UNSUPERVISED MACHINE LEARNING: CLUSTERING SONGS

WBS CODING SCHOOL
GROUP 1: ASLAM, HUY, ÍCARO & NABIL

Who is Moosic?

Moosic is a little start up that creates curated playlists done by music experts and specialists in old and new trends.

Objectives

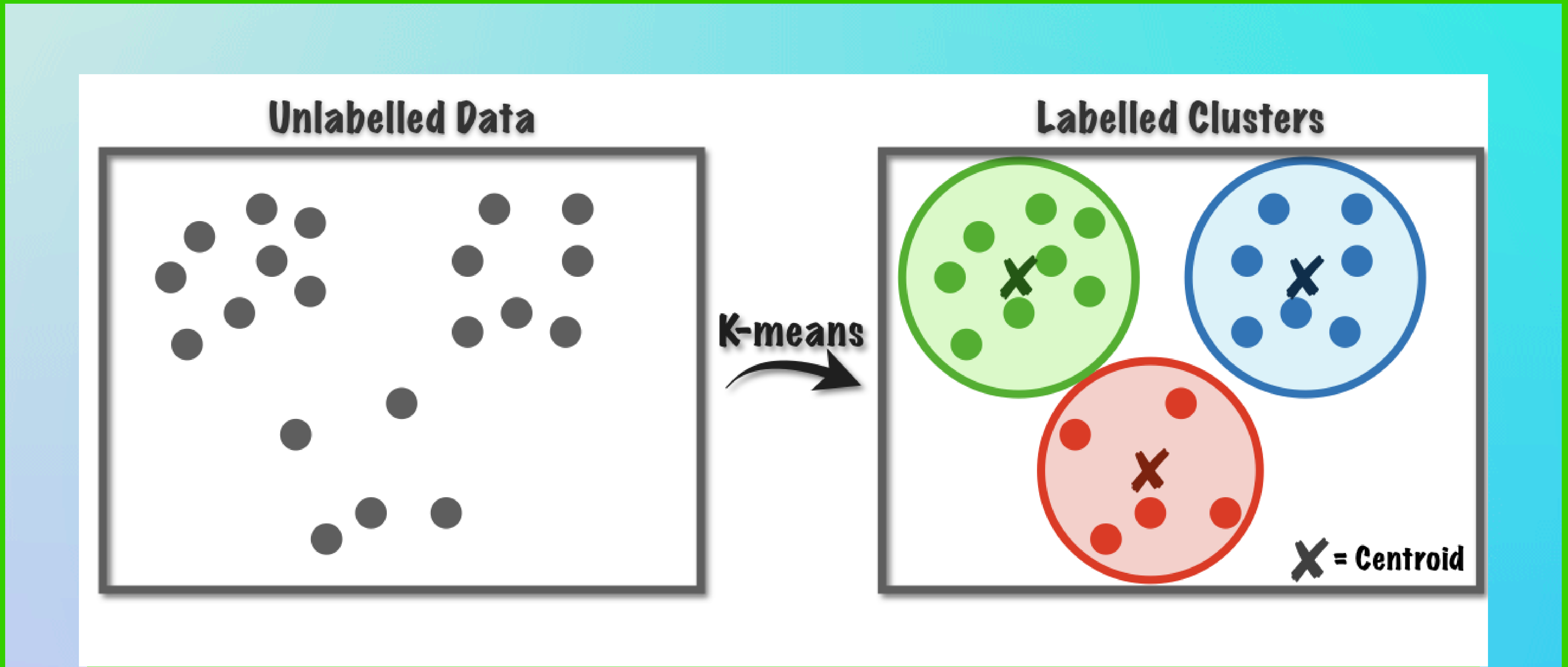
- **Automate the creation of playlists for Spotify using as parameters the features created by Spotify.**
  - **Are Spotify's audio features able to identify "similar songs", as defined by humanly detectable criteria?**
  - **Is K-Means a good method to create playlists?**

