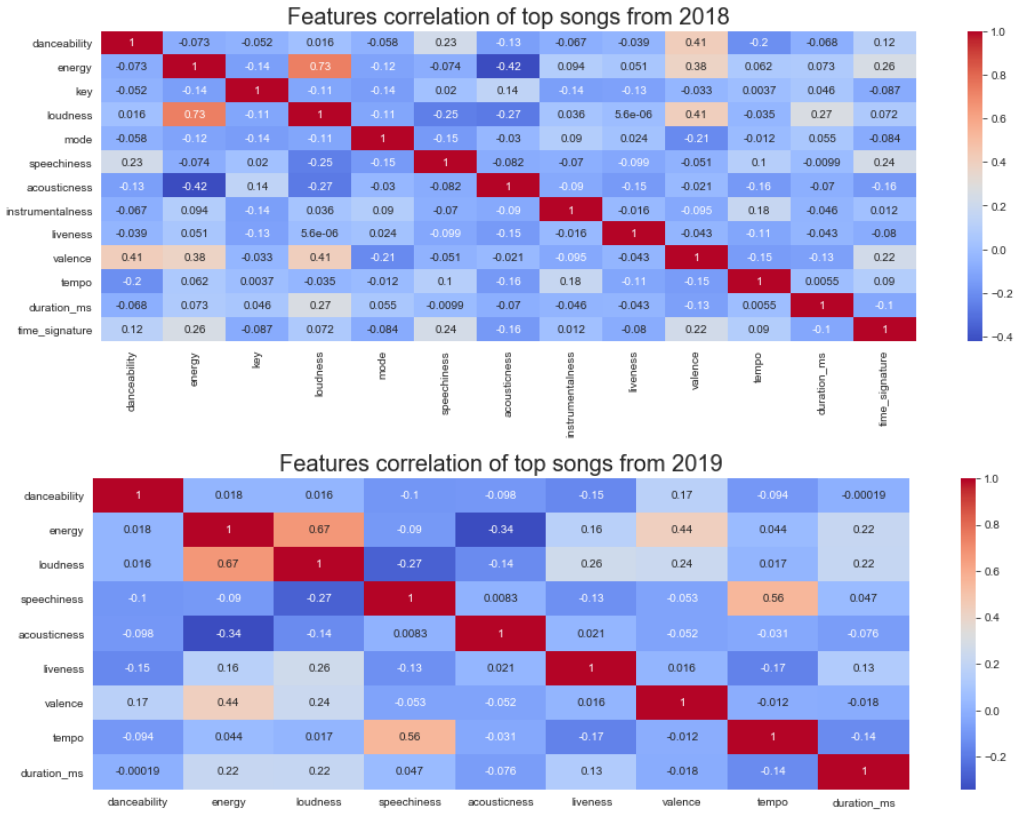
[6] Nadin Tamer. Top Spotify Tracks of 2018. [online] [Accessed on: 2.06.2020]. Available under: `https://www.kaggle.com/nadintamer/top-spotify-tracks-of-2018`

[7] Leonardo Henrique. Top 50 Spotify Songs - 2019. [online] [Accessed on: 2.06.2020]. Available under: `https://www.kaggle.com/leonardopena/top50spotify2019`

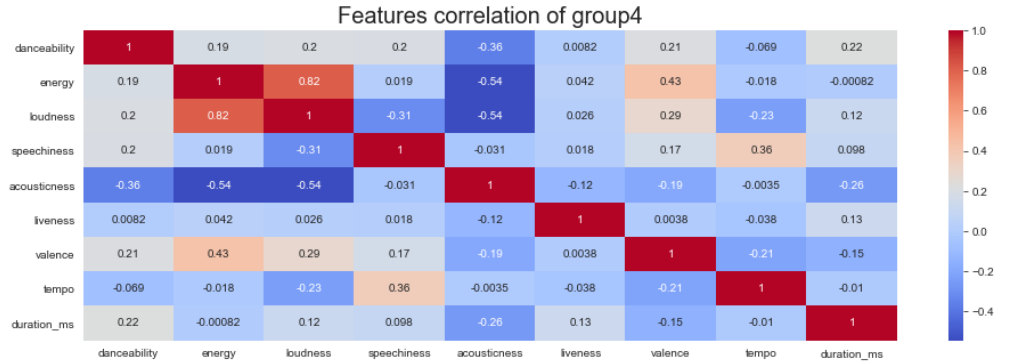
[8] Yamaç Eren Ay. Spotify Dataset 1921-2020, 160k+ Tracks. [online] [Accessed on: 20.06.2020]. Available under: `https://www.kaggle.com/yamaerenay/spotify-dataset-19212020-160k-tracks`

