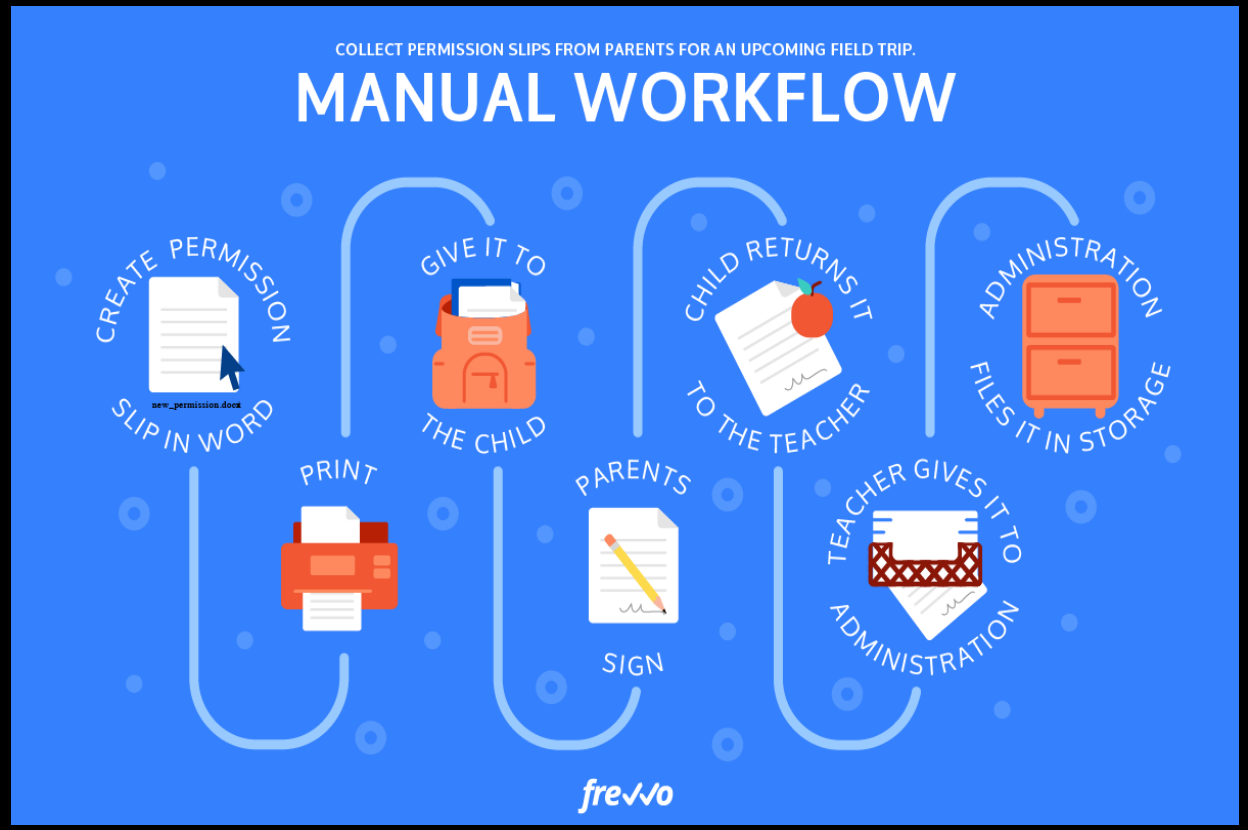
Building Music Recommendation Using Audio Features From Spotify Dataset

Muhammad Ramzy
Job Connector Program - Data Science and Machine Learning
Purwadhika Digital Technology School

Presentation Outline

About Music Streaming Service

Important Features in Music Streaming Service



Workflow

Spotify for Developers

The Dataset

Building Recommendation System

Conclusion and Recommendation

About Music Streaming Service

- MP3 founded in the 90s
- Napster in 1999 provided a straightforward peer-to-peer file-sharing system
- Apple's iTunes Music Store launched in 2003
- Last.fm number-crunched your listening habits, made comparisons with everyone else's data, and served up recommendations, launched in 2002
- The final component of modern streaming: listening to specific tracks or albums on demand, was brought by Soundcloud in 2007

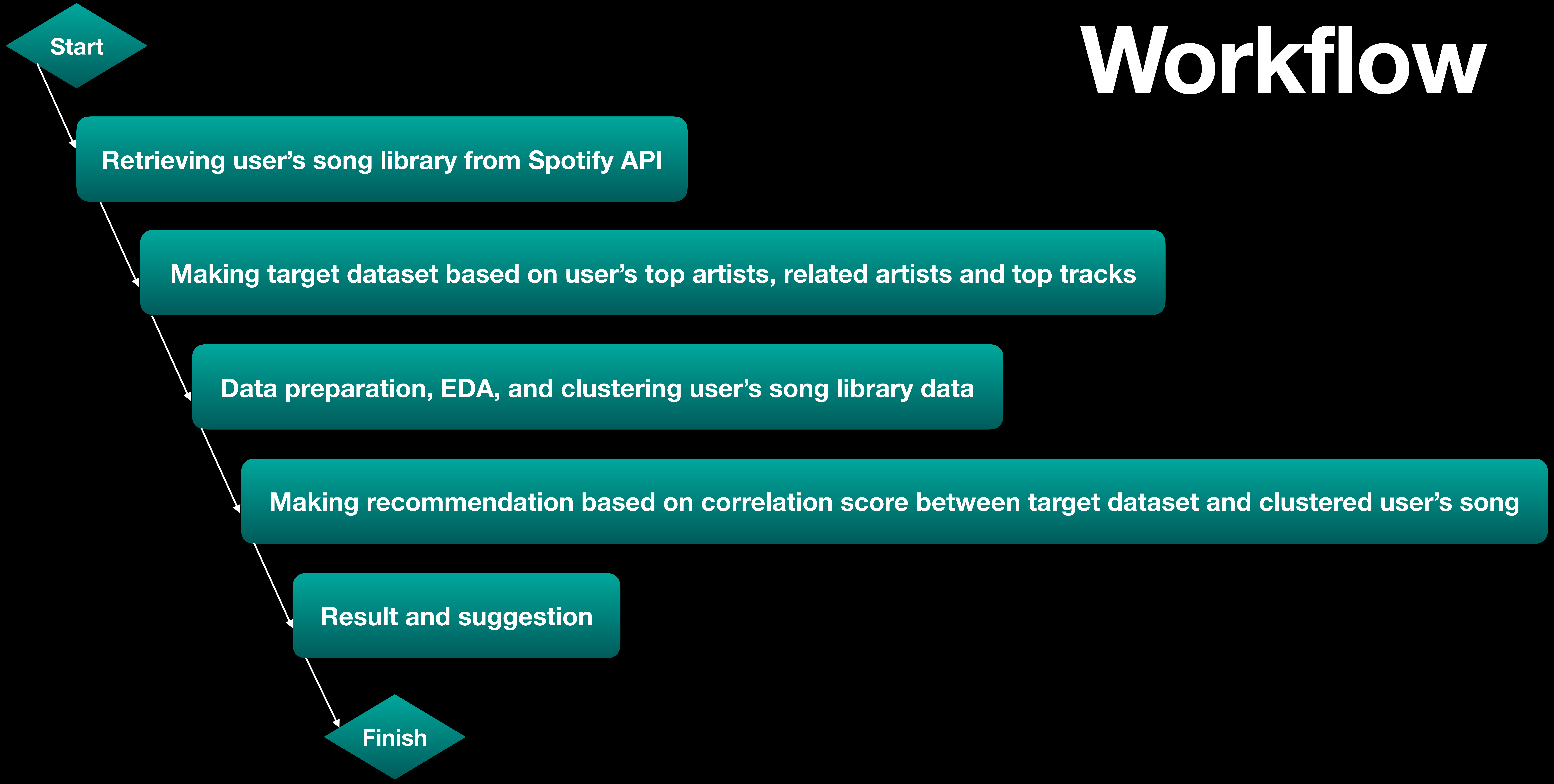
Important Features in Music Streaming Service

Features That
Sells Streaming Service

- Music Catalogue
- Recommendation System
- Hi-Res Quality Music
- Exclusive Content
- Data Analysis

Workflow

Workflow



Start
Retrieving user’s song library from Spotify API
Making target dataset based on user’s top artists, related artists and top tracks
Data preparation, EDA, and clustering user’s song library data
Making recommendation based on correlation score between target dataset and clustered user’s song
Result and suggestion
Finish

The Dataset

Spotify for Developers

DISCOVER DOCS CONSOLE COMMUNITY

WEB API

QUICK START GUIDES LIBRARIES REFERENCE

API ENDPOINT REFERENCE

- Albums
- Artists
- Browse
- Episodes
- Follow
- Library
  - Check User's Saved Albums

Psst! Check out our brand-new Web API Reference in beta!
And be sure to tweet us your feedback at @SpotifyPlatform on Twitter!

Get a User's Saved Tracks

Get a list of the songs saved in the current Spotify user's 'Your Music' library.

WEB API

QUICK START GUIDES LIBRARIES REFERENCE

API ENDPOINT REFERENCE

- Albums
- Artists
- Browse
- Episodes
- Follow
- Library
- Personalization

Psst! Check out our brand-new Web API Reference in beta!
And be sure to tweet us your feedback at @SpotifyPlatform on Twitter!

Get Audio Features for a Track

Get audio feature information for a single track identified by its unique Spotify ID.

```
GET https://api.spotify.com/v1/me/tracks
```

the endpoints that being used for user's library dataset

```
GET https://api.spotify.com/v1/audio-features/{id}
```

The result of merging the track and audio feature

from audio feature

	artist_name	track_name	track_id	popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness
0	Koes Plus	Anak Manja	23xDflyvmkfkKDm6TLTC6c	15	0.797	0.718	10	-5.544	1	0.0286	0.561	0.000027
1	Koes Plus	Cinta Buta	2mA3rND8OfaweQmNzbCC6k	15	0.766	0.580	0	-8.993	1	0.0311	0.425	0.000431
2	Koes Plus	Cinta Mulia	0sls6RZiAhwqUdZfieQaz9	19	0.823	0.569	0	-7.859	1	0.0402	0.176	0.000000
3	Koes Plus	Cintamu Tlah Berlalu	0mwtB9lS4tWCHbgRUnQzbL	20	0.533	0.411	2	-9.421	1	0.0260	0.689	0.000030
4	Koes Plus	Cobaan Hidup	3DnduT3dPwe9t9YIncUWjF	15	0.558	0.506	2	-6.221	1	0.0258	0.304	0.005420

liveness	valence	tempo	uri	duration_ms	time_signature
0.160	0.955	114.094	spotify:track:23xDflyvmkfkKDm6TLTC6c	220298	4
0.221	0.781	139.936	spotify:track:2mA3rND8OfaweQmNzbCC6k	163693	4
0.136	0.917	129.081	spotify:track:0sls6RZiAhwqUdZfieQaz9	172678	4
0.272	0.461	142.849	spotify:track:0mwtB9lS4tWCHbgRUnQzbL	123493	4
0.292	0.481	137.967	spotify:track:3DnduT3dPwe9t9YIncUWjF	185281	4

from user's saved track

Popularity

The popularity of the track. The value will be between 0 and 100, with 100 being the most popular.

Danceability

Danceability describes how suitable a track is for dancing based on a combination of musical elements including tempo, rhythm stability, beat strength, and overall regularity. A value of 0.0 is least danceable and 1.0 is most danceable.

Energy

Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.

Key

The estimated overall key of the track. Integers map to pitches using standard Pitch Class notation . E.g. 0 = C, 1 = C♯/D♭, 2 = D, and so on.

Loudness

The overall loudness of a track in decibels (dB).

Mode

Mode indicates the modality (major or minor) of a track, the type of scale from which its melodic content is derived.

Speechiness

Speechiness detects the presence of spoken words in a track. The more exclusively speech-like the recording (e.g. talk show, audio book, poetry), the closer to 1.0 the attribute value. Values above 0.66 describe tracks that are probably made entirely of spoken words.

Accousticness

A confidence measure from 0.0 to 1.0 of whether the track is acoustic.

Instrumentalness

Predicts whether a track contains no vocals. “Ooh” and “aah” sounds are treated as instrumental in this context. Rap or spoken word tracks are clearly “vocal”. The closer the instrumentalness value is to 1.0,

Liveness

Detects the presence of an audience in the recording. Higher liveness values represent an increased probability that the track was performed live.

Valance

A measure from 0.0 to 1.0 describing 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

The overall estimated tempo of a track in beats per minute (BPM).

Duration_ms

The duration of the track in milliseconds.

Time_signature

An estimated overall time signature of a track. The time signature (meter) is a notational convention to specify how many beats are in each bar (or measure).

Personalization

Endpoints for retrieving information about the user's listening habits.

Base URL: `https://api.spotify.com/v1`

METHOD	ENDPOINT	USAGE	RETURNS
GET	/v1/me/top/{type}	Get a User's Top Artists and Tracks	artists or tracks

Get an Artist's Related Artists

Get Spotify catalog information about artists similar to a given artist. Similarity is based on analysis of the Spotify community's listening history.

Get an Artist's Top Tracks

Get Spotify catalog information about an artist's top tracks by country.

For target dataset

1. Get user's top artists
2. Get related artists based on user's top artists
3. Get related artists' top tracks
4. Complete the dataset with audio features

Building Recommendation

Feature Selection

redundant features

	artist_name	track_name	track_id	popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness
0	Koes Plus	Anak Manja	23xDflyvmkfkKDm6TLTC6c	15	0.797	0.718	10	-5.544	1	0.0286	0.561	0.000027
1	Koes Plus	Cinta Buta	2mA3rND8OfaweQmNzbCC6k	15	0.766	0.580	0	-8.993	1	0.0311	0.425	0.000431
2	Koes Plus	Cinta Mulia	0sls6RZiAhwqUdZfieQaz9	19	0.823	0.569	0	-7.859	1	0.0402	0.176	0.000000
3	Koes Plus	Cintamu Tlah Berlalu	0mwtB9lS4tWCHbgRUnQzbL	20	0.533	0.411	2	-9.421	1	0.0260	0.689	0.000030
4	Koes Plus	Cobaan Hidup	3DnduT3dPwe9t9YIncUWjF	15	0.558	0.506	2	-6.221	1	0.0258	0.304	0.005420

Not a musical feature and there are a lot of nonsensical 0 value

also redundant

unnecessary

liveness	valence	tempo	uri	duration_ms	time_signature
0.160	0.955	114.094	spotify:track:23xDflyvmkfkKDm6TLTC6c	220298	4
0.221	0.781	139.936	spotify:track:2mA3rND8OfaweQmNzbCC6k	163693	4
0.136	0.917	129.081	spotify:track:0sls6RZiAhwqUdZfieQaz9	172678	4
0.272	0.461	142.849	spotify:track:0mwtB9lS4tWCHbgRUnQzbL	123493	4
0.292	0.431	137.967	spotify:track:3DnduT3dPwe9t9YIncUWjF	185281	4

Data Filtering

So, there are 6711 songs in the dataset, which is every song that I have saved ever since I used Spotify, and I think that's a bit too much.

Remember earlier that we got user's top artists? Let's use that to filter our data.
I can gather 158 user's top artists, in no particular order.

Besides that, I also filtering my data by dropping duplicates, because there are chance that there are one song but have different album, therefore, it's possible that my data have a duplicate

After filtering, my data shrink from 6711 rows to 3707 rows.

	artist_name	genres
0	TURBO	['korean pop']
1	Vampire Weekend	['baroque pop', 'indie pop', 'indie rock', 'mo...
2	Yves Tumor	['art pop', 'chamber psych', 'chillwave', 'esc...
3	The Strokes	['alternative rock', 'garage rock', 'modern ro...
4	Car Seat Headrest	['alternative rock', 'anti-folk', 'art pop', '...
...	...	...
153	Project Pop	['indonesian pop']
154	MF DOOM	['alternative hip hop', 'east coast hip hop', ...
155	Daft Punk	['electro', 'filter house']
156	SCALLER	['indonesian indie']
157	Parquet Courts	['alternative rock', 'brooklyn indie', 'denton...