Features of the musics on Spotify

- **Acousticness**
  - A confidence measure from 0.0 to 1.0 of whether the track is acoustic. 1.0 represents high confidence the track is acoustic.
- **Danceability**
  - Danceability describes how suitable a track is for dancing
- **Energy**
  - *Energy is a measure from 0.0 to 1.0 and represents a perceptual measure of intensity and activity.*
- **Instrumentalness**
  - Predicts whether a track contains no vocals.
- **Key**
  - *The key the track is in. E.g. 0 = C, 1 = C♯/D♭, 2 = D*
- **Liveness**
  - *Detects the presence of an audience in the recording.*
- **Loudness**
  - *The overall loudness of a track in decibels (dB).*
- **Mode**
  - *Mode indicates the modality (major or minor) of a track*
- ***Speechiness***
  - *Speechiness detects the presence of spoken words in a track.*
- *Tempo*
- *The overall estimated tempo of a track in beats per minute (BPM).*
- ***Time Signature***
  - *An estimated time signature.*
- ***Valence***
- *A measure from 0.0 to 1.0 describing the musical positiveness conveyed by a track.*

Selective songs using KMeans number

- We used the KMeans method to split the Dataframe in X numbers of clusters.
- Each cluster will be made in a Playlist. For this playlist we choose the 20 songs closer to the centroid.
- The distance between the songs and the centroids were calculated using the Manhattan and Euclidean method.



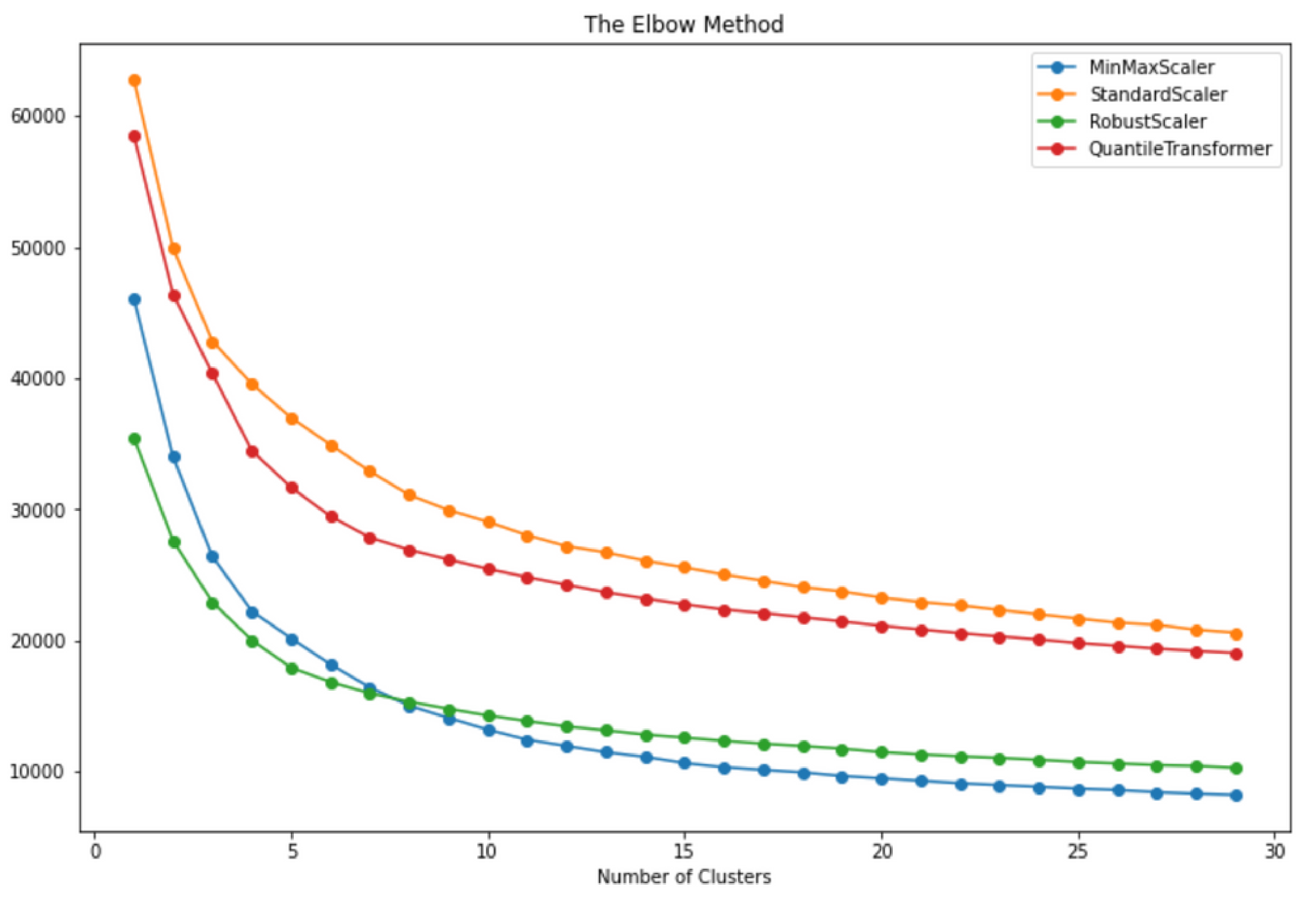
Observing the correlation between the features