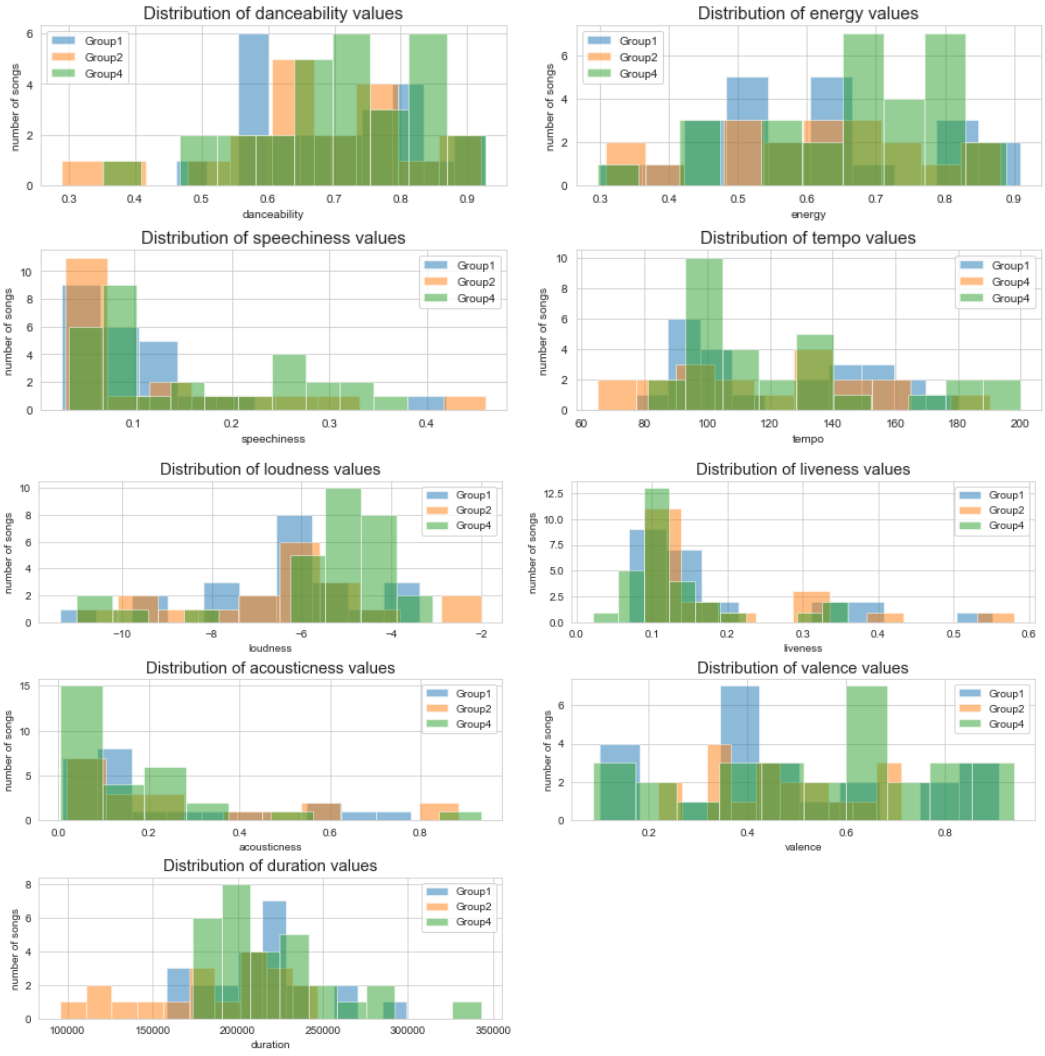
Appendices

A Heatmaps of dataset 2018 and 2019

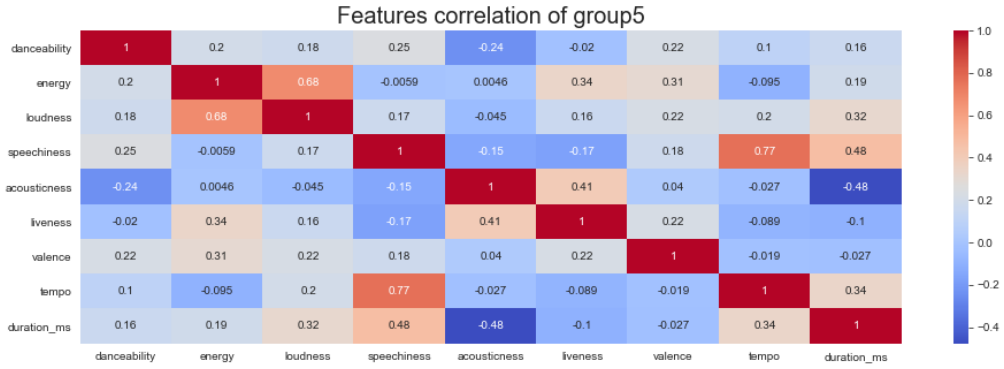


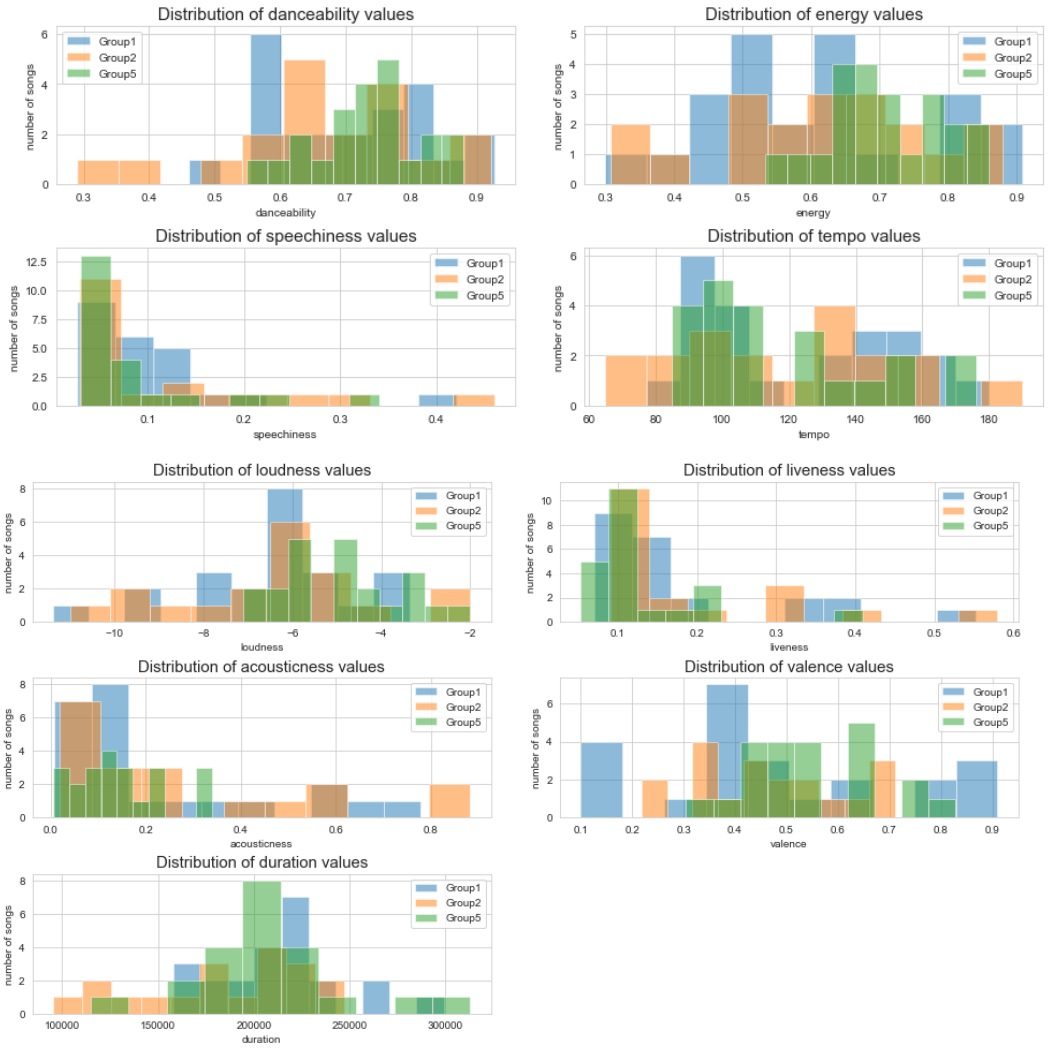
B Group4




Distribution of danceability values
Group1
Group2
Group4
number of songs
danceability
Distribution of energy values
energy
Distribution of speechiness values
speechiness
Distribution of tempo values
tempo
Distribution of loudness values
loudness
Distribution of liveness values
liveness
Distribution of acousticness values
acousticness
Distribution of valence values
valence
Distribution of duration values
duration

C Group5


Features correlation of group5
danceability 1 0.2 0.18 0.25 -0.24 -0.02 0.22 0.1 0.16
energy 0.2 1 0.68 -0.0059 0.0046 0.34 0.31 -0.095 0.19
loudness 0.18 0.68 1 0.17 -0.045 0.16 0.22 0.2 0.32
speechiness 0.25 -0.0059 0.17 1 -0.15 -0.17 0.18 0.77 0.48
acousticness -0.24 0.0046 -0.045 -0.15 1 0.41 0.04 -0.027 -0.48
liveness -0.02 0.34 0.16 -0.17 0.41 1 0.22 -0.089 -0.1
valence 0.22 0.31 0.22 0.18 0.04 0.22 1 -0.019 -0.027
tempo 0.1 -0.095 0.2 0.77 -0.027 -0.089 -0.019 1 0.34
duration_ms 0.16 0.19 0.32 0.48 -0.48 -0.1 -0.027 0.34 1
danceability
energy
loudness
speechiness
acousticness
liveness
valence
tempo
duration_ms


Distribution of danceability values
Group1
Group2
Group5
number of songs
danceability
Distribution of energy values
energy
Distribution of speechiness values
speechiness
Distribution of tempo values
tempo
Distribution of loudness values
loudness
Distribution of liveness values
liveness
Distribution of acousticness values
acousticness
Distribution of valence values
valence
Distribution of duration values
duration