158 rows × 3 columns

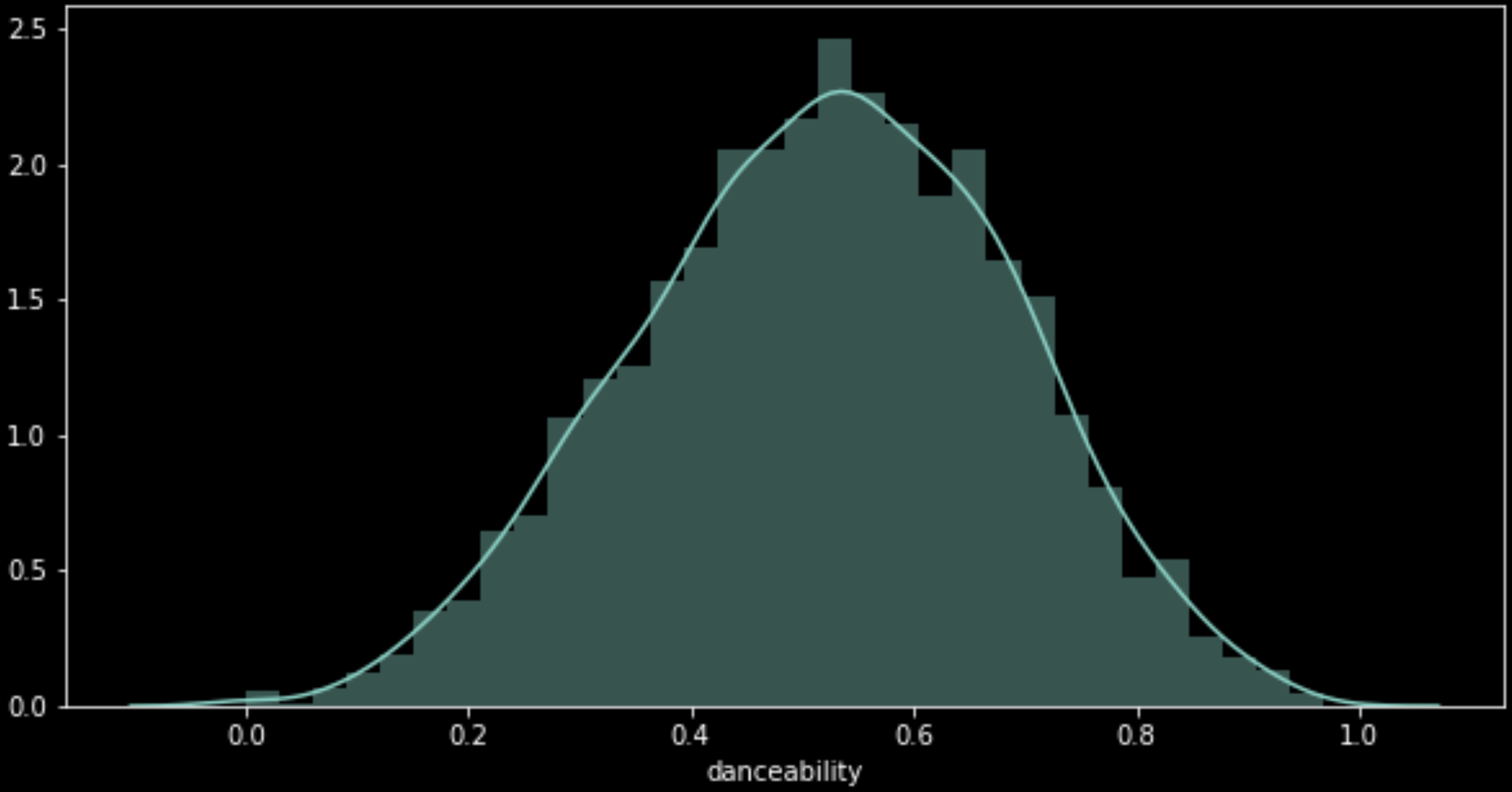
Personalization

Endpoints for retrieving information about the user's listening habits.

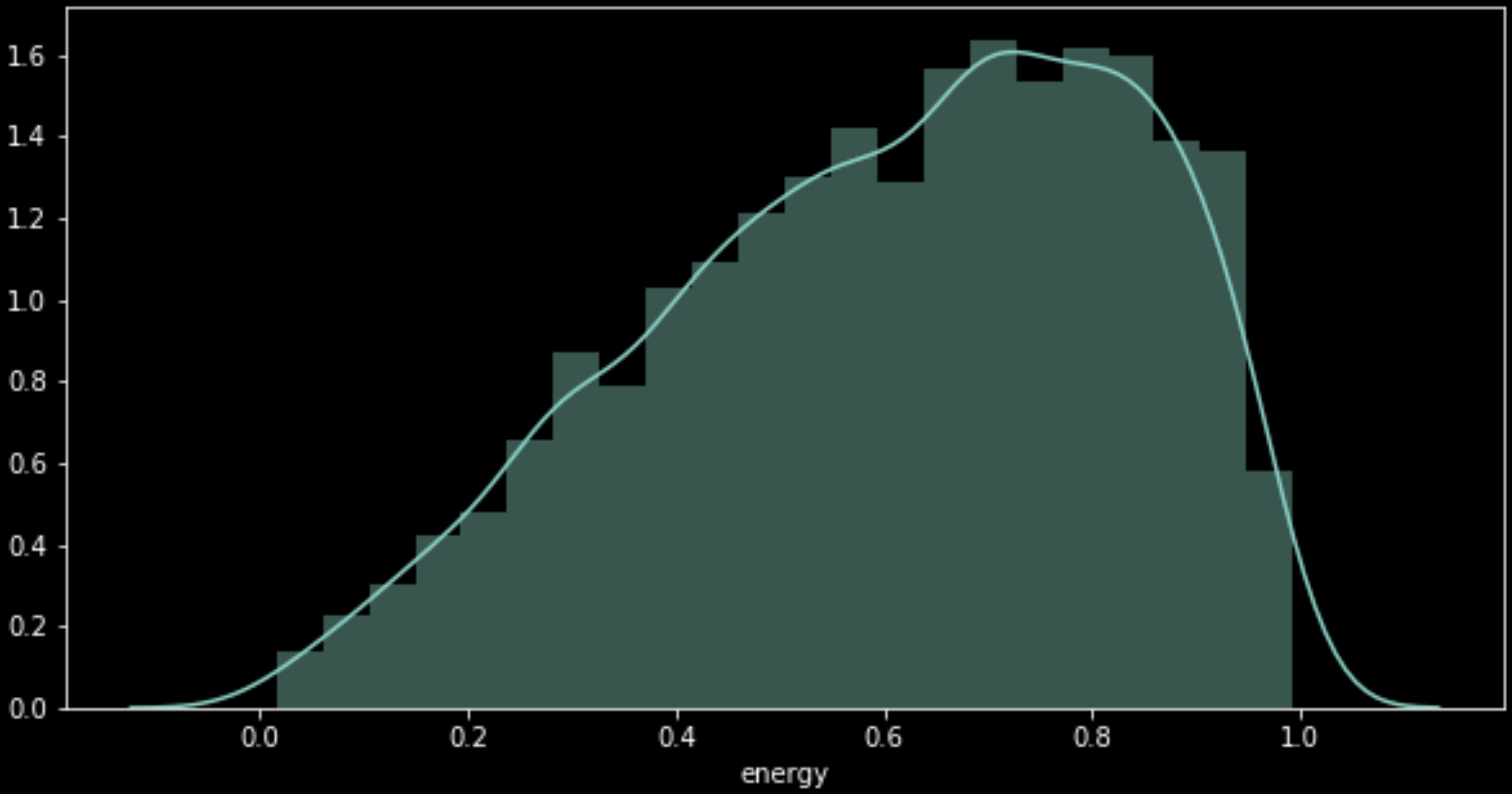
Base URL: `https://api.spotify.com/v1`

METHOD	ENDPOINT	USAGE	RETURNS
GET	/v1/me/top/{type}	Get a User's Top Artists and Tracks	artists or tracks