	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
danceability	1.000000	0.040491	0.002152	0.358328	-0.088908	0.036121	-0.111151	-0.573800	-0.032534	0.680097	-0.009585	0.215498
energy	0.040491	1.000000	0.029702	0.786860	-0.008461	0.303940	-0.850469	-0.169923	0.170642	0.159101	0.211617	0.162435
key	0.002152	0.029702	1.000000	0.027082	-0.155697	0.027547	-0.024794	-0.016775	0.025193	-0.018109	-0.002370	0.007796
loudness	0.358328	0.786860	0.027082	1.000000	-0.030855	0.233609	-0.697709	-0.471786	0.134788	0.335754	0.213228	0.215875
mode	-0.088908	-0.008461	-0.155697	-0.030855	1.000000	-0.041282	0.028854	-0.003017	-0.009712	0.005966	0.004739	-0.013039
speechiness	0.036121	0.303940	0.027547	0.233609	-0.041282	1.000000	-0.265754	-0.064754	0.081963	-0.011395	0.064255	0.060871
acousticness	-0.111151	-0.850469	-0.024794	-0.697709	0.028854	-0.265754	1.000000	0.194941	-0.103144	-0.130646	-0.187994	-0.163980
instrumentalness	-0.573800	-0.169923	-0.016775	-0.471786	-0.003017	-0.064754	0.194941	1.000000	-0.051664	-0.500584	-0.071945	-0.160122
liveness	-0.032534	0.170642	0.025193	0.134788	-0.009712	0.081963	-0.103144	-0.051664	1.000000	0.007272	0.036370	0.025039
valence	0.680097	0.159101	-0.018109	0.335754	0.005966	-0.011395	-0.130646	-0.500584	0.007272	1.000000	0.098783	0.189048
tempo	-0.009585	0.211617	-0.002370	0.213228	0.004739	0.064255	-0.187994	-0.071945	0.036370	0.098783	1.000000	0.024075
time_signature	0.215498	0.162435	0.007796	0.215875	-0.013039	0.060871	-0.163980	-0.160122	0.025039	0.189048	0.024075	1.000000

