

Capstone Project 1 - Milestone Report

Spotify Audio Features

Objective and Goal:

Question: Is there a relationship between a track's audio features and its popularity?

The goal of this project is to use the audio features of music tracks to predict the songs' popularity among listeners. Being able to predict a song or album's popularity before it is released has many applications in music business. As an example, "The Music Fund", a company based in San Francisco, CA, takes advantage of models developed using data to predict popularity of a newly produced album by an artist. If the model predicts high probability of popularity, the company acts as a label and sponsors the musician. This is a purely data-oriented decision without the need for the sponsor company to know the artist or even like their music work.

Predicting popularity of art work productions, such as music or film, also has applications in recommender systems used in digital entertainment. "Popular on Netflix" and "Recommended for you" features are examples of such applications.

Description of the Data:

The dataset used for this project is called "Spotify Audio Features" and is made available via the official Spotify web API. I accessed the dataset from the Kaggle competitions website(<https://www.kaggle.com/tomigelo/spotify-audio-features>) in CSV format. This is a dataset of 116,372 music tracks (one row per track) with 17 variables. The variables are Artist name, Track ID, Track name, Popularity (an integer between 0 and 100 that is calculated based on the number of streams a song had) as well as the following 13 audio features that characterize each song:

Acousticness, danceability, duration_ms, energy, instrumentalness, key, liveness, loudness, mode, speechiness, tempo, time_signature and valence. For more details on the meaning of the audio features visit:

<https://developer.spotify.com/documentation/web-api/reference/tracks/get-audio-features/>

Data Cleaning and Wrangling:

As expected from a kaggle dataset, this data was relatively clean. The following steps were taken to ensure creation of an analyses ready dataset:

1. Tidiness held true:

- Each attribute is measured in a separate variable,
- Each observation (i.e. a track) is in a separate row
- There are no row duplicates in the data

2. Missingness:

There were no missing values in any column.

3. Examining Outliers:

Histograms of all numeric variables were examined. Of the numeric variables, only duration_ms (length of the track in milliseconds) seemed to have extremely large values (e.g. 90 minutes length). A sample of these observations were checked on Spotify's application. The duration values in all the examined tracks were correct. Some examples are:

- Artist name: John, Tracks: Whatever, 5,040,048 ms (~84 mins) and Ever, 5,610,020 ms (~93.5 mins)
- Artist name: Gentle Whispering, Track: Fluffy Sleepy Whispers, 2,097,245 (~35 mins)

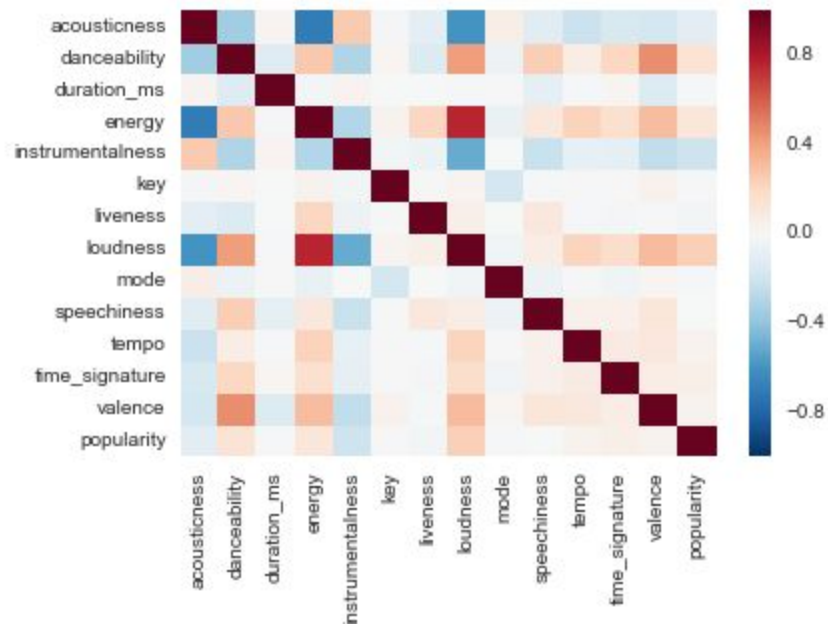
Exploratory Analysis and Initial Findings:

The Main Question: Is there a relationship between audio features of songs and their popularity?