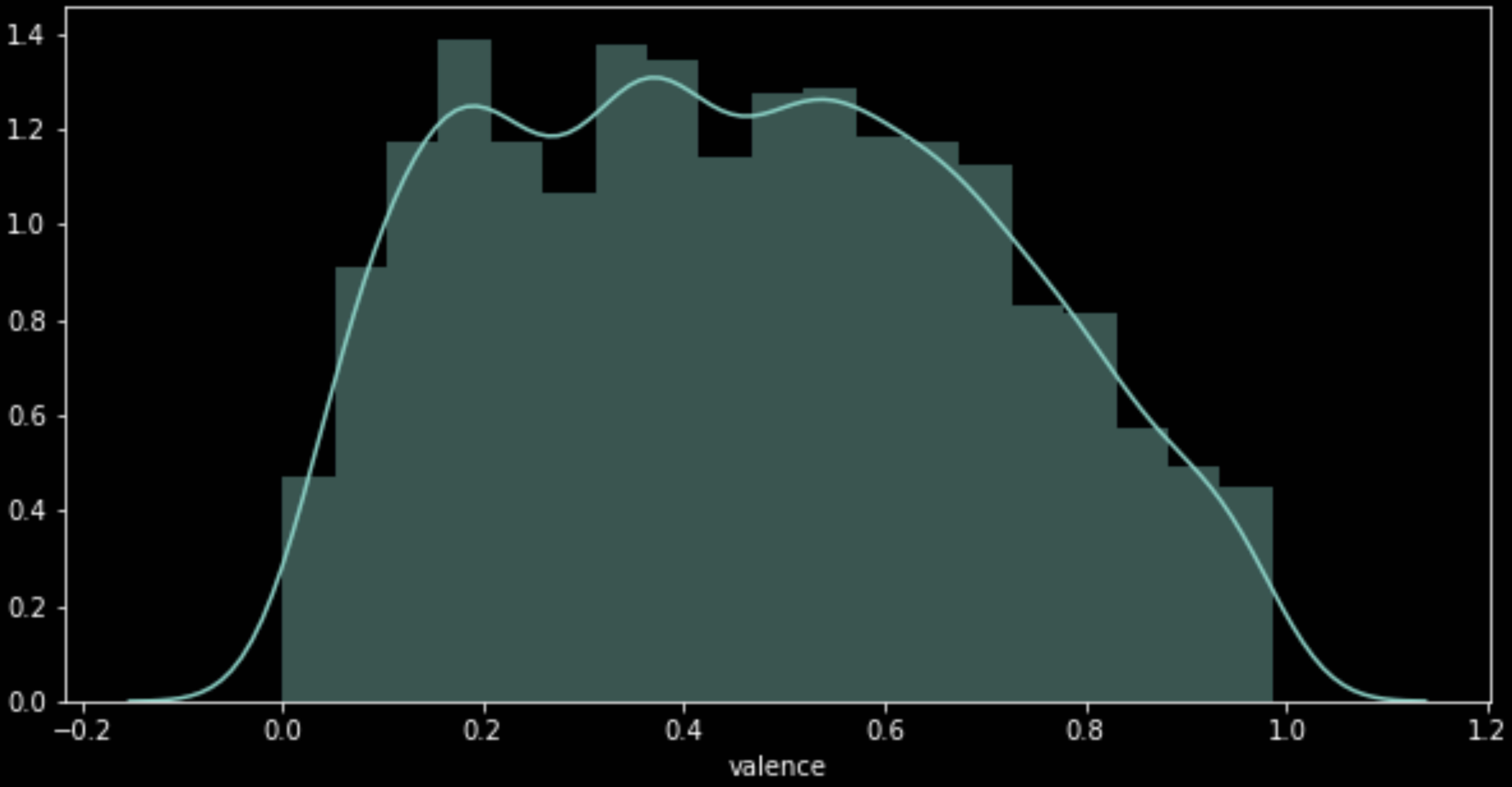
Data Distribution



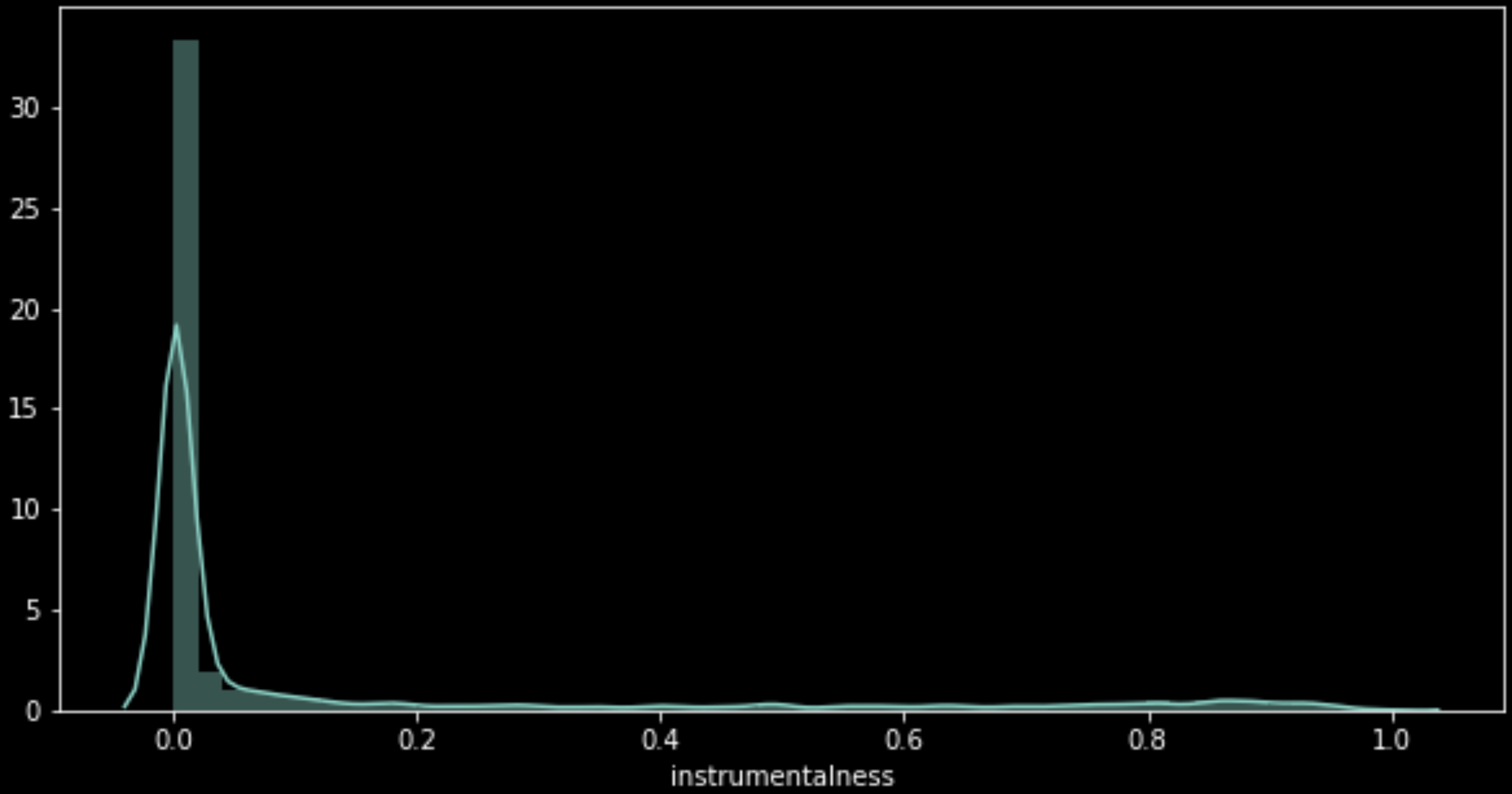
Data Distribution



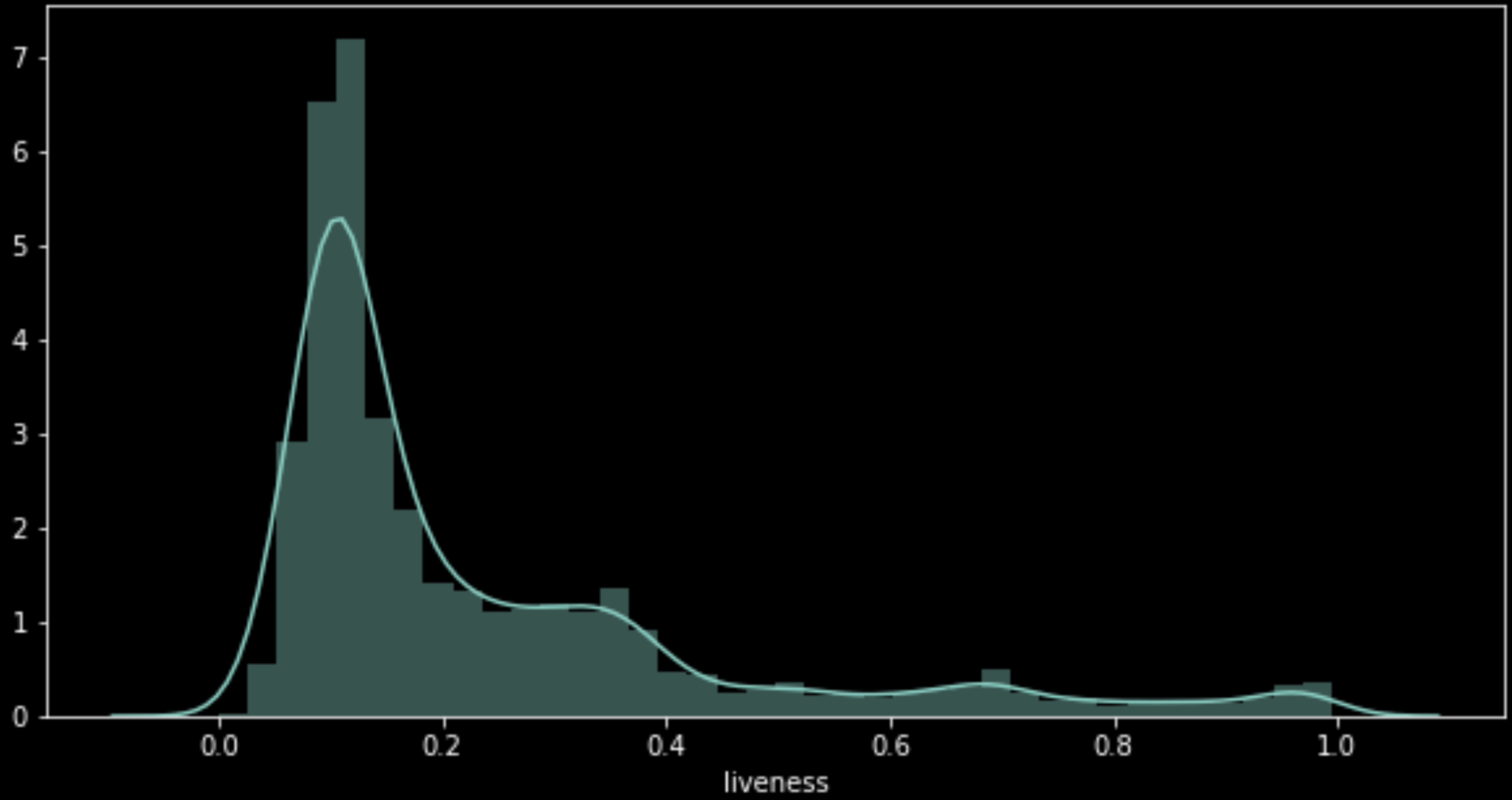
Data Distribution



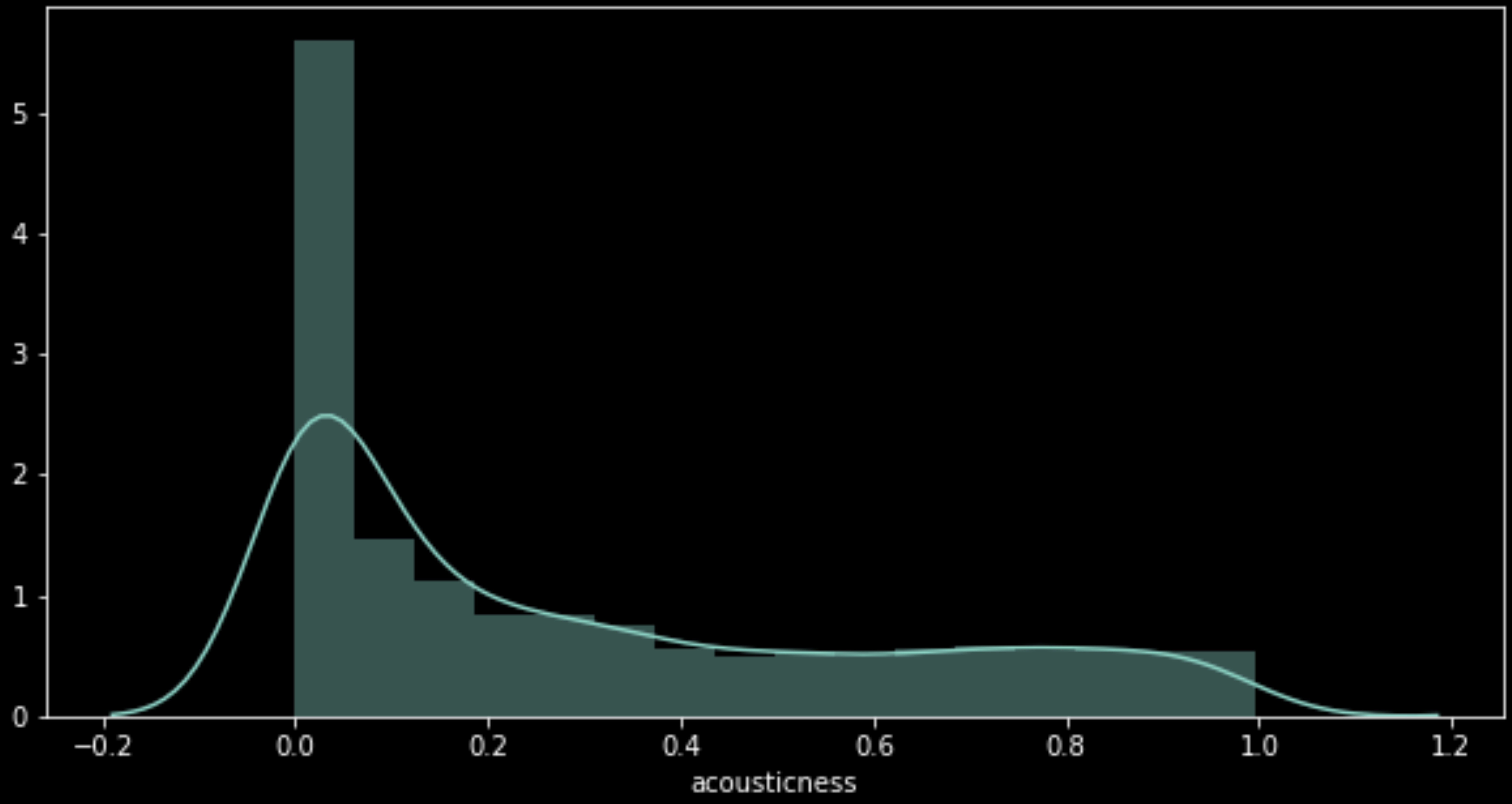
Data Distribution



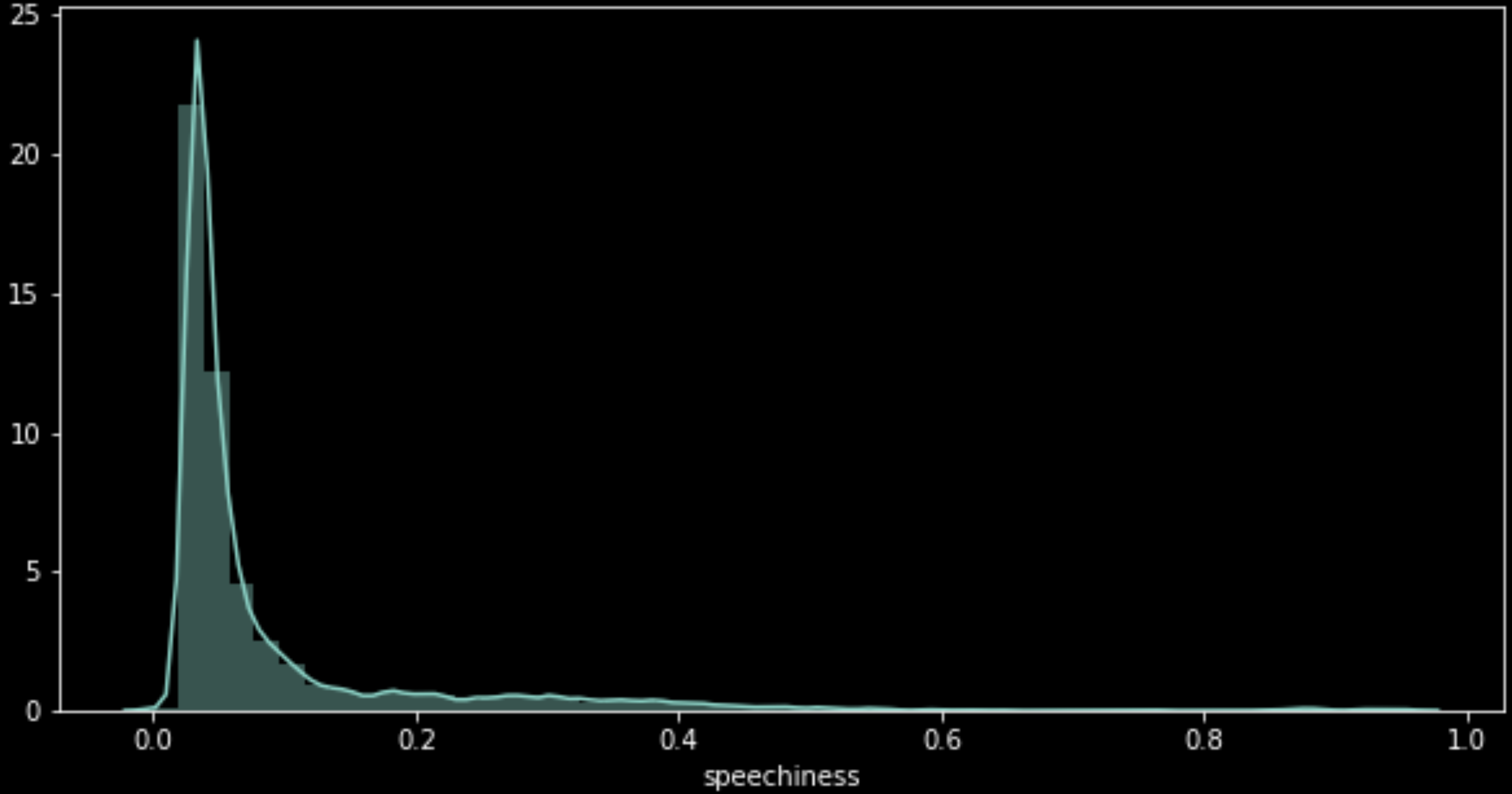
Data Distribution



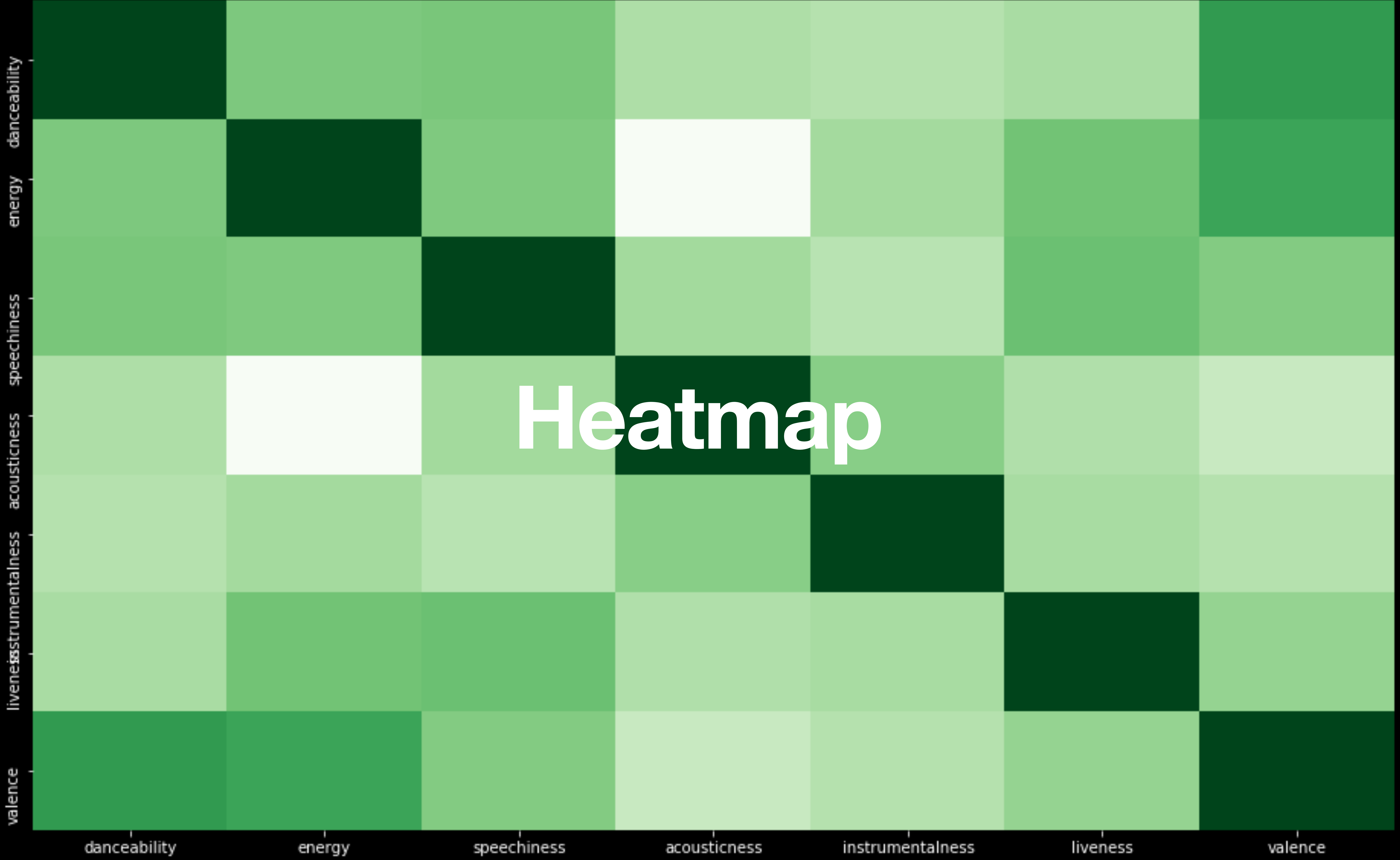
Data Distribution



Data Distribution

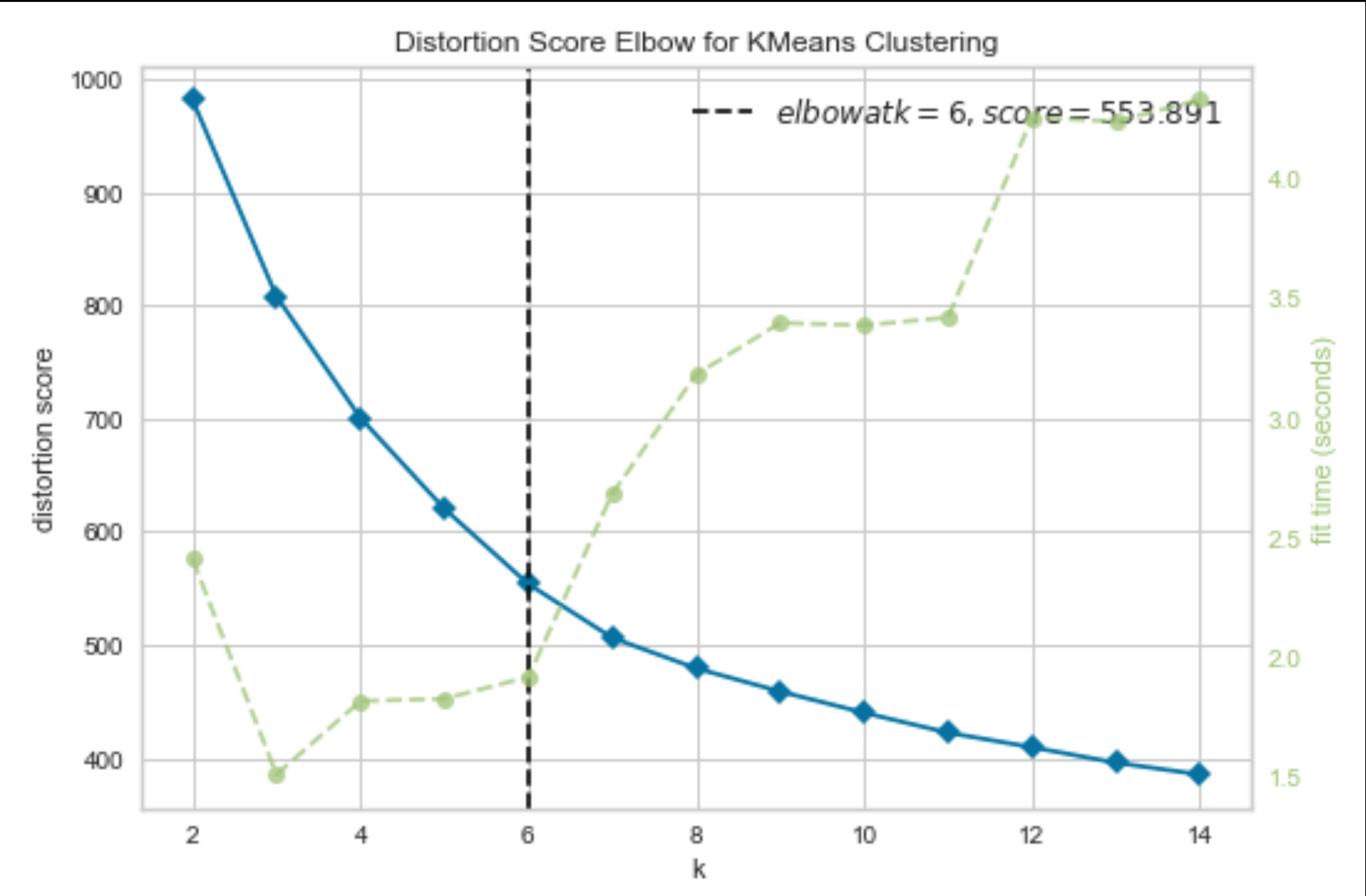


Heatmap


danceability
energy
speechiness
acousticness
instrumentalness
liveness
valence
danceability
energy
speechiness
acousticness
instrumentalness
liveness
valence