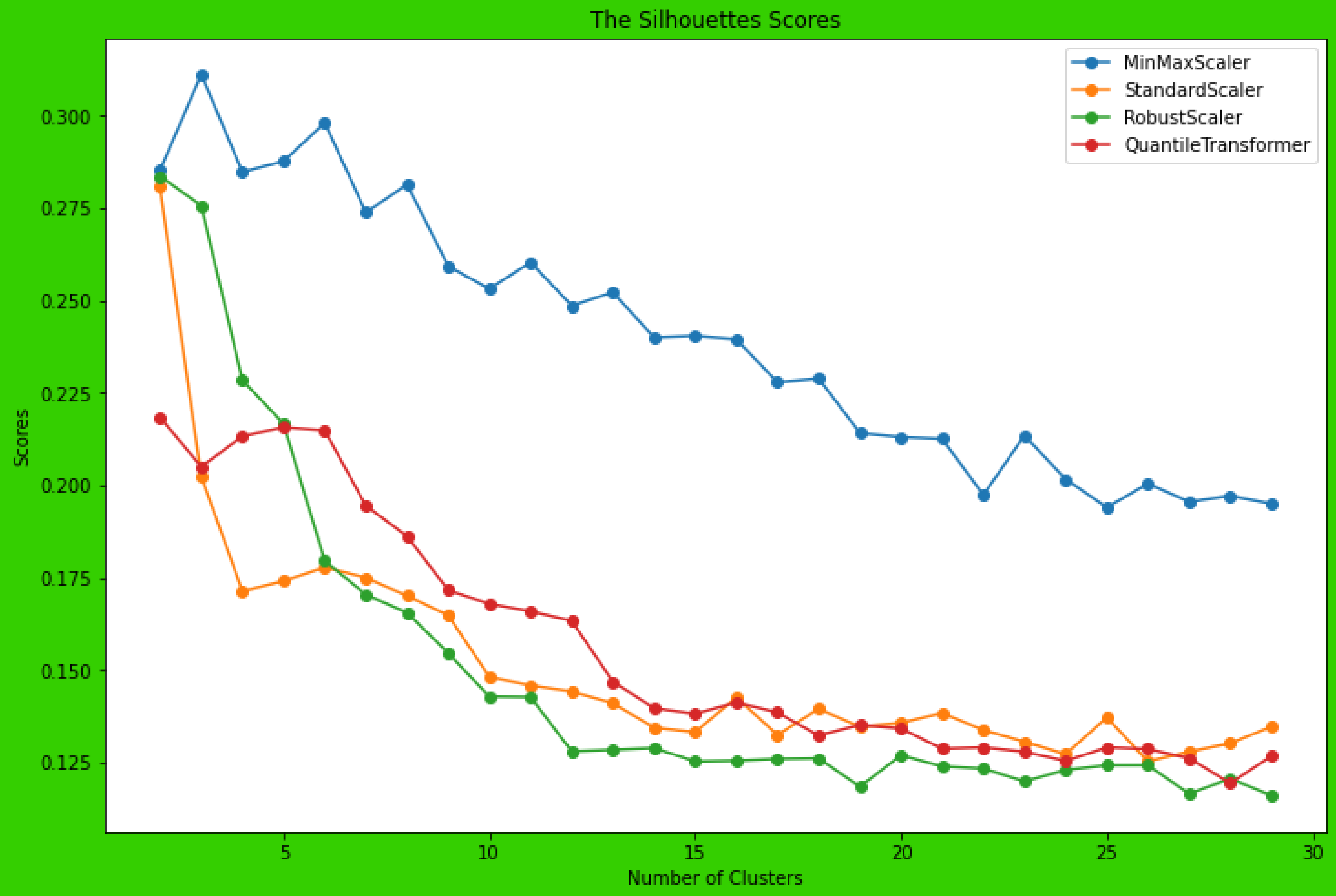
Finding the best cluster number

Elbow method: We use different scalers to find the best suitable cluster number as show in figure