Numerical Variables:

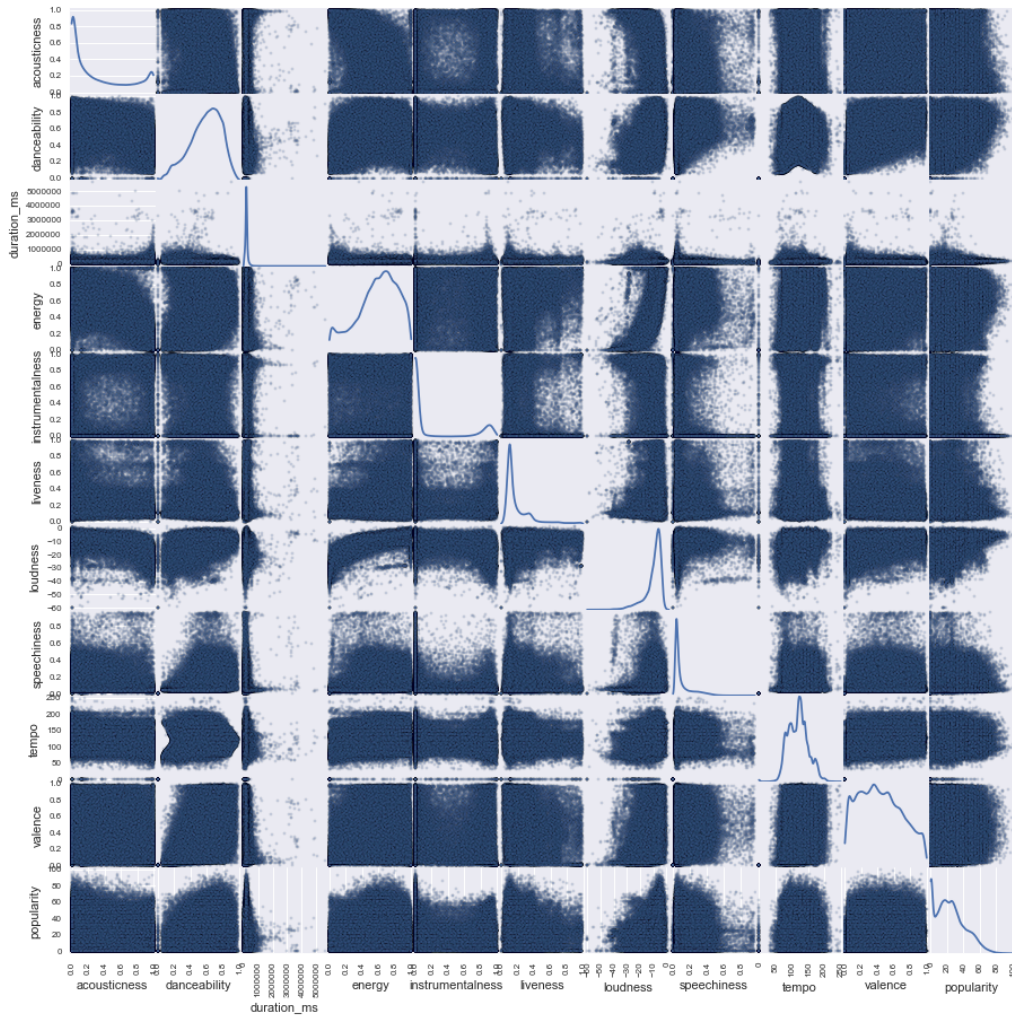
| | acousticness | danceability | duration_ms | energy | instrumentalness | liveness | loudness | speechiness | tempo | valence | popularity |
|-------|--------------|--------------|-------------|----------|------------------|----------|----------|-------------|----------|----------|------------|
| count | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 | 116372.0 |
| mean | 0.3 | 0.6 | 212546.2 | 0.6 | 0.2 | 0.2 | -9.9 | 0.1 | 119.6 | 0.4 | 24.2 |
| std | 0.3 | 0.2 | 124320.8 | 0.3 | 0.4 | 0.2 | 6.5 | 0.1 | 30.2 | 0.3 | 17.9 |
| min | 0.0 | 0.0 | 3203.0 | 0.0 | 0.0 | 0.0 | -60.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 25% | 0.0 | 0.5 | 164049.0 | 0.4 | 0.0 | 0.1 | -11.8 | 0.0 | 96.1 | 0.2 | 10.0 |
| 50% | 0.2 | 0.6 | 201773.0 | 0.6 | 0.0 | 0.1 | -8.0 | 0.1 | 120.0 | 0.4 | 22.0 |
| 75% | 0.6 | 0.7 | 240268.5 | 0.8 | 0.5 | 0.2 | -5.7 | 0.1 | 139.8 | 0.6 | 35.0 |
| max | 1.0 | 1.0 | 5610020.0 | 1.0 | 1.0 | 1.0 | 1.8 | 1.0 | 250.0 | 1.0 | 100.0 |