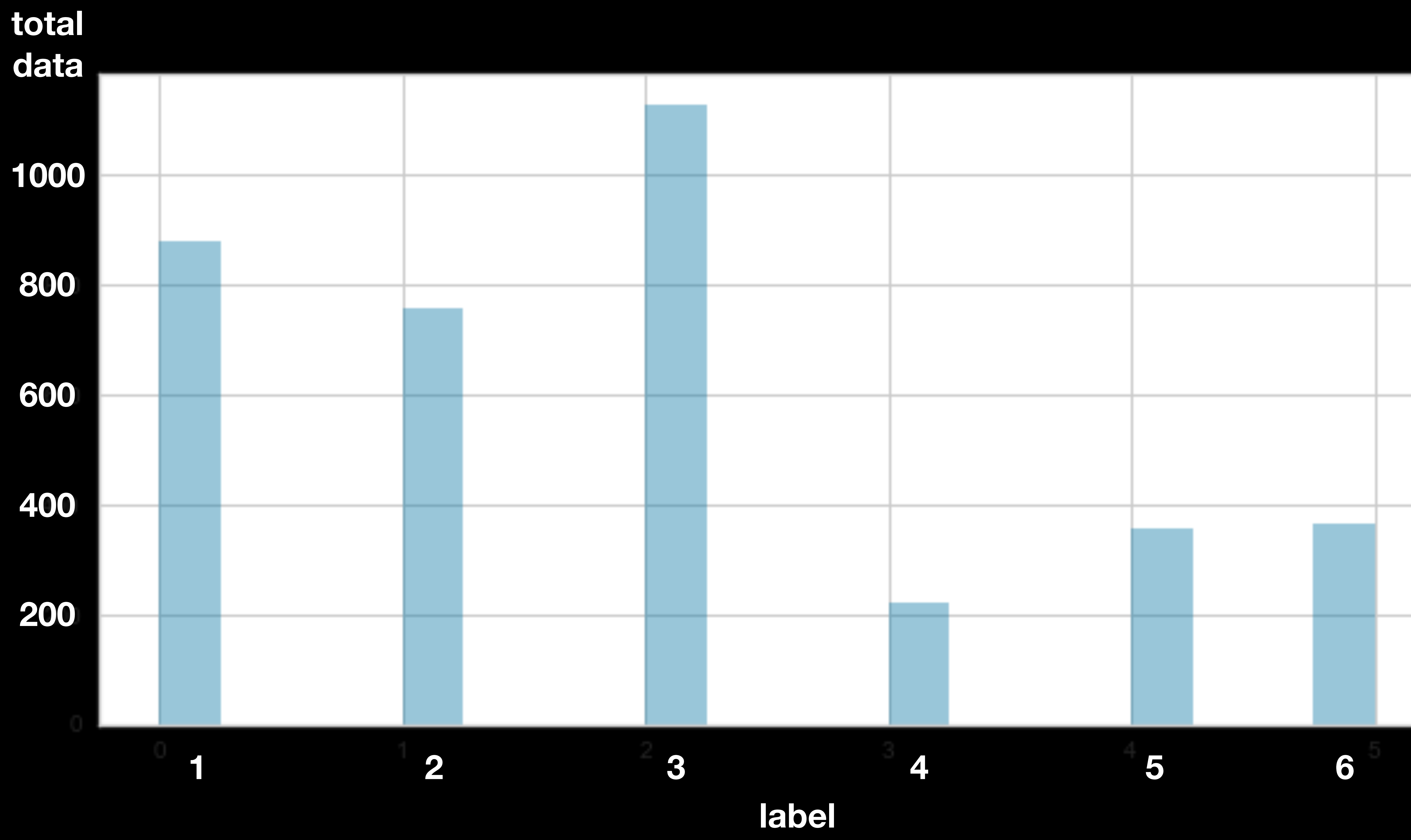
Clustering

Determining the amount of cluster using elbow method

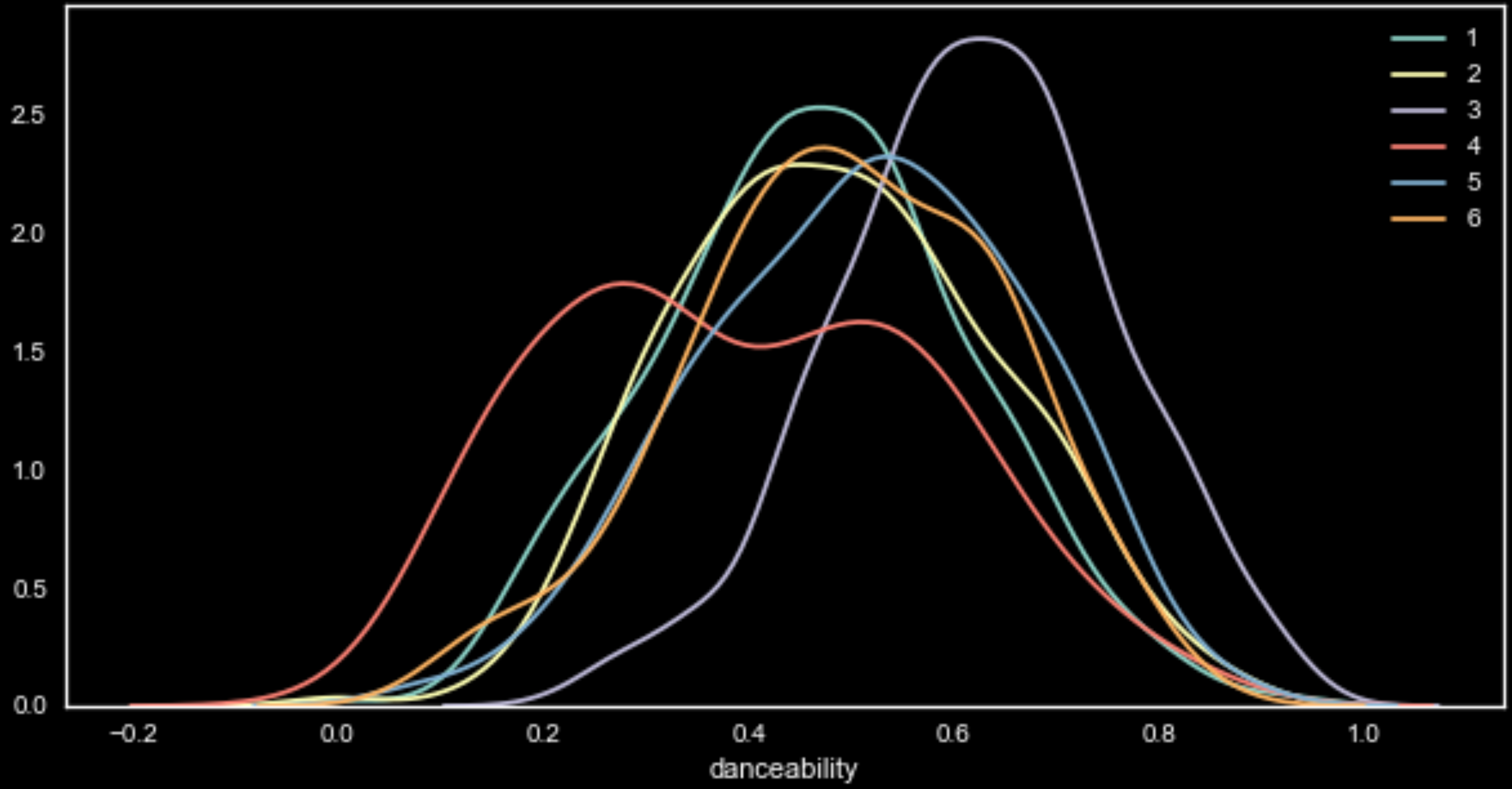


Clustering Result

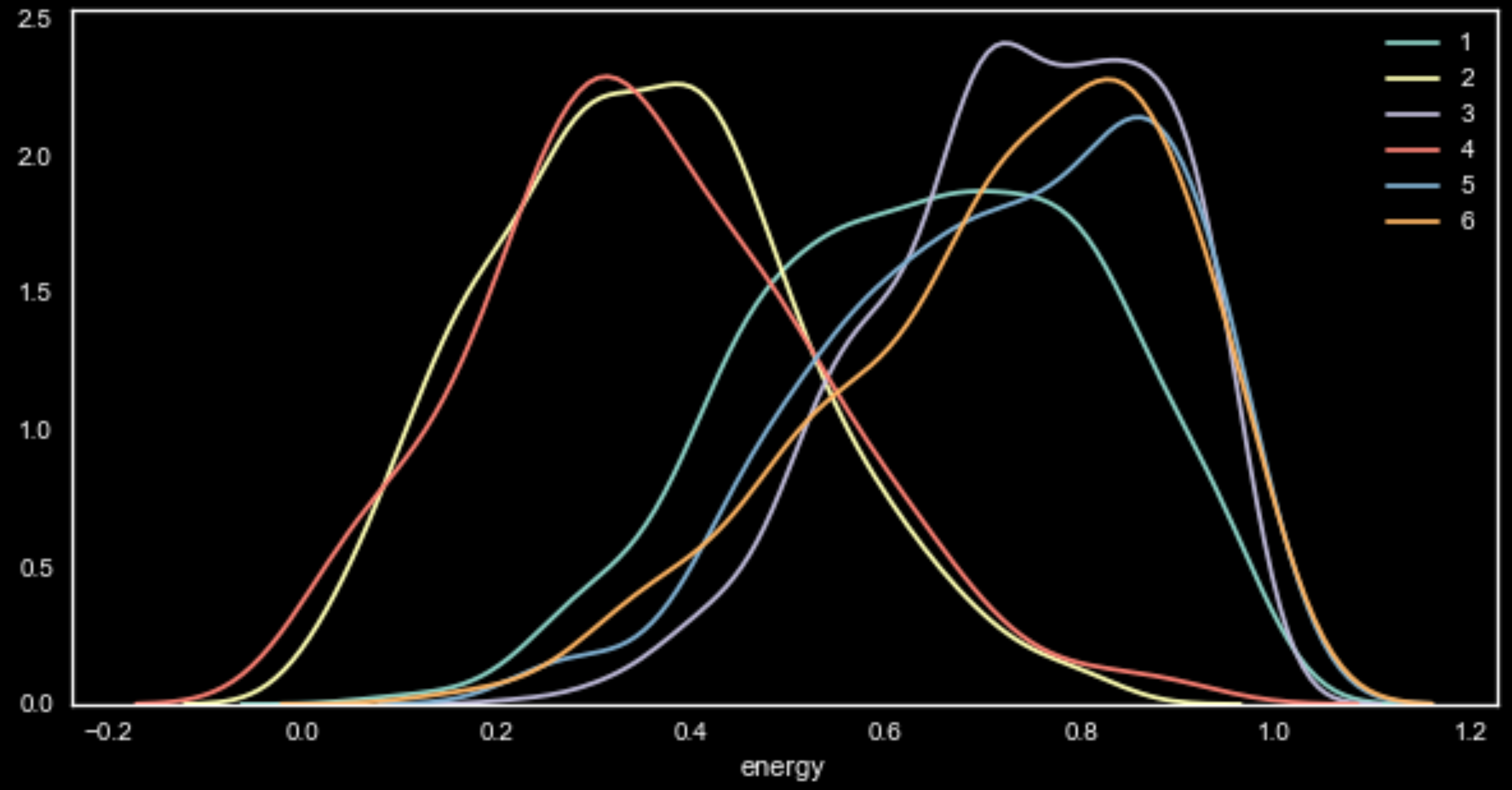
Total data per label



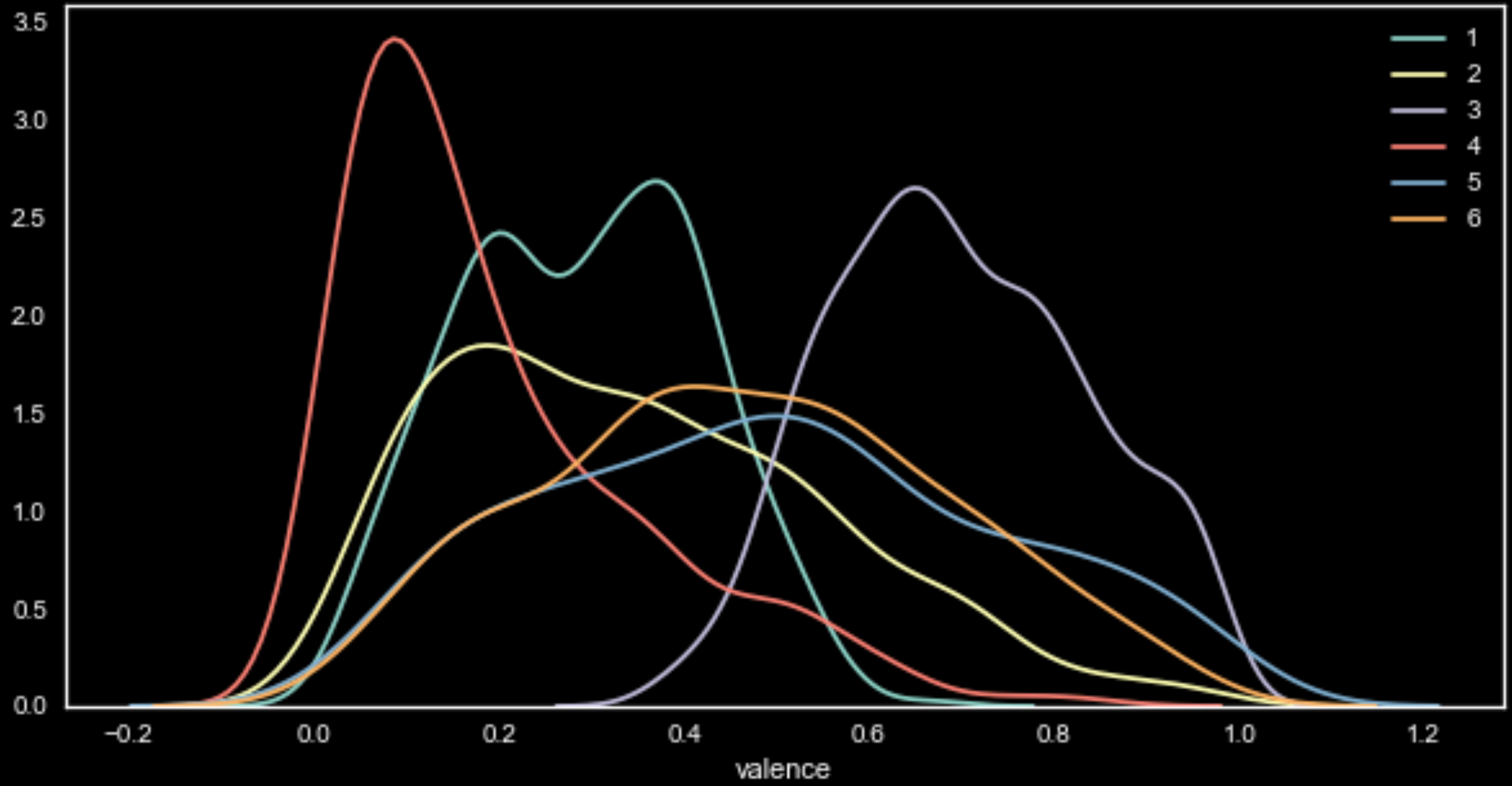
Clustering Result



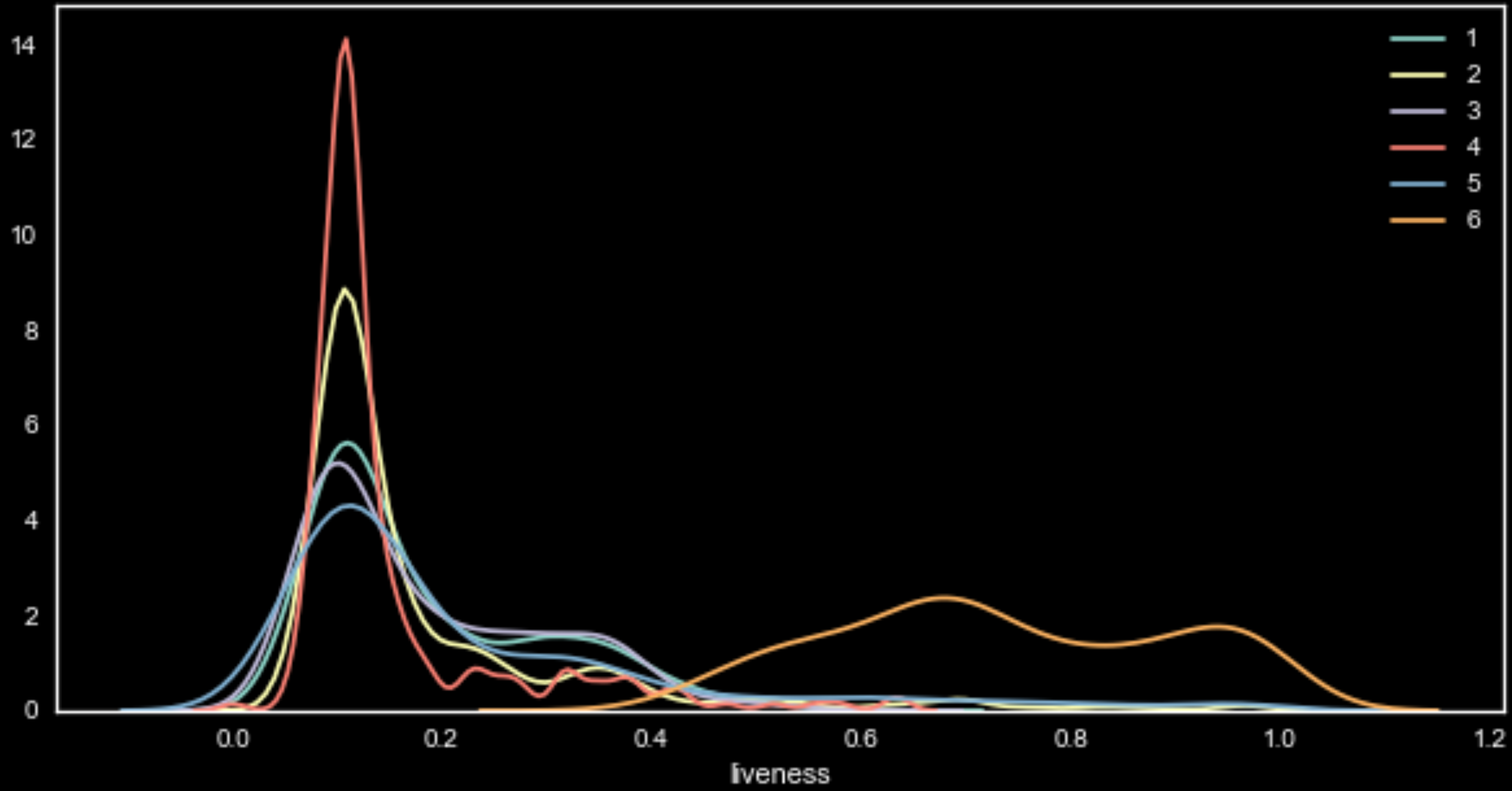
Clustering Result



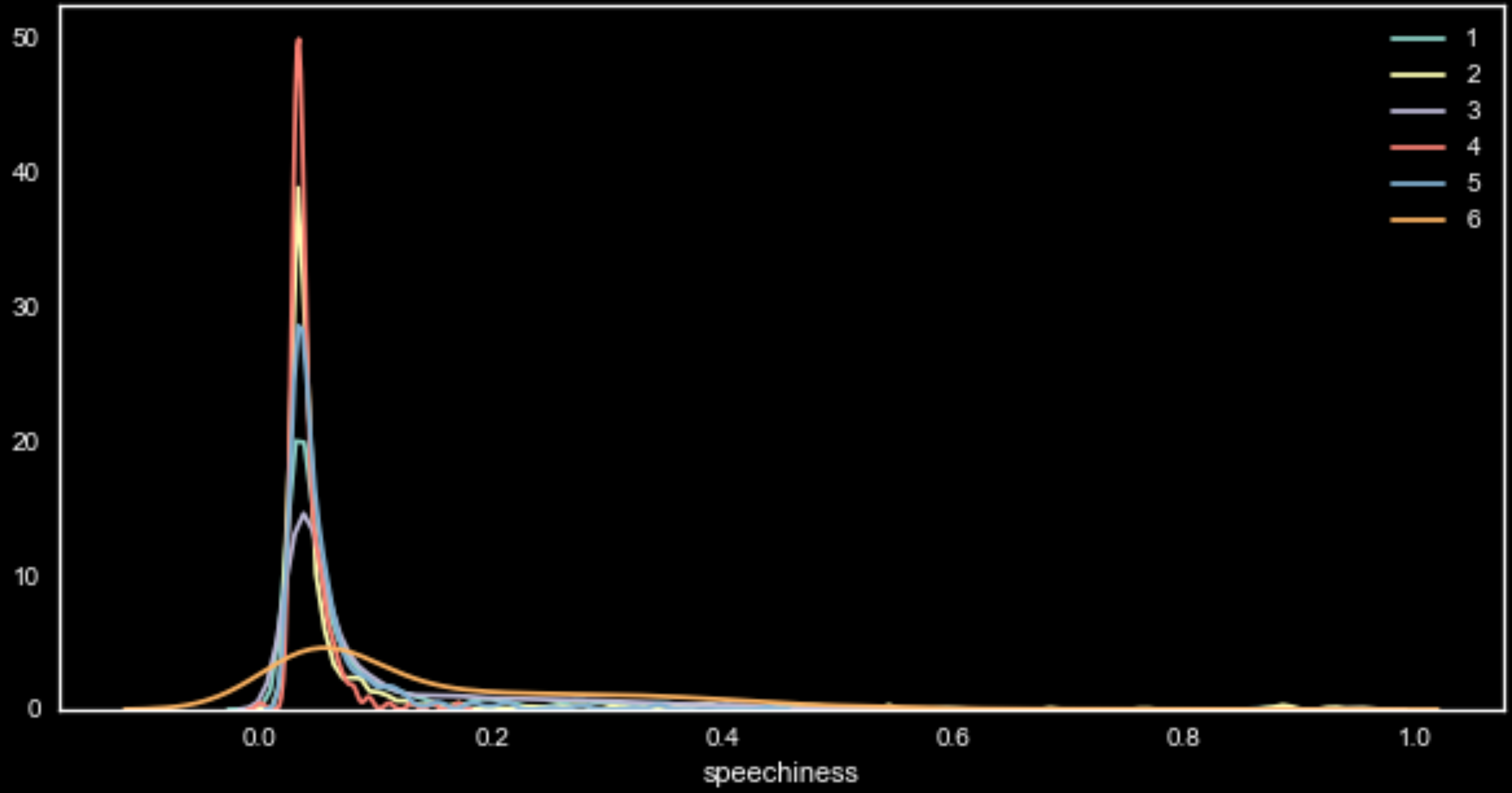
Clustering Result



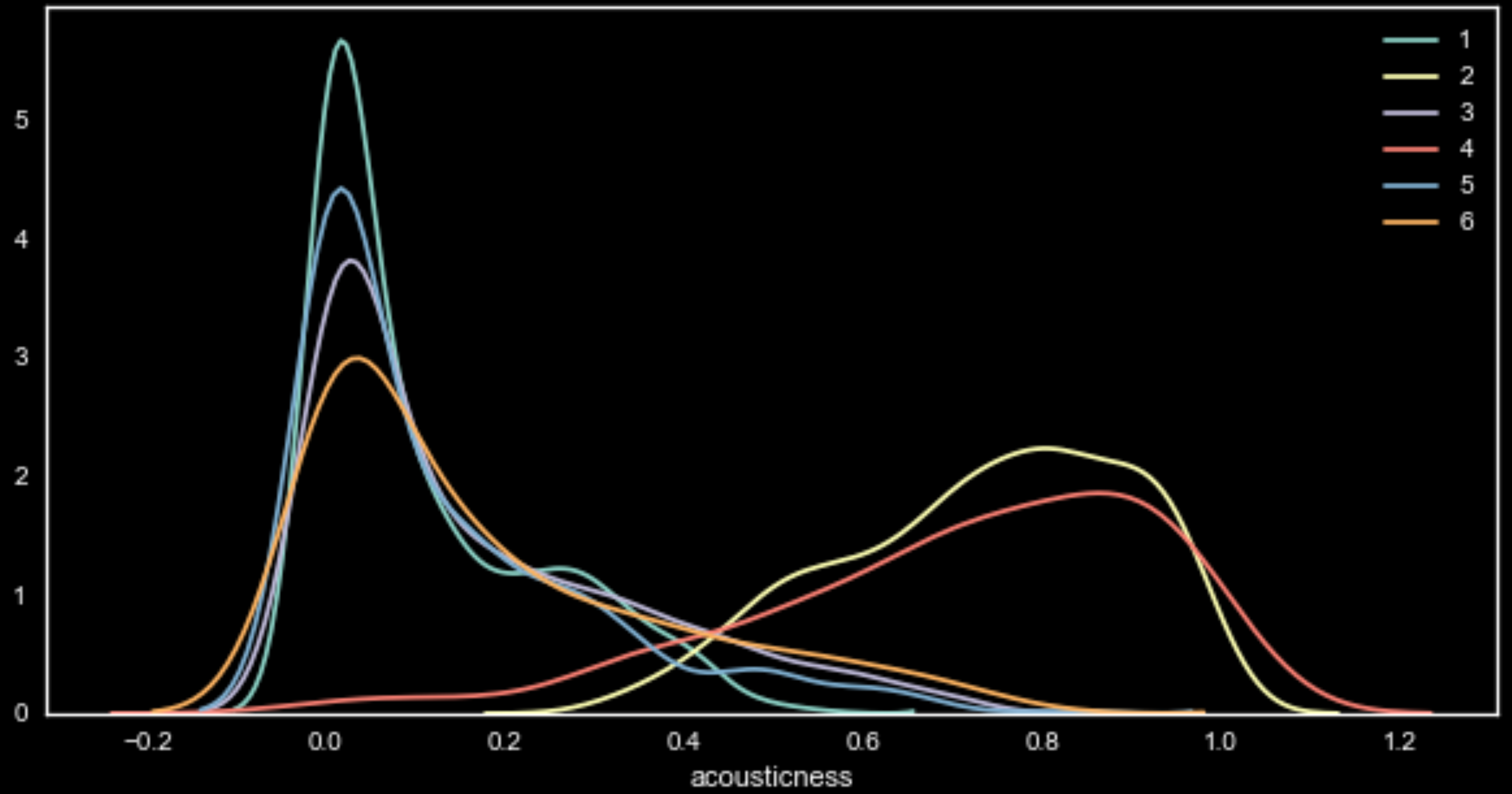
Clustering Result



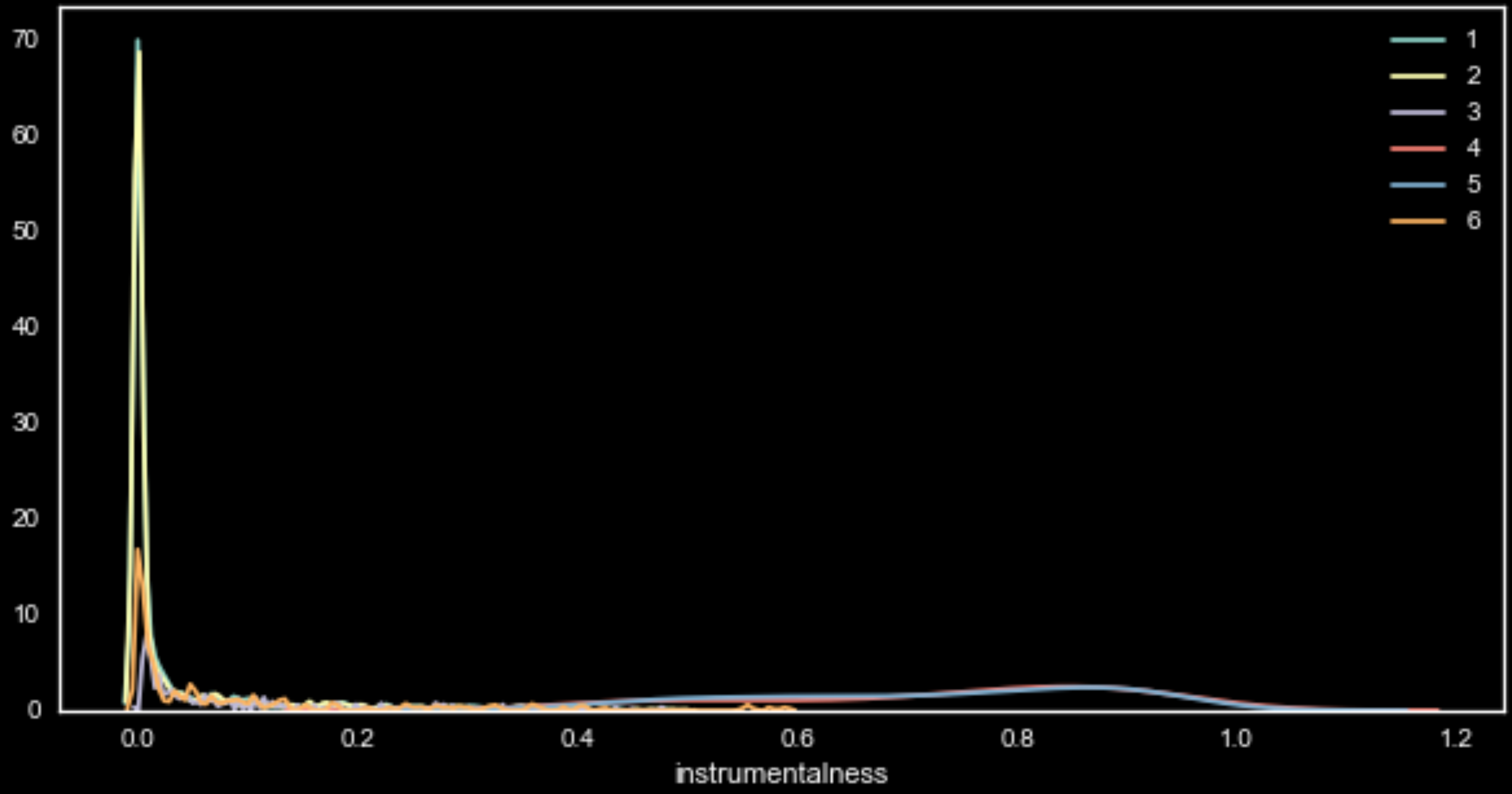
Clustering Result



Clustering Result



Clustering Result



Clustering Result

	label 1	label 2	label 3	label 4	label 5	label 6
danceability			+	-		
energy	+	-	+	-	+	+
valence	-	-	+	- -		
liveness						+
speachiness						+
acousticness	-	+	-	+	-	-
instrumentalness	-	-				

Determining Correlation Score