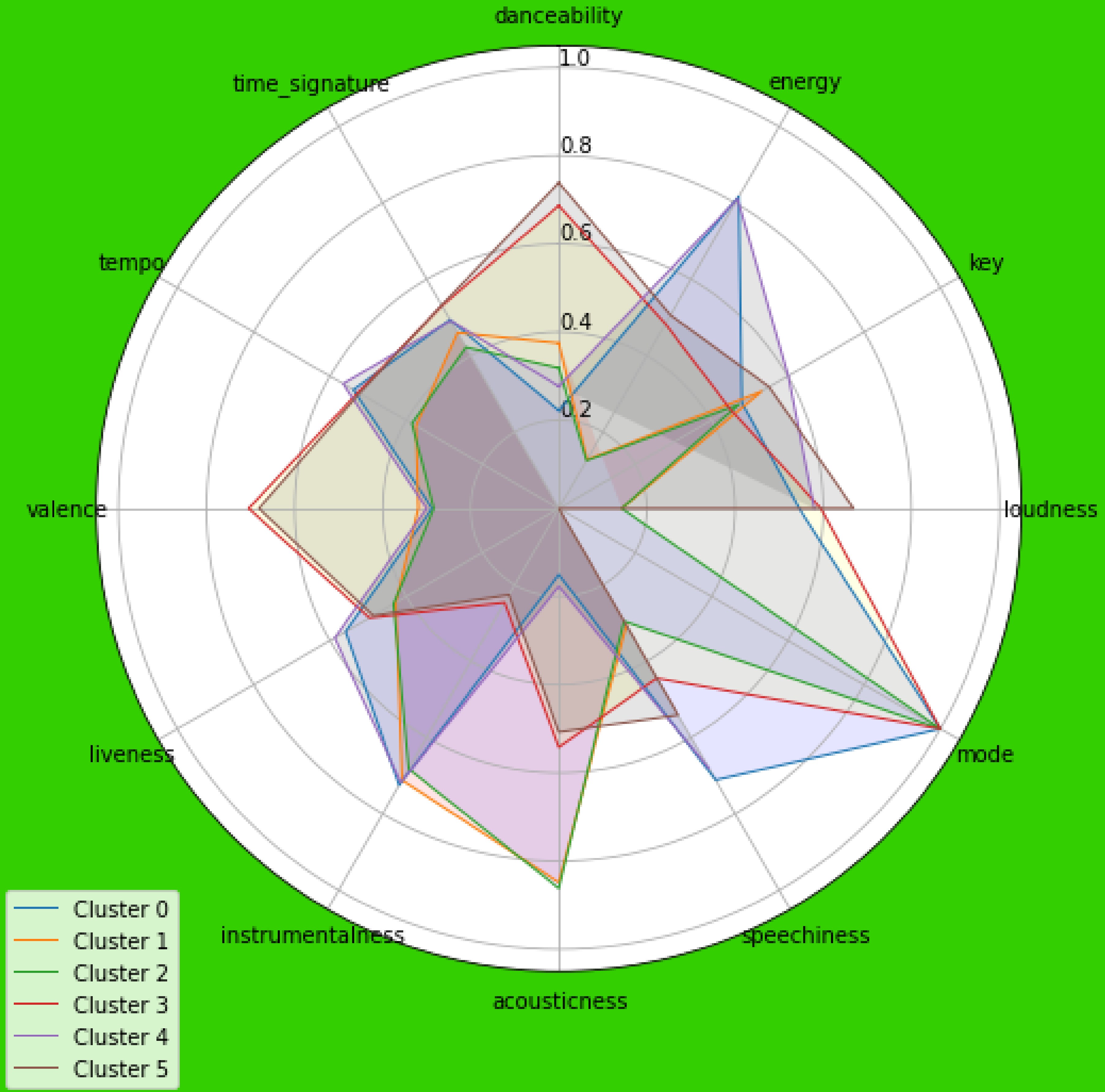
Silhouette method: We use different scalers to find the best suitable cluster number as show in figure



The MinMax and Quantile were multiplied by 10 to scale it with the others methods



Cluster exploration



A distribution of features within the 6 clusters

Creating the Playlists

- **Calculated the euclidian and Manhattan distance**
- **Then we choose any number of songs closer to the centroid**

name	artist	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature	cluster	eucl_dist	manh_dist
Game Of Pricks	Guided By Voices	0.331	0.866	9	-5.525	1	0.0456	0.000414	0.024300	0.1300	0.347	139.086	4	1	0.610374	1.664310
Faithless - B-Side	City and Colour	0.367	0.973	11	-2.191	1	0.0985	0.000428	0.000015	0.3700	0.617	170.043	4	1	1.041156	2.835442
San Francisco	Foxygen	0.341	0.552	8	-10.503	1	0.0423	0.000080	0.616000	0.0591	0.486	121.361	4	1	0.935938	2.324685
The Stars Keep On Calling My Name	Mac DeMarco	0.464	0.815	10	-6.371	1	0.0368	0.008710	0.201000	0.1620	0.467	161.845	4	1	0.859137	2.310139
Red Eyes	The War On Drugs	0.419	0.880	5	-6.019	1	0.0301	0.029500	0.876000	0.1350	0.521	150.792	4	1	0.866719	2.088307
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
Man On Fire	Idahams	0.772	0.687	2	-7.398	0	0.1640	0.044500	0.000006	0.0480	0.794	96.034	4	6	0.767915	2.101065
If He Did It Before....Same God - Live	Tye Tribbett	0.608	0.790	3	-5.413	0	0.1720	0.036000	0.000004	0.0566	0.680	159.869	4	6	0.743332	1.896114
Blessed & Highly Favored - Live	The Clark Sisters	0.502	0.759	5	-4.065	0	0.1260	0.311000	0.000000	0.9850	0.382	102.302	4	6	0.804118	2.222771
You Brought The Sunshine -	The Clark	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

We connect to the Spotify API and used the euclidian distance to select the 20 songs closer to the centroid

Naming the Playlists

- **Playlist 1 - Black Metal - Headbanger Mode On / Neighbors love it Pt.1**
- **Playlist 2 - Jazz/Classic - To chill and code**
- **Playlist 3 - Jazz/Classic -To chill (without to code)**
- **Playlist 4 - Pop Mix Songs - Dancing in Summer**
- **Playlist 5 -Eddie Munson Metal songs**
- **Playlist 6 -R&B - RoadTrip**

Conclusion

Spotify's audio features are able to identify "similar songs", as defined by humanly detectable criteria.