Correlation plot(matrix):



- Loudness, energy and danceability are positively correlated with popularity
 - Louder and more energetic musics are more commonly used for parties and dancing
- Instrumentalness and acousticness are negatively correlated with popularity
 - Instrumentalness (absence of words in the track) limits the audience of the track to listeners who find the music itself appealing. So high instrumentalness limits the audience.
 - Acousticness is a measure of how much acoustic instruments are used in the track, as opposed to electronically produced music. Over the past decade, specially among younger generation, electronic music has become increasingly popular. This explains in negative correlation between acousticness and popularity.

Scatter Matrix:



- Tracks with very long length (e.g. 80 or 90 minutes) tend to be soft and meditative. They are not everyday music. Hence they have limited audience.
- Live tracks are less popular in spotify. Even though listening to live music as it is being played is entertaining, because of the lower quality of sound, the recorded versions of live performances are not popular compared to studio versions.
- Tracks with extreme tempo levels (< 50 BPM or > 250 BPM) are not popular

Formal Tests of Correlation for Numerical Variables:

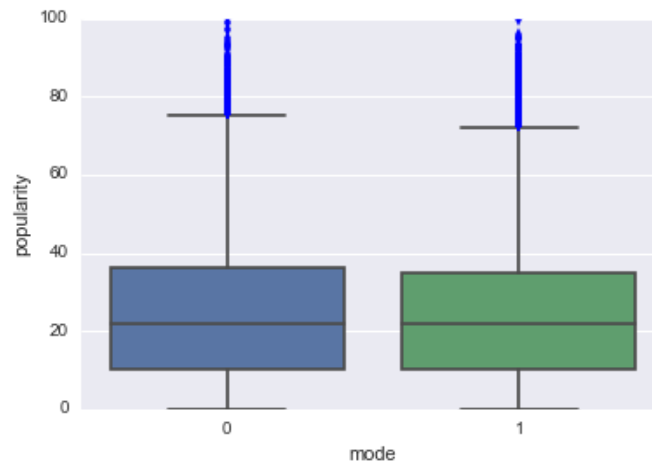
Using the bootstrap method, the following hypothesis tests were performed on the correlation between some of the numerical variables and the outcome (i.e. popularity):

| | Hypothesis to Test | Observed correlation | P_value | Conclusion |
|---------------------------------------|--|----------------------|-----------------|--------------------|
| Loudness vs popularity | H0: No Correlation Ha: Positive correlation | 0.235 | ≈ 0.0 | Rejecting the null |
| Danceability vs popularity | H0: No Correlation Ha: Positive correlation | 0.134 | ≈ 0.0 | Rejecting the null |
| Instrumentalness vs popularity | H0: No Correlation Ha: Negative correlation | -0.21 | ≈ 0.0 | Rejecting the null |
| Track duration vs popularity | H0: No Correlation Ha: Negative correlation | -0.009 | ≈ 0.001 | Rejecting the null |
| Liveness vs popularity | H0: No Correlation Ha: Negative correlation | -0.028 | ≈ 0.001 | Rejecting the null |