To make a recommendation, I use the Pearson correlation score between every song in target dataset with every clustered data mean

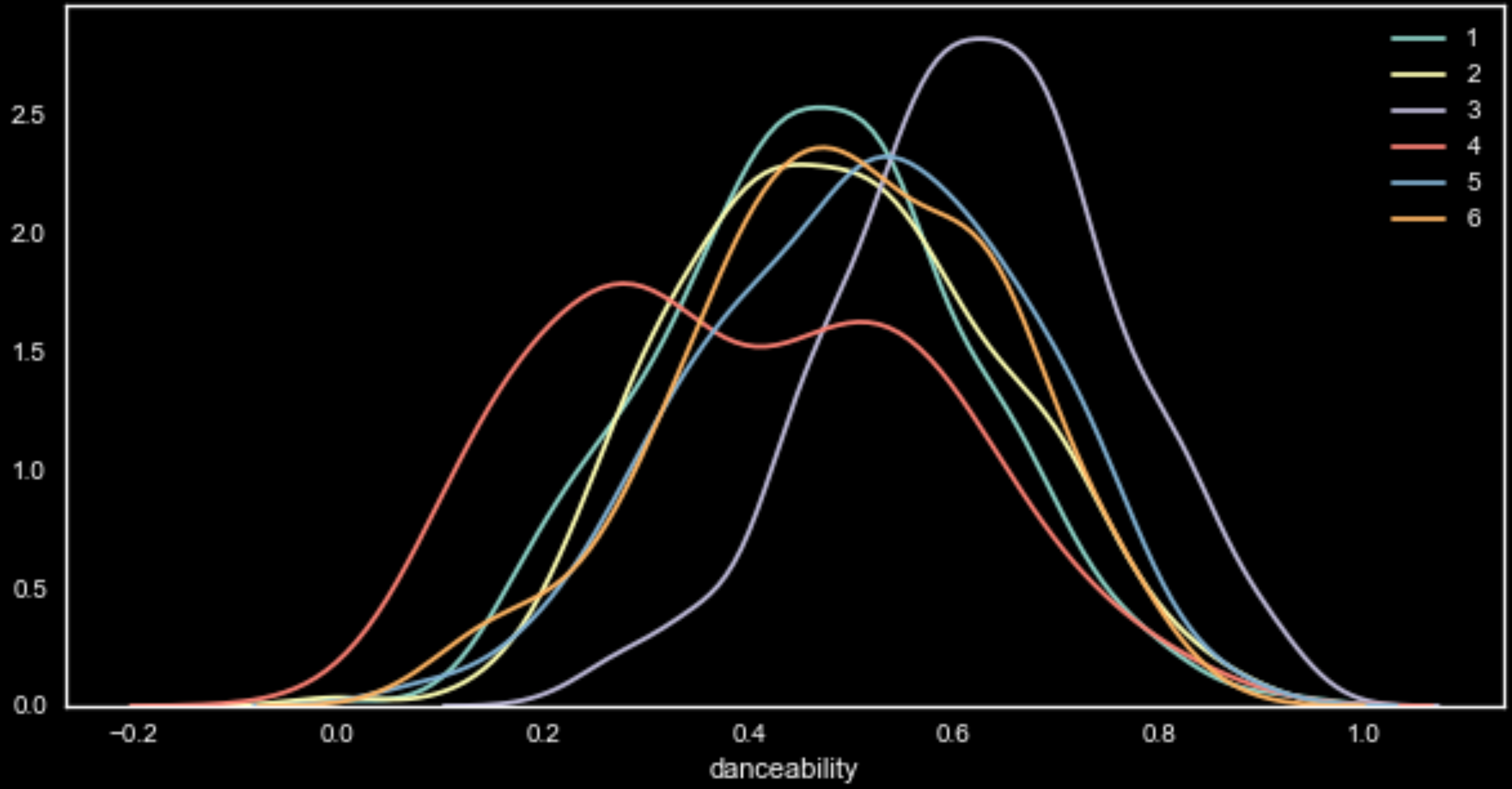
	danceability	energy	speechiness	acousticness	instrumentalness	liveness	valence	playlist_num
count	364.000000	364.000000	364.000000	364.000000	364.000000	364.000000	364.000000	364.0
mean	0.493619	0.720467	0.155265	0.180223	0.037332	0.734964	0.465631	6.0
std	0.154033	0.175032	0.151905	0.198193	0.096114	0.159638	0.214546	0.0
min	0.073200	0.150000	0.024200	0.000261	0.000000	0.392000	0.036300	6.0
25%	0.392500	0.603750	0.044000	0.013900	0.000000	0.616000	0.319250	6.0
50%	0.495500	0.742500	0.083750	0.107000	0.000119	0.705500	0.458500	6.0
75%	0.613250	0.852250	0.240000	0.297500	0.012475	0.885000	0.613250	6.0
max	0.849000	0.990000	0.877000	0.787000	0.589000	0.995000	0.938000	6.0

