Spotify metadata: Pop music is really less diverse

Dominik Glandorf
Matrikelnummer 6007407
dominik.glandorf@student.uni-tuebingen.de

Felix Gross
Matrikelnummer 6001480
fel.gross@student.uni-tuebingen.de

GitHub repository: https://github.com/dominikglandorf/spotify-analysis

Abstract

In this work we investigated if popular music is less diverse in terms of its variability. To assess the cliché of contemporary popular music "being all the same", metadata of 28,195 tracks on Spotify was collected, notably their popularity and audio features. To compare variability within pop versus non-popular songs, a Principal Component Analysis and F-tests were conducted. A 2D-plot visually showed a difference between the groups. An analysis of variance confirmed that the diversity is indeed less in pop music, strenghtening our initial hypothesis. Conclusively, we discussed our findings' generalizability, challenged by a potential sampling bias.

1 Introduction

Music is one of the oldest and most valued forms of communication and therefore of particular interest. Yet, its underlying structure has been eluded systematic analysis until recently. Accelerated computing and freely accessible databases allow now for systematic inquiry of numerous features across thousands of songs. Thus, it is only now possible to hold common clichés about music proof to the real world data. One of these clichés is, that pop music is all the same and less diverse then non-popular music [Serrà et al., 2012]. Thus, our hypothesis is, that the variability of pop music is smaller than the one observed in non-popular music.

Since Spotify is the world's most used music streaming platform and offers one of the largest corpora of music [SpotifyInc., 2021], it promises to be representative for global listening preferences. For this reason, we rely on this single platform to define what is considered as music, as well as what is considered as popular music. Spotify's massive amount of listening data allows to assess which songs are popular and its meta knowledge eases the examination of music out of a quantitative perspective. We define variability (and synonymously diversity) in the statistical sense of variance, i.e. the average squared distance from the mean. This variability between songs, is expected to be reflected in objective measures like tempo, or more subjective ones like "danceability". The latter ambiguous properties are calculated by Spotify utilising usage statistics and audio analysis.

2 Method

2.1 Sample

Our sample was obtained from the Spotify audio endpoint in the following way: For each possible single Latin letter and two letter combination, the 50 first search results including their popularity score were stored. After eliminating duplicates in the result lists, the total sample encompassed 28,195 songs. To the list of songs were then added the metadata of the Spotify Web API [Spotify,

Feature	Description
danceability	Describes how suitable a track is for dancing from 0.0 to 1.0
energy	Represents a perceptual measure of intensity and activity from 0.0 to 1.0
liveness	Reflects the presence of an audience in the recording from 0.0 to 1.0
loudness	The overall loudness of a track in decibels (dB)
speechiness	Detects the amount of spoken words in a track from 0.0 to 1.0
acousticness	A confidence measure from 0.0 to 1.0 of whether the track is acoustic
instrumentalness	Predicts whether a track contains no vocals from 0.0 to 1.0
tempo	The overall estimated tempo of a track in beats per minute (BPM)
valence	Describes the musical positiveness conveyed by a track from 0.0 to 1.0
popularity	A value from 0 to 100 is assigned to each song, 100 being the most popular

Table 1: Description of the metadata provided by the Spotify

2022]. The obtained features are explained in Table 1. We defined a song being a pop song, if its popularity value is equal or above 90. 70 songs fulfilled this criterion, and were then assigned to the pop song group. The remainder of 28,125 was assigned to the non-pop song group.

2.2 Analysis

To ensure data quality, we first graphically and numerically checked the distributions of our audio features and the popularity in the sample. As stated above, we were interested in the variability of pop songs compared to the remainder. Hence, we first employed an exploratory and visual approach before running statistical tests. In order to search for an underlying structure in the audio features and to be able to get a first impression of the data as a whole, we decided to do a dimensionality reduction via Principal Component Analysis (PCA). We did not use a stochastic neighbour embedding such as t-SNE since we only have nine dimensions and assumed a global structure that might not be preserved by this class of methods. To be able to plot the principal components we chose the two components that explained the most variance. Since the PCA is agnostic to the grouping we indicated the class