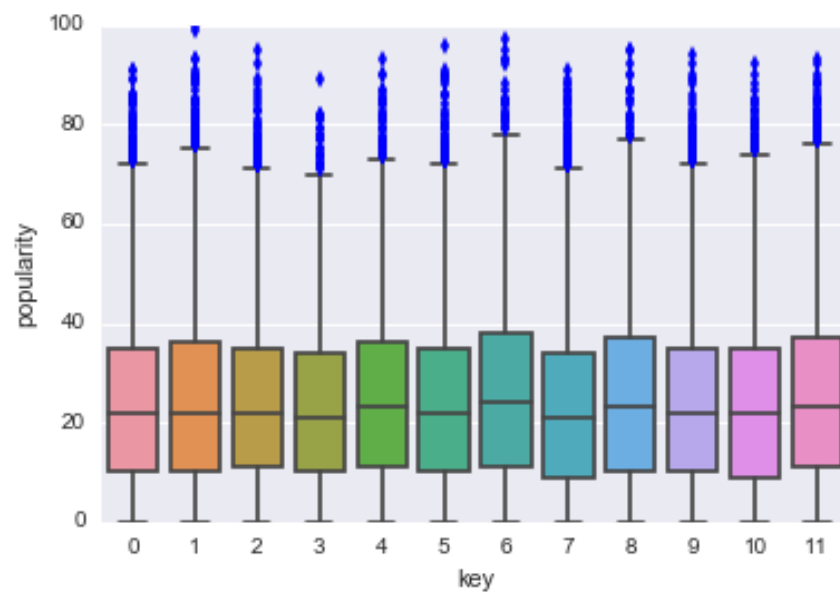
Categorical Variables (Mode, Time Signature, Key):

Variables mode, key and time_signature are represented as integer variables in the dataset. However, since these variables do not have ordinal nature, we will use them as categorical variables:

Mode (minor or major) vs Popularity: No visible difference between the two mode groups

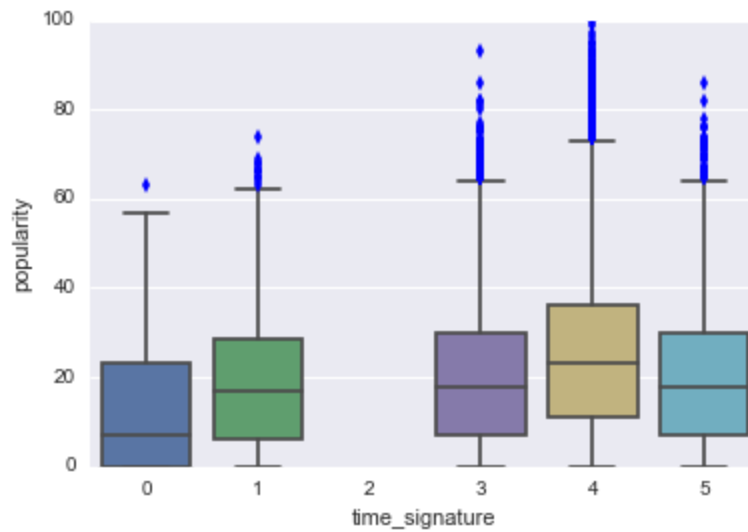


Key vs Popularity: No visible difference between the 12 key groups



Time Signature vs Popularity:

Rhythms with 4 beats per measure seem to be more popular. This could be due to the fact that 4-beat rhythm is more common than any other rhythm. This is the most familiar beat to human ear and is the most frequent time_signature in the dataset (~ 86%)



Formal 2-Sample Test of 4-beat Time Signature vs Other Time Signatures:

Bootstrap method was used to test the following hypothesis at 5% significance level:

H₀: There is no difference in popularity between tracks with 4-beat time signature and other time signatures

H_a: Tracks with 4-beat time signature are, on average, more popular than other time signatures

Observed difference in means ($\mu_{\text{four}} - \mu_{\text{other}}$): **4.56**

P_value \approx **0.0**

Conclusion: Given the small p_value, we reject the null and conclude that 4-beat time signatures, on average, are more popular than other time signatures