Determining Correlation Score

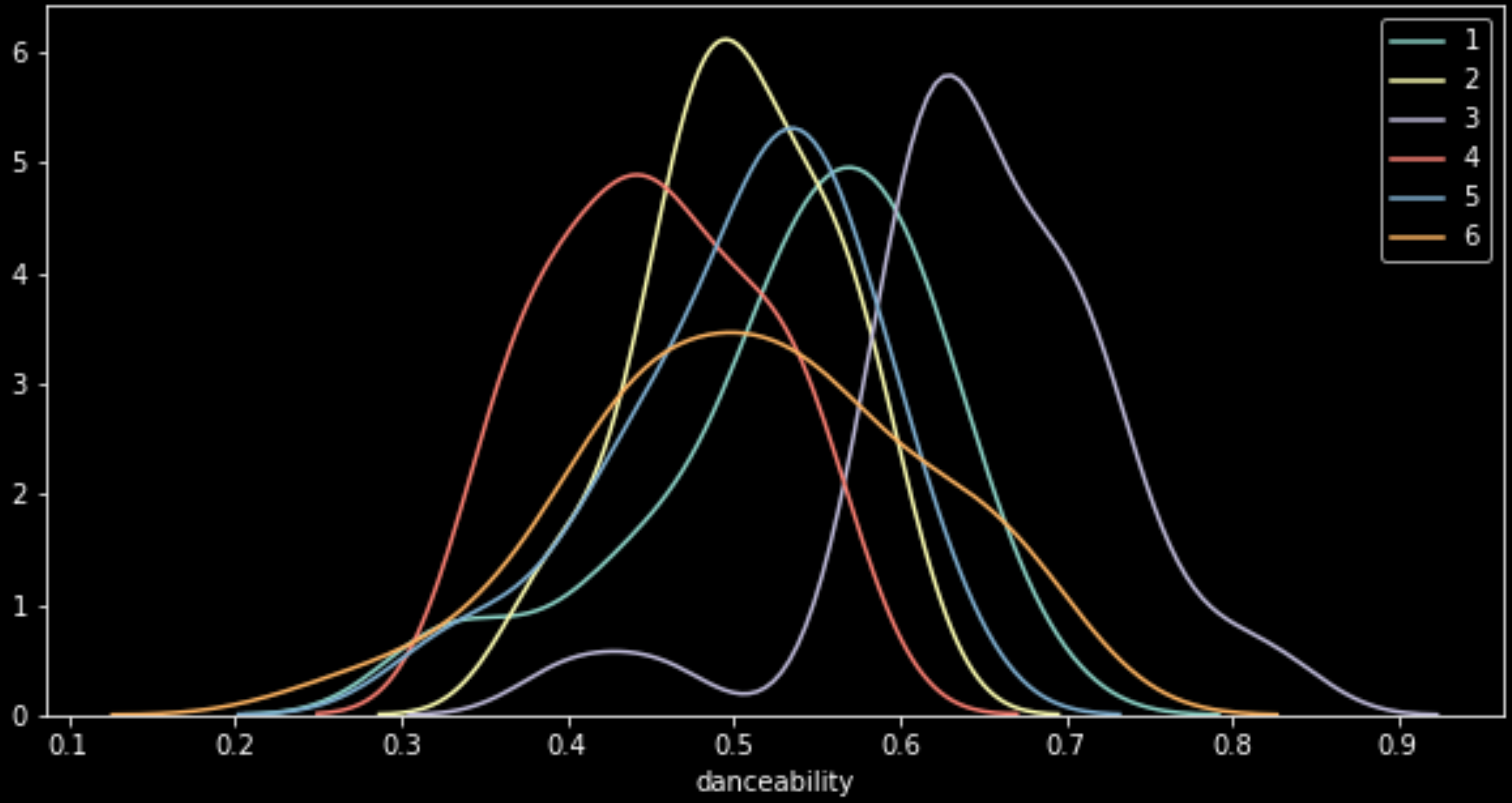
Then, we rank each correlation column, I'm taking top 30 song to make it into the playlist

correlation_1	correlation_2	correlation_3	correlation_4	correlation_5	correlation_6
0.822570	0.738156	0.852581	0.057260	0.290536	0.565410
0.837383	0.685244	0.909335	0.063879	0.388162	0.476459
0.881624	0.200310	0.949654	-0.380858	0.474921	0.824998
0.881463	0.231824	0.981656	-0.325617	0.518497	0.732041
0.849496	0.222263	0.976049	-0.331488	0.512848	0.715090
...	...	...	...	...	...
0.853761	0.116421	0.912979	-0.364646	0.514513	0.749616
0.408200	-0.379333	0.227522	0.311487	0.901277	0.207451
0.335697	-0.249639	0.154019	0.532645	0.899407	-0.036174
0.804424	-0.052076	0.610055	0.052455	0.749859	0.467943
0.928998	0.098818	0.738307	-0.138943	0.580388	0.684367