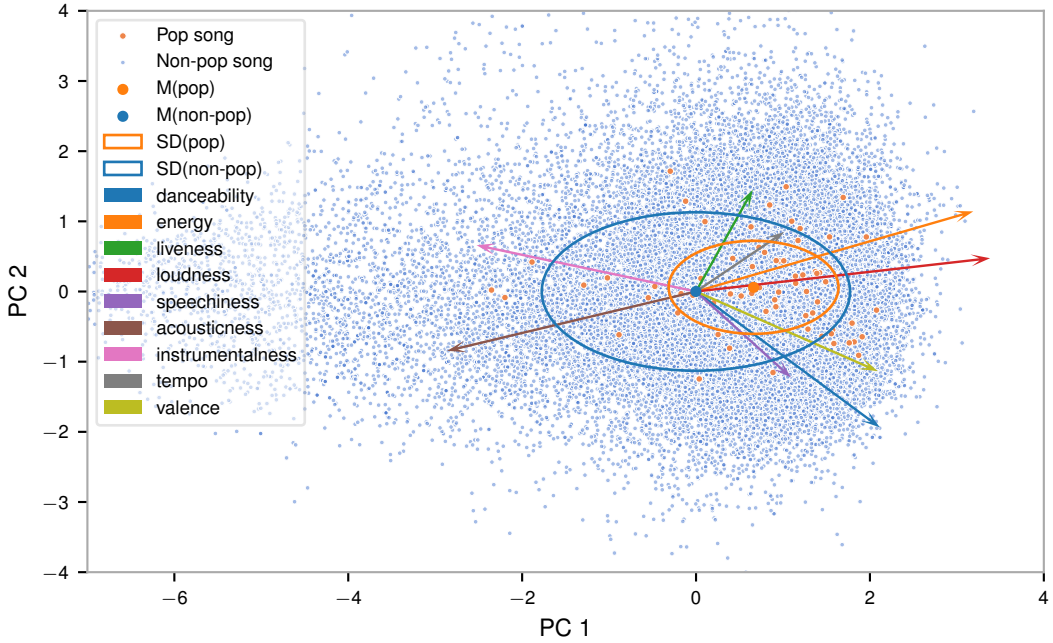

Figure 1: Although the projected data points of the two groups fully overlap in the first two principal components of the PCA, there is a difference in variance, indicated by the ellipses.

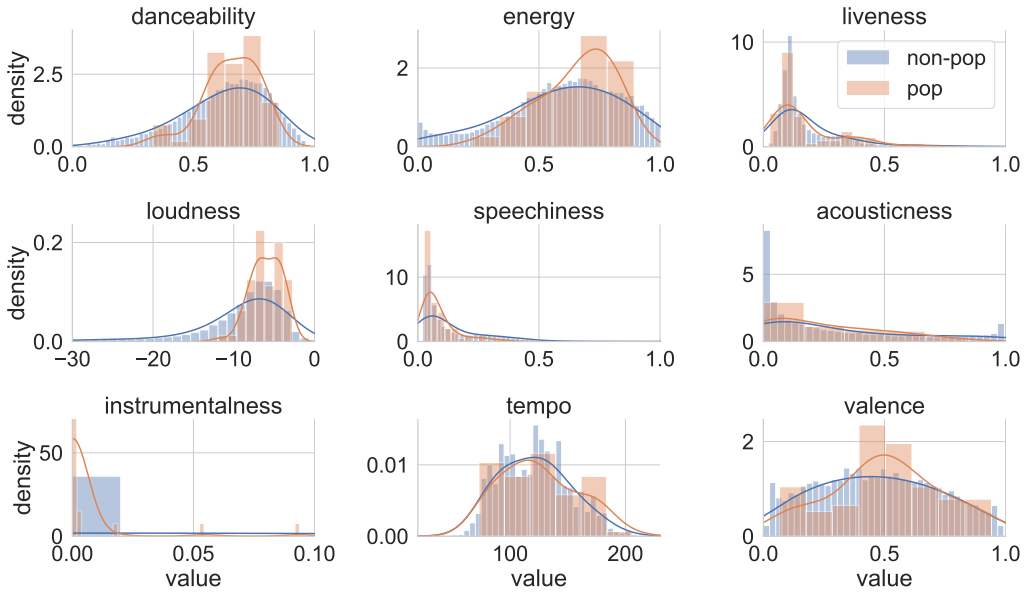

Figure 2: The variances of the single features are indicated by the spread of estimated kernel densities that are based on the histograms.

labels in the visualization, in hope to see a structure. If there are common components in the feature space it is probable that we already see a difference in variance in the visualized data.

Independently of the results of the data exploration, we planned to conduct a two sample variance comparison on each audio feature [Snedecor and Cochran, 1989]. Since our hypothesis was not comprised out of mean differences in the two groups but rather on variance differences, the method of choice was a classical F-test for independent samples. To account for our multiple tests, we corrected our critical f-value using the Bonferroni method. To further assess robustness of our point estimation, we ran a bootstrap sampling from both groups and calculated the variance quotient, i.e. the F-statistic for each of the 10,000 resampled pairs of groups. According to Kruschke [2014], we calculated the 95% highest density interval (HDI) that contains the 95% most often encountered variance ratios obtained in the bootstrap procedure. This allows for corrobating the F-test results, because we get an interval, in which we expect our statistics to lie with a high confidence after resampling the data.

3 Results

In Figure 1 a 2D-visualization of the projected data points is shown. The popular songs only spread over a smaller area within the bandwidth of the remaining songs. The standard variance within the first two principal components (PC) is visualized via the ellipses around the group means. This underlines the visual impression of lower variance of the popular songs. The mean of them is positively shifted mainly on the first PC. To understand how the original features load onto the PCs, their loadings are indicated in an arbitrary scale via the arrows originating from the data's origin. The first PC is mainly determined by loudness and energy (positive loading) as well as instrumentalness and acousticness (negative loading). This PC could be seen as the "pop" dimension since today's popular songs are generally known to be energetic and well-produced (which results in a higher loudness feature) and less instrumental or acoustic (since those are perceived as stripped versions of the original music). The second PC shows the same pattern of less variance in the pop songs, even though it is less prominent. It is mainly based on liveness, speechiness, valence, danceability and tempo. Due to the very small mean difference those features seem not to be different in popular songs, yet less variant. From the dimensionality reduction we can conclude that popular music seems to be actually less diverse in terms of its audio features and shows average differences.

The distributions of the feature values are shown in Figure 2. All features except for liveness, tempo and valence had a significantly lower variance among the pop songs. The highest ratio by far was

Feature	F-value	95% HDI
danceability	2.50**	[1.79; 3.38]
energy	2.26**	[1.79; 2.77]
liveness	1.23	[0.85; 1.80]
loudness	9.27**	[7.53; 11.26]
speechiness	3.57**	[1.95; 6.92]
acousticness	1.80*	[1.44; 2.20]
instrumentalness	642.00**	[111; >9.99×10^5]
tempo	0.83	[0.69; 0.98]
valence	1.21	[1.01; 1.43]

Table 2: Variance statistics and HDIs of bootstrapped samples. The HDI for instrumentalness needed to be clipped due to the extreme F-statistic values. *: $p < \frac{0.05}{9}$, **: $p < \frac{0.001}{9}$

observed in the instrumentalness feature indicating there is literally no variance within the pop songs. The bootstrap-based robustness estimation showed that the significant results are mostly robust to resampling and will result in at least 95% in the same decision. One exception was the feature acousticness, that was not significantly different in the whole HDI.

4 Discussion

Confirming the impression of the exploratory analysis, the variance of six of the evaluated nine features was significantly less in the pop-music group, if compared to the non-pop music group, hereby mostly supporting our hypothesis of pop music being all the same. Even though we could not gather any intelligence about how the feature values were measured on Spotify's side, we still opted to use them. An own synthesis of the raw data into interpretable features would have been beyond the scope of this report. Our specific sampling strategy is very likely to have caught all of the songs, one would consider popular, nevertheless the remainder is sampled inadequately. For each letter combination we only took the fifty most popular songs, thereby omitting the really unpopular songs. Thus, our reference group is rather comprised out of "mediocre songs", then out of songs, which are not popular at all. The size of our pop song group reflects the length of Spotify's own Top Playlist and resembles the present radio repetoire. Furthermore, being popular often encompasses being sparse. So the tremendous sample size differences are inherent to our analysis. We accounted for it by using F-tests, which are quite robust against said differences. Furthermore, bootstrapping does not make any assumption about the distribution at all. Further studies might explore how a different interpretation of popularity would affect the analysis. For example our quite arbitrarily chosen threshold could be lowered, or popularity could be treated as categorically or even continuous variable. Albeit disproportionate growth in non western markets, these are still under-represented and a specific focus on the diversity of the music there, could be fruitful. Lastly a better way of testing our hypothesis was to elicit the data over time - which we had not. The resulting time series would allow for a better robustness, then the methods applied by us. Aiding in better understanding one of our oldest forms of communication - music.

References

Joan Serrà, Álvaro Corral, Marián Boguñá, Martín Haro, and Josep Ll Arcos. Measuring the evolution of contemporary western popular music. *Scientific reports*, 2(1):1–6, 2012.

SpotifyInc. Financial results for fourth quarter, 2021. URL https://investors.spotify.com/home/default.aspx.

Spotify. Spotify API Documentation, 2022. URL https://developer.spotify.com/documentation/web-api/reference/operations/get-audio-features.

George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, eighth edition, 1989.

John Kruschke. *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press, 2014.