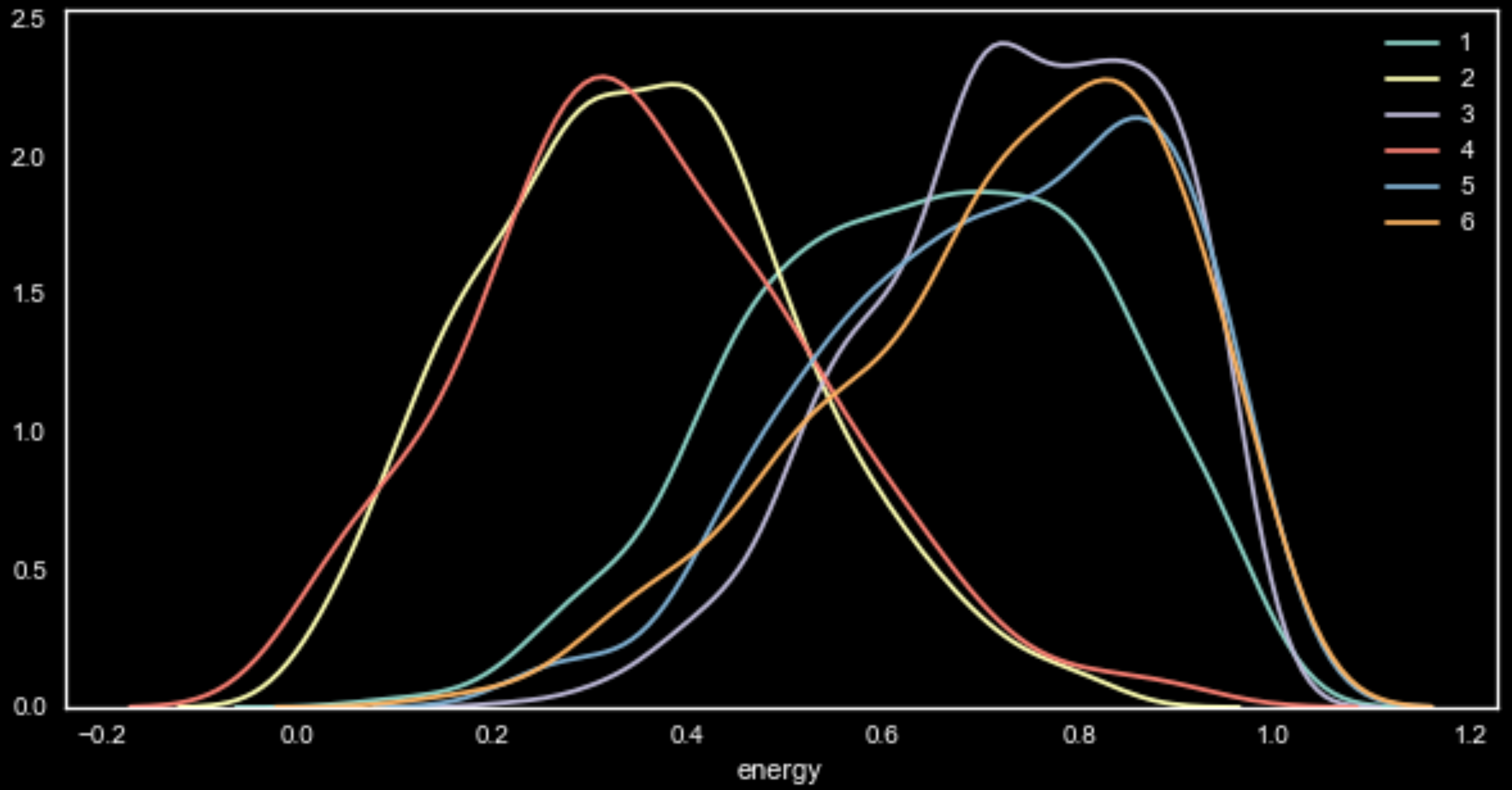
Feature Comparison - From Library



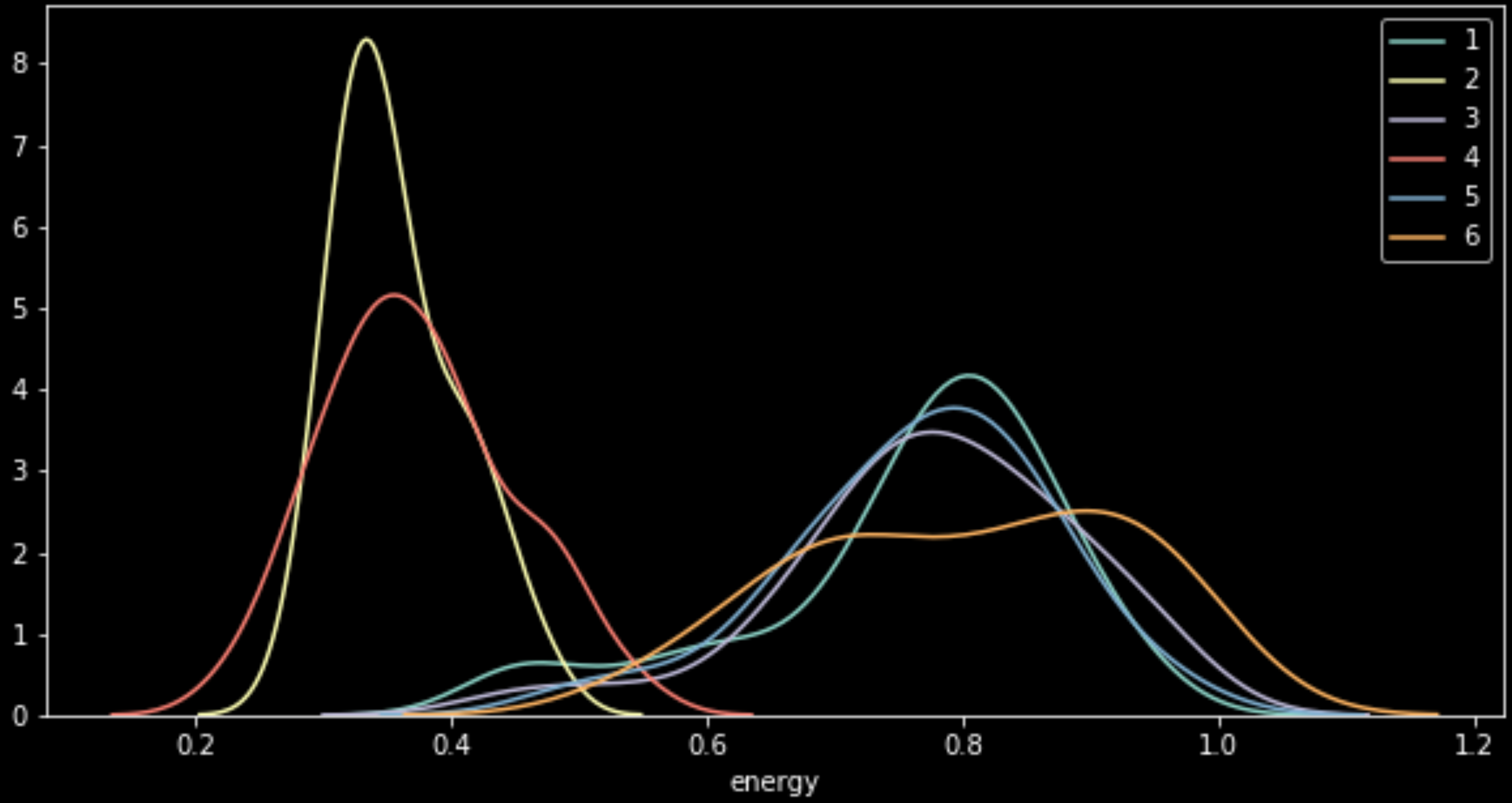
Feature Comparison - New Playlist



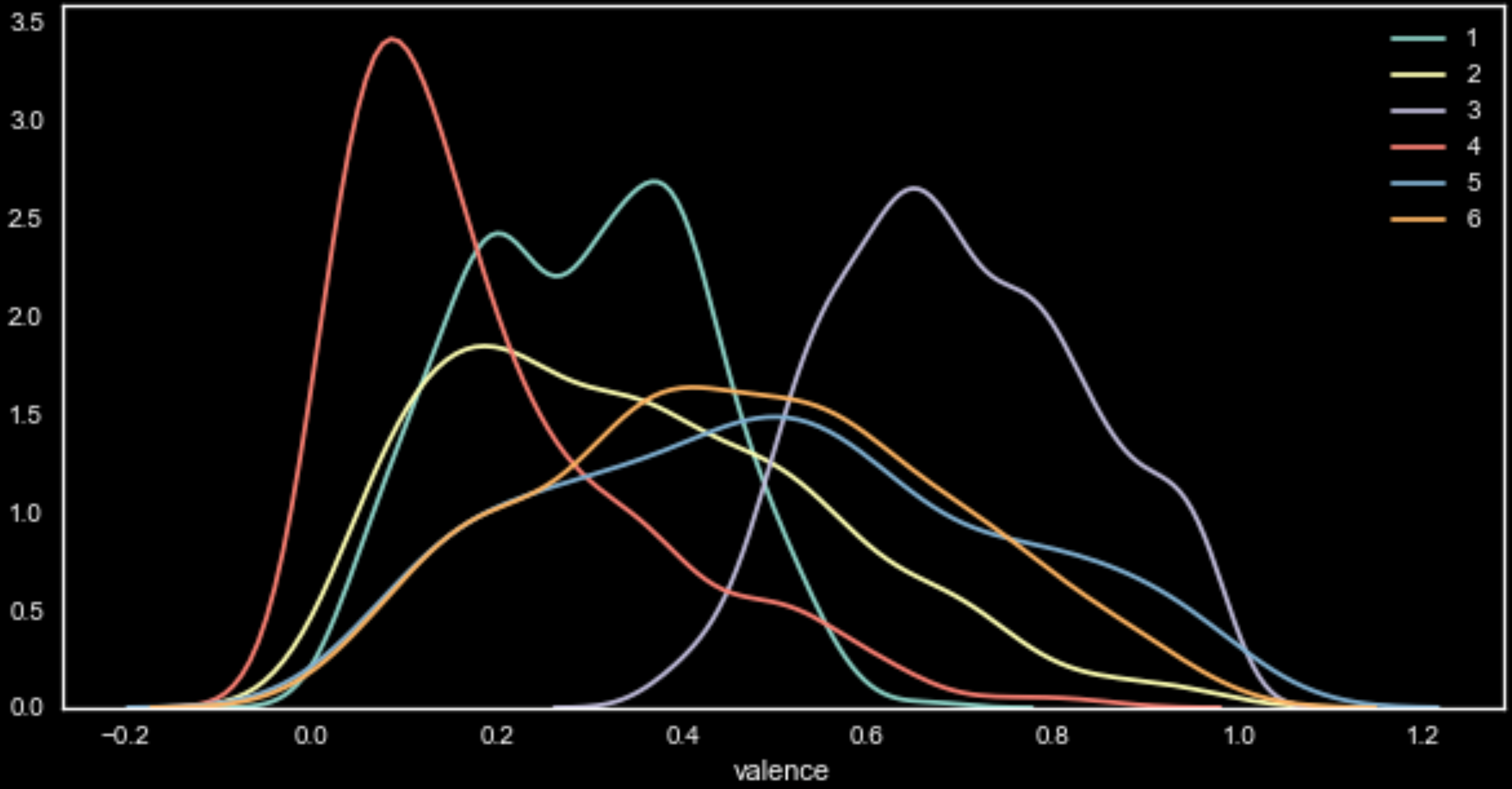
Feature Comparison - From Library



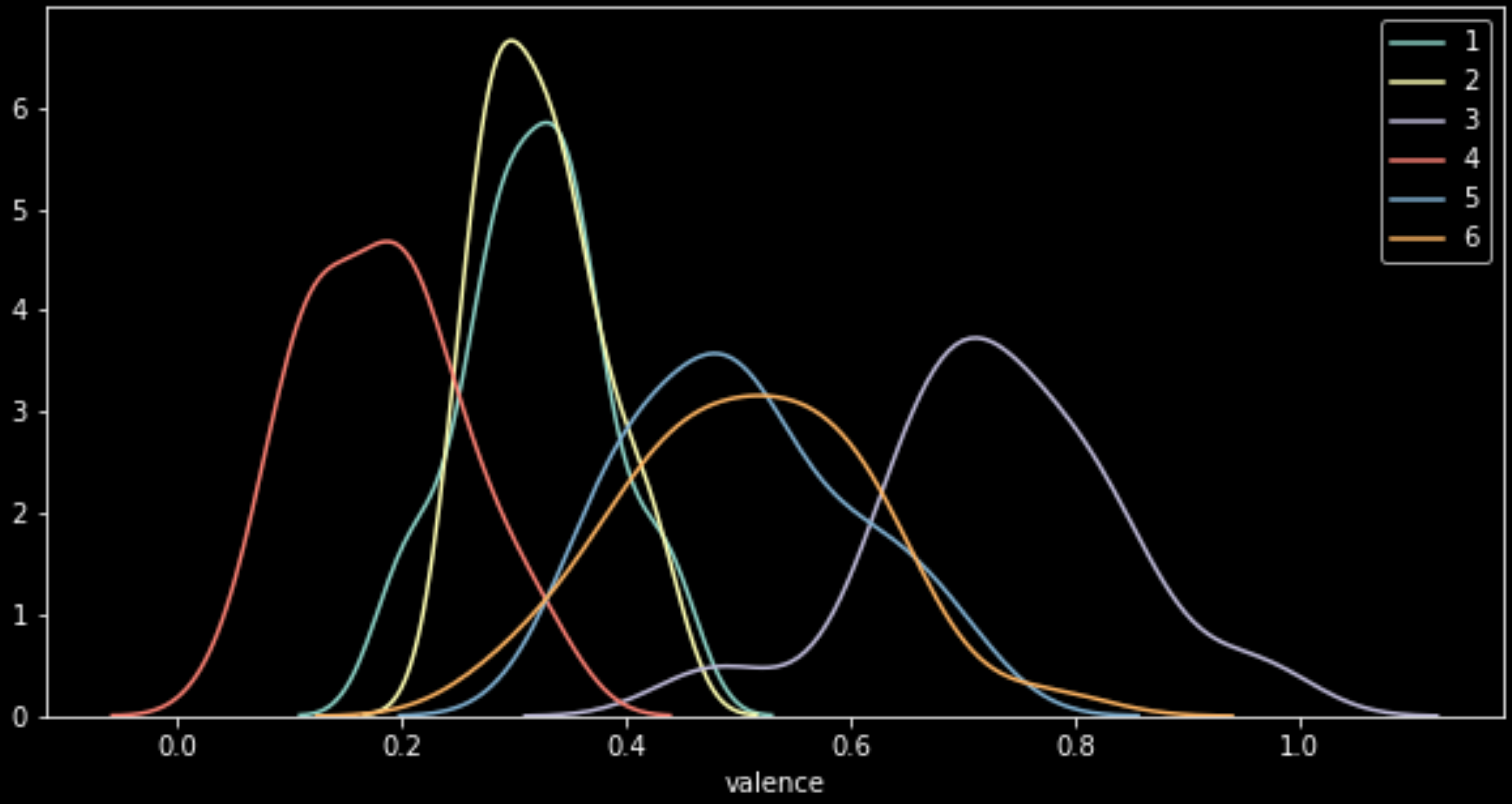
Feature Comparison - New Playlist



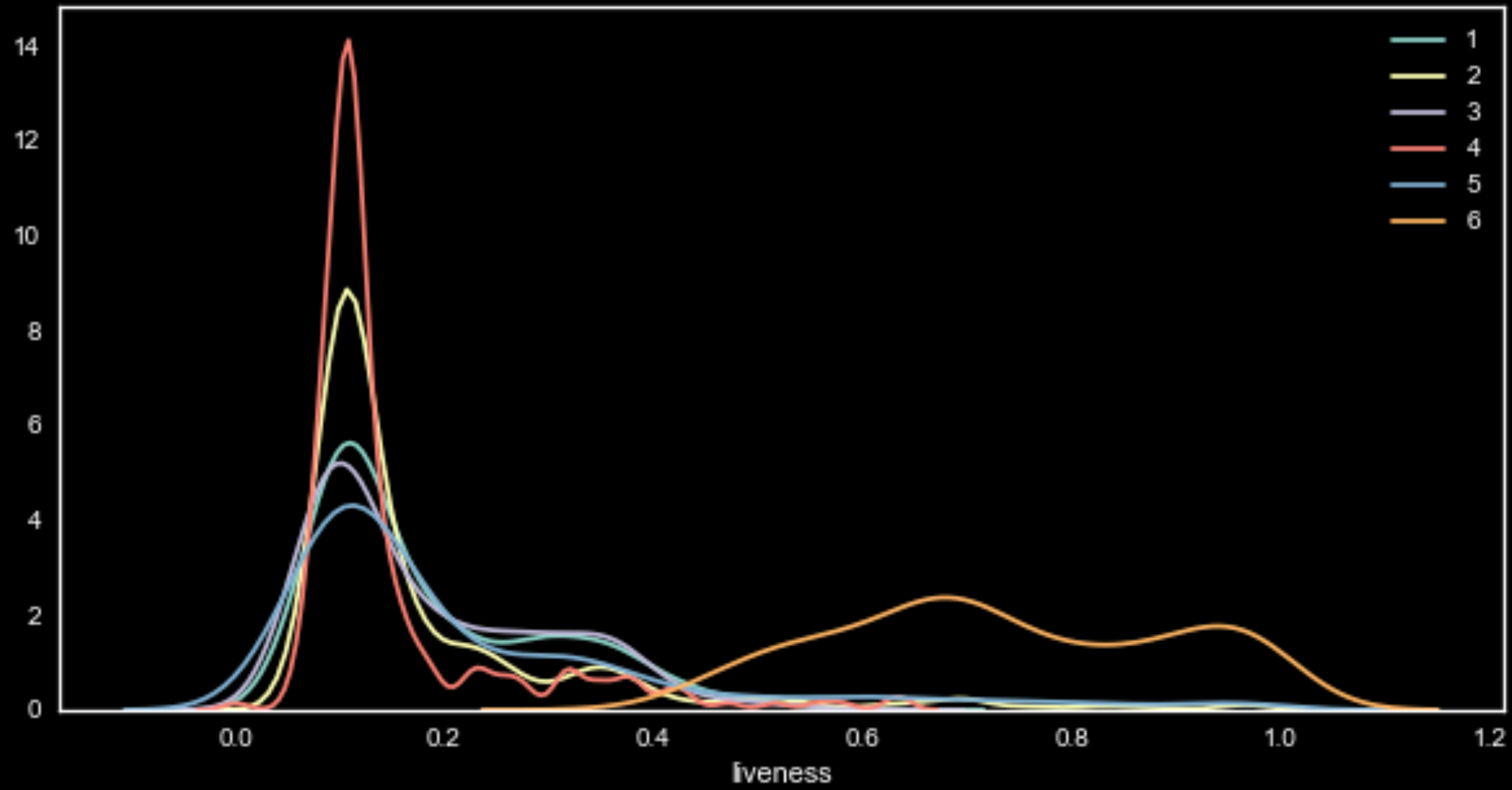
Feature Comparison - From Library



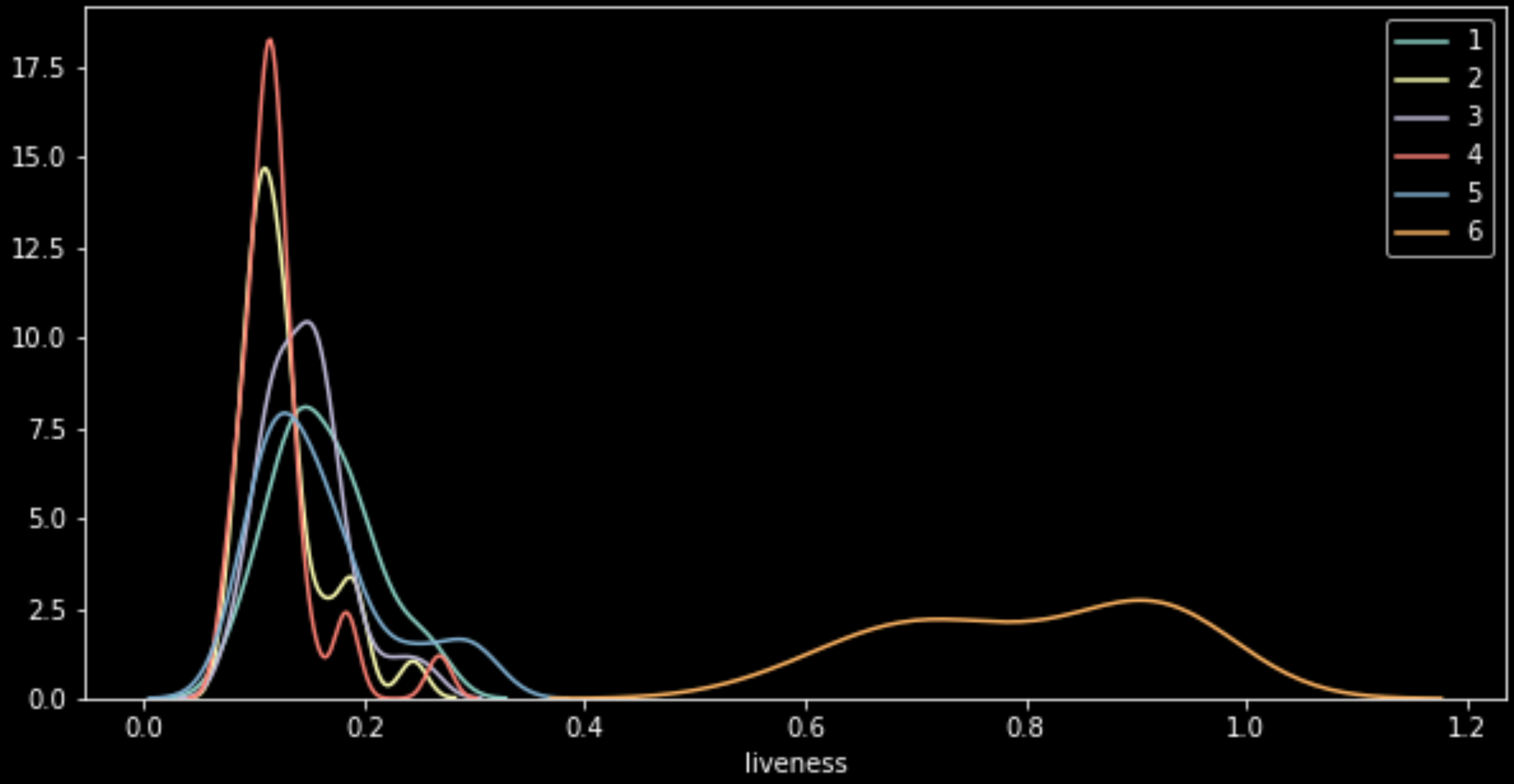
Feature Comparison - New Playlist



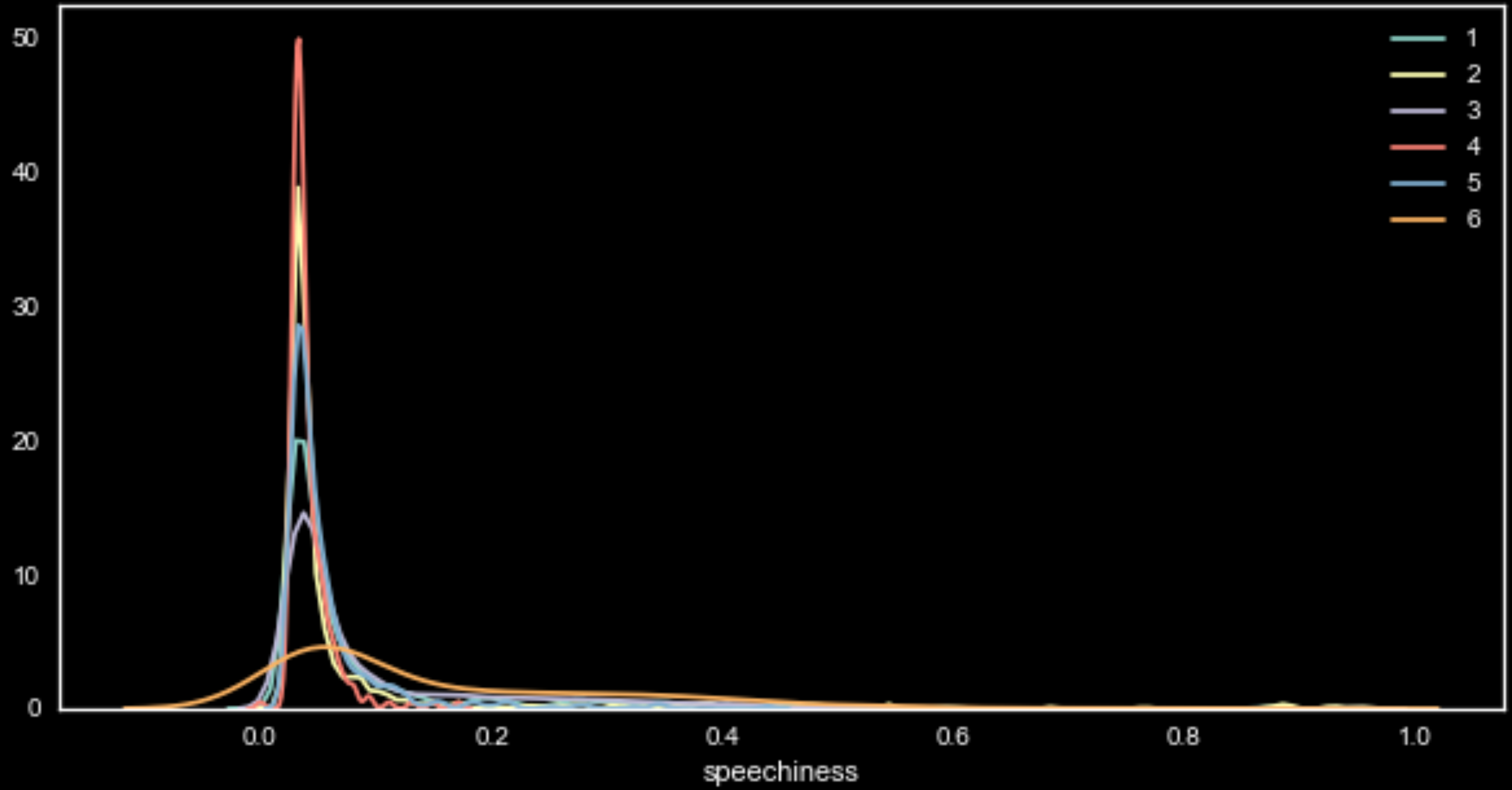
Feature Comparison - From Library



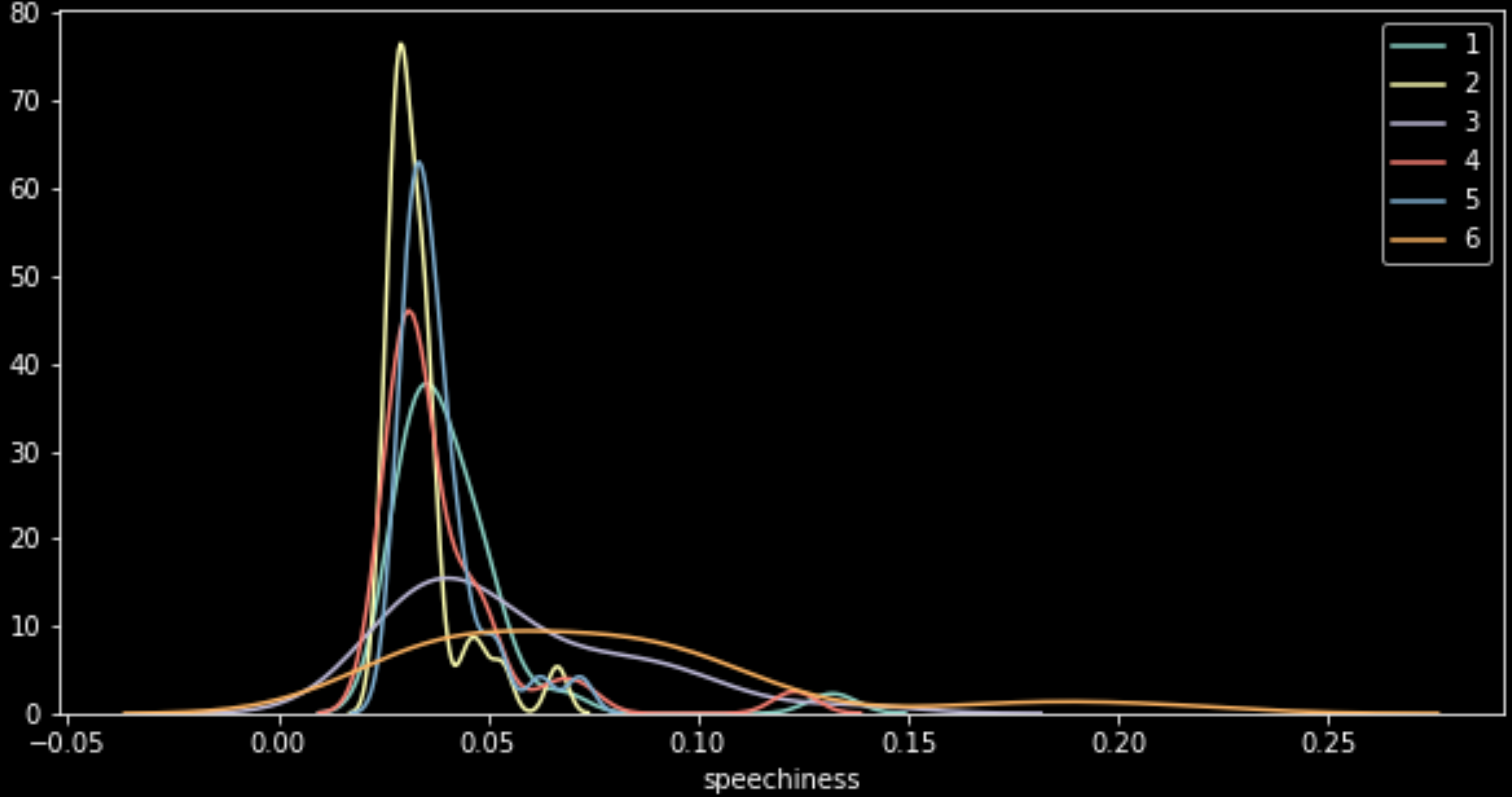
Feature Comparison - New Playlist



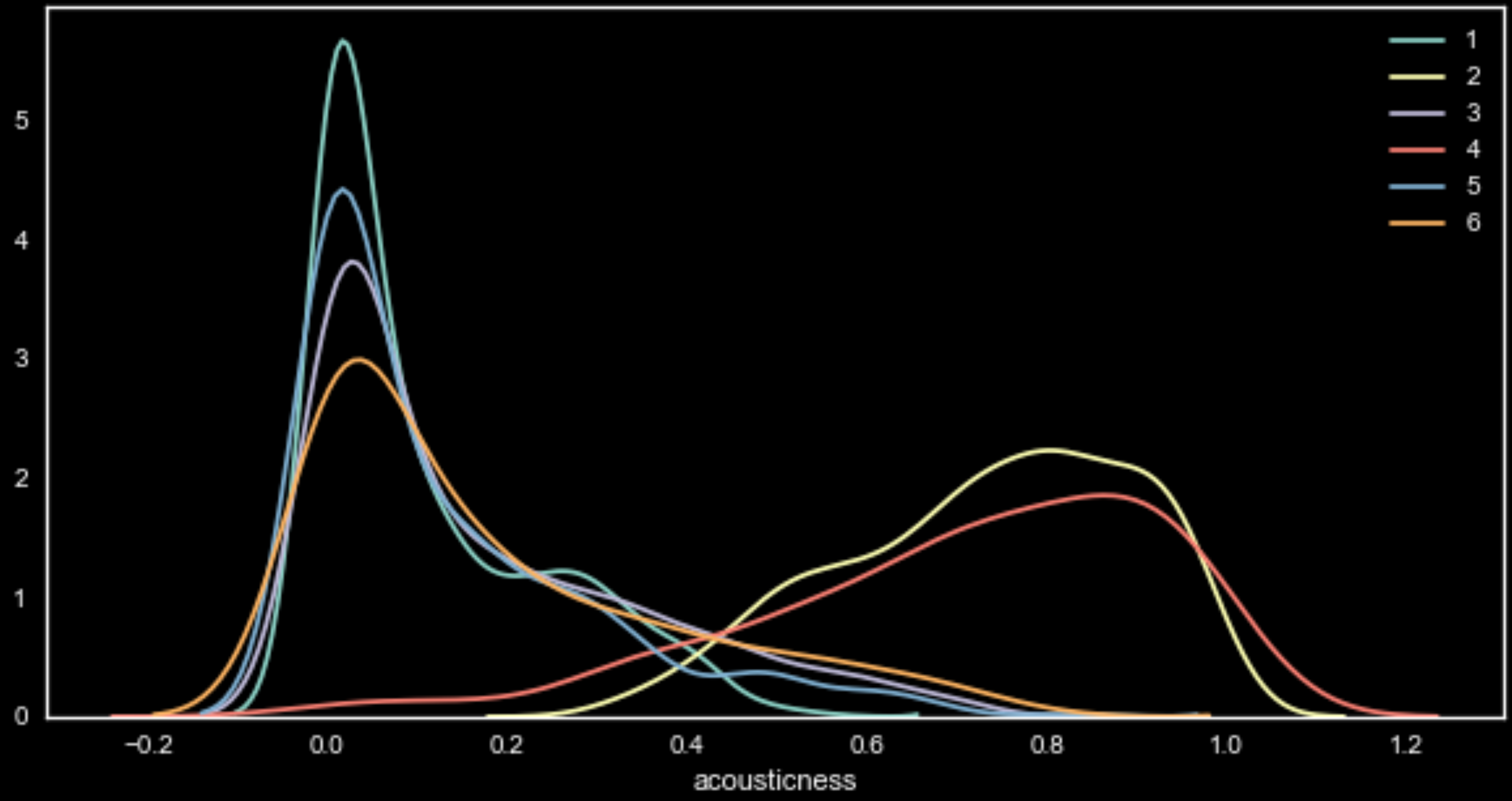
Feature Comparison - From Library



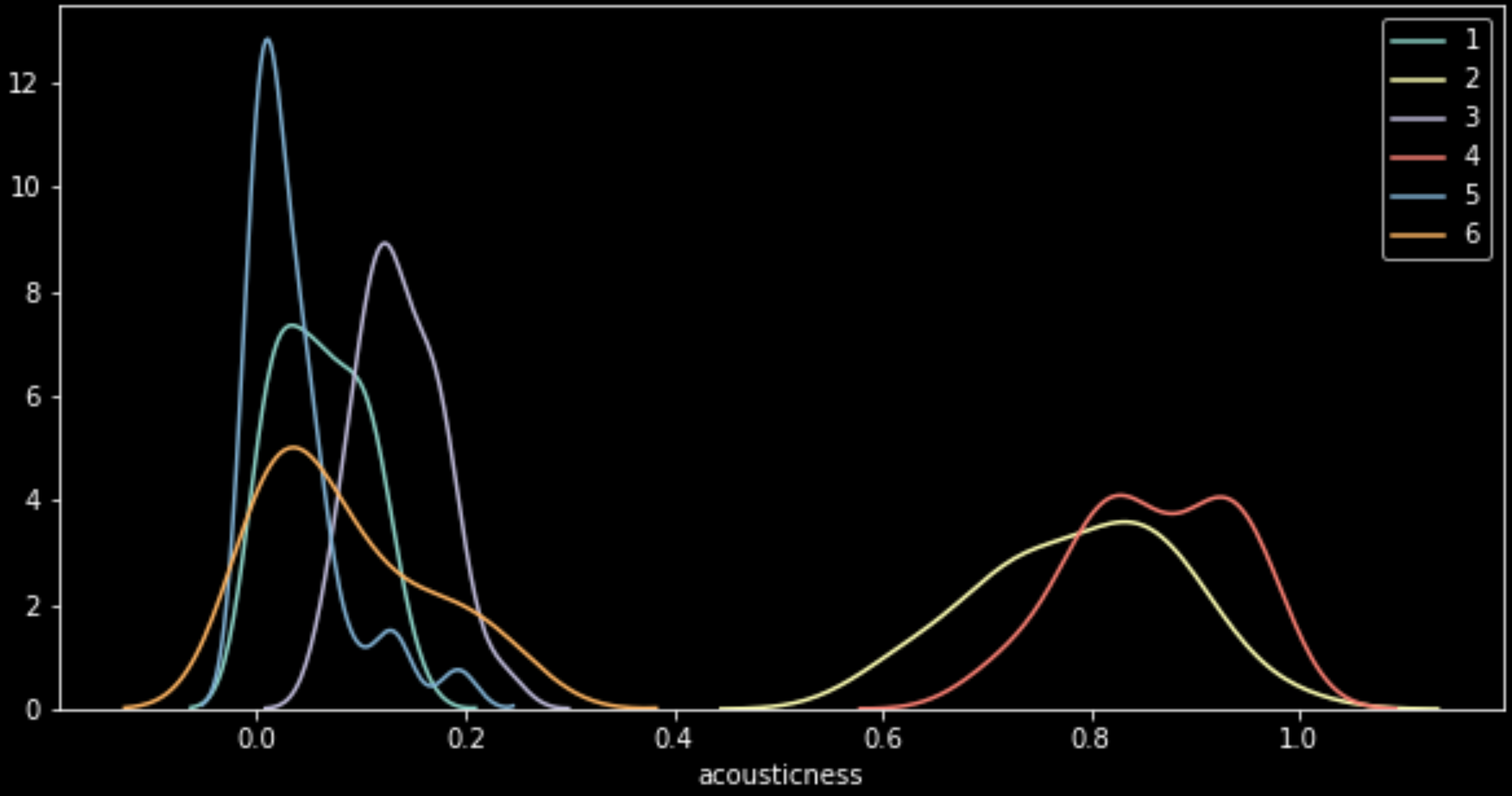
Feature Comparison - New Playlist



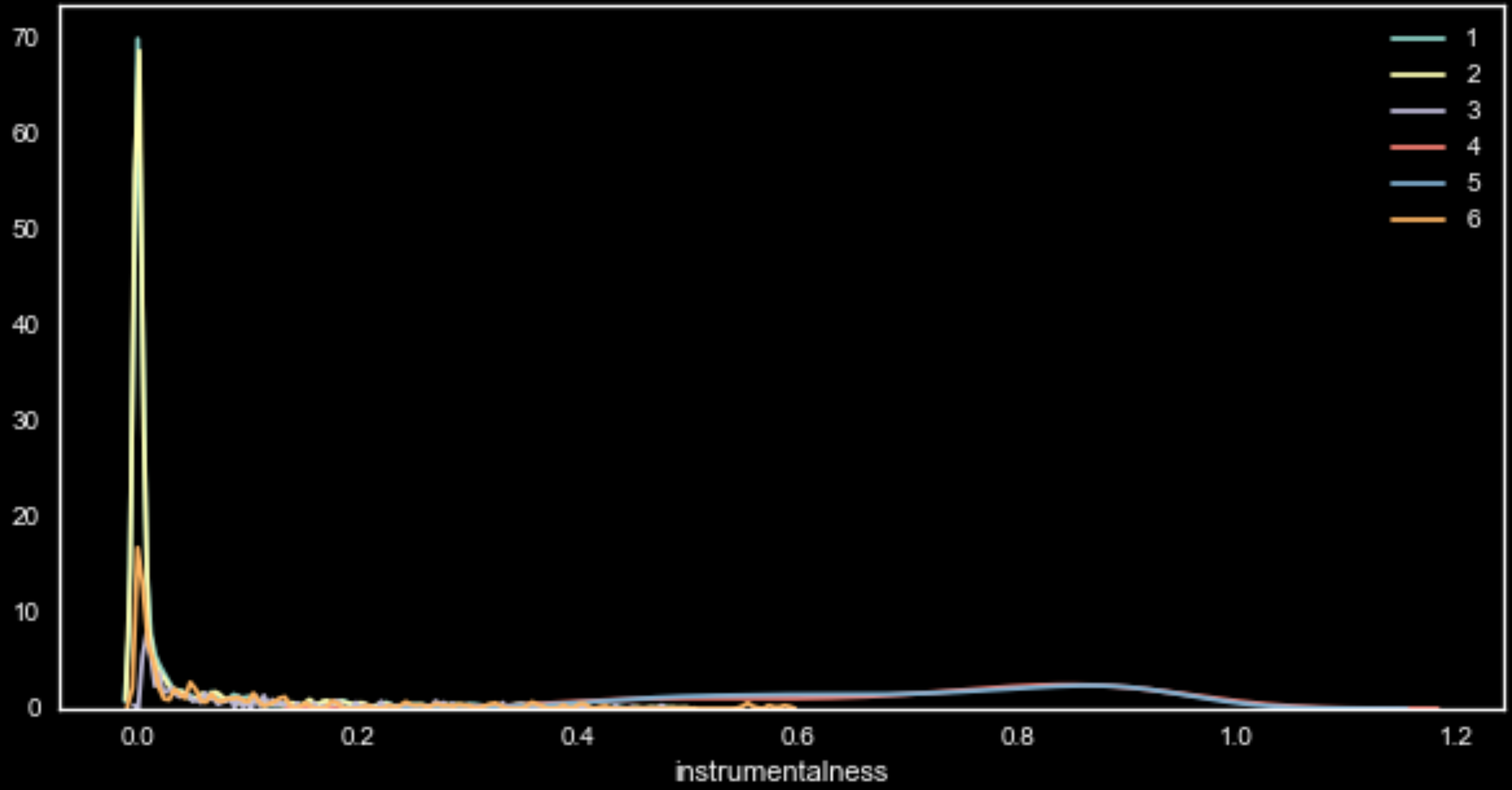
Feature Comparison - From Library



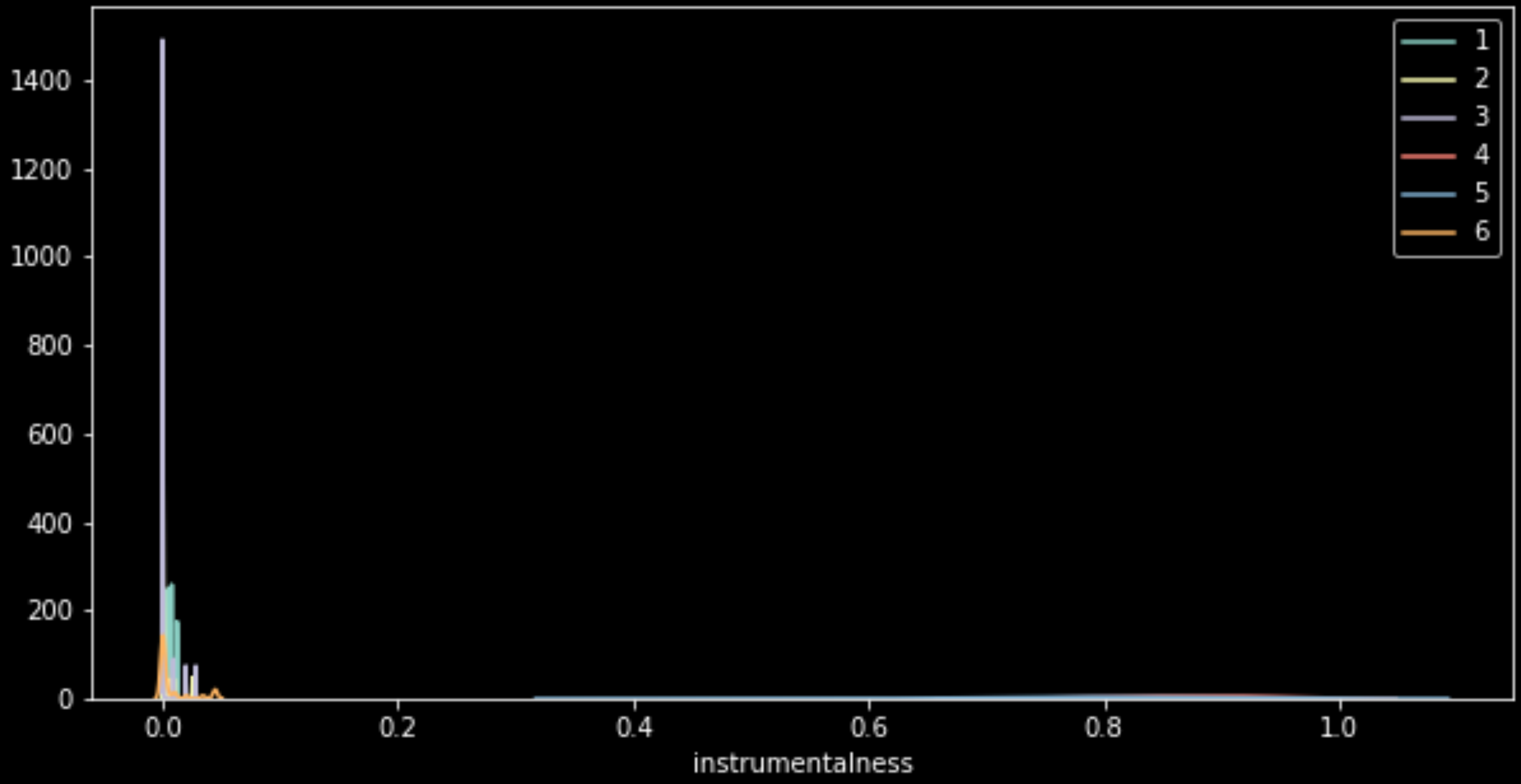
Feature Comparison - New Playlist



Feature Comparison - From Library



Feature Comparison - New Playlist



Playlist Link

Music From Final Project 1: https://open.spotify.com/playlist/2evgGdaDagUHvDqhNNG7pM?si=jBFY-D0wTTO9kSq2Qknitg

Music From Final Project 2: https://open.spotify.com/playlist/0vcZPqe7JOKqWOkRi62vid?si=MyRaBV0ZQsWiRW-jH6utWQ

Music From Final Project 3: https://open.spotify.com/playlist/2yoeICv7TkSqnh220UMwIE?si=nD99ZQfzT12Y3tLtmT1X7Q

Music From Final Project 4: https://open.spotify.com/playlist/6Buvk2gEbhZ4lJwlIQGUhE?si=W3usy4IiRo-aF5KT_dTcxQ

Music From Final Project 5: https://open.spotify.com/playlist/2HmaeG1YCp4fzvextlLepg?si=DzW60NWHQ1aOHz5HiTrvgA

Music From Final Project 6: https://open.spotify.com/playlist/4pdGZJ8j5GIcAkfkrtkm5H?si=LYEtoo9fRTiasOEU09IpSw

Conclusion and Recommendation

Conclusion

This recommendation system focus on the feeling that clustered from user's library, because of that, each playlist have a similar feeling and seldom go out of that corridor. The weakness is there are no content filtering such as genre filtering, which is from what I can guess, is the first priority in Spotify recommendation system. Because of that, some song or artists in the playlist seems unfamiliar.

Recommendation

For next attempt at making music recommendation, I suggest to include content filtering, so the correlation score can be determine after content filtering. Because content filtering usually easier to identify

Thank You