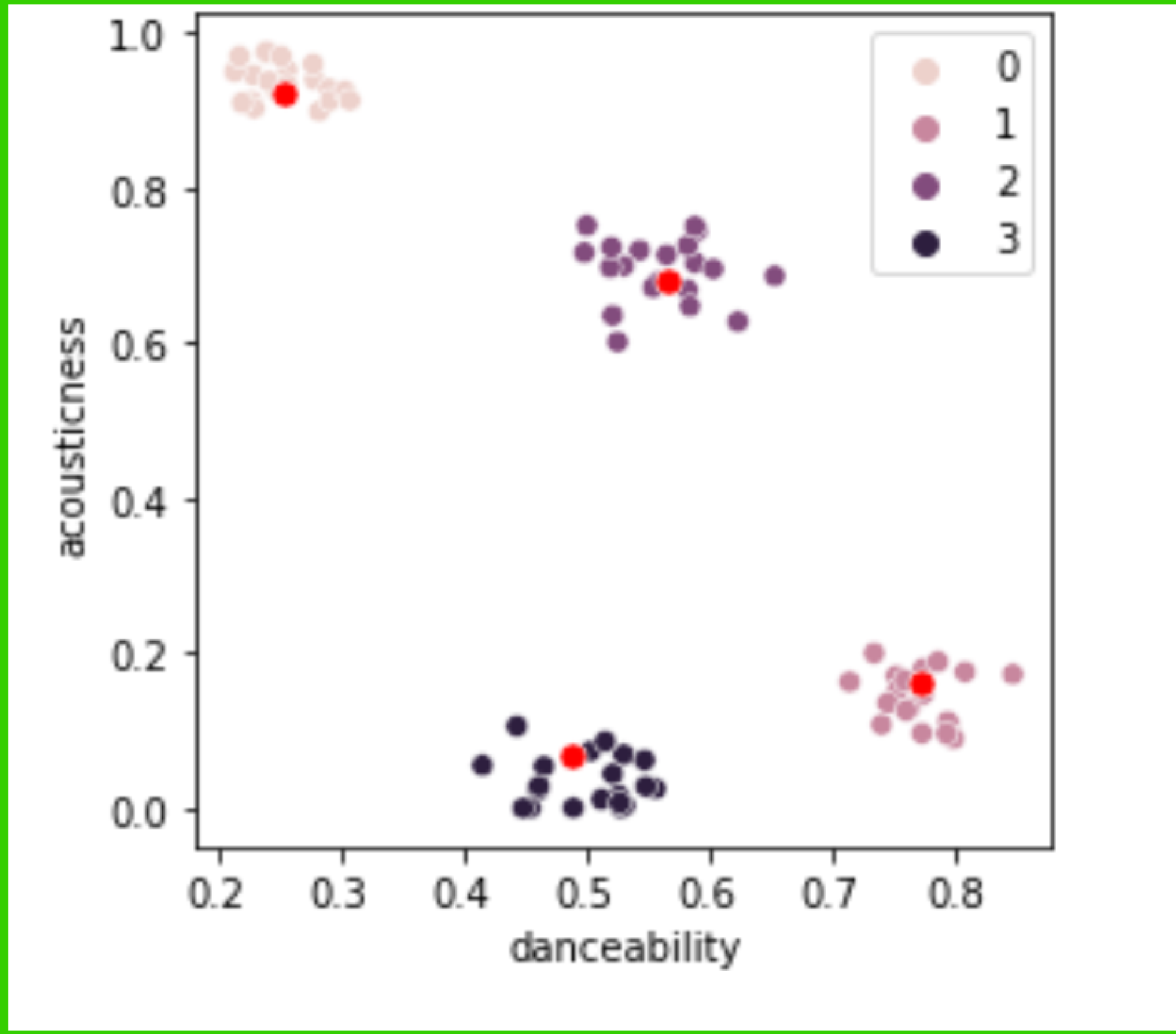
K-Means is a good method to create playlists, but … only to a certain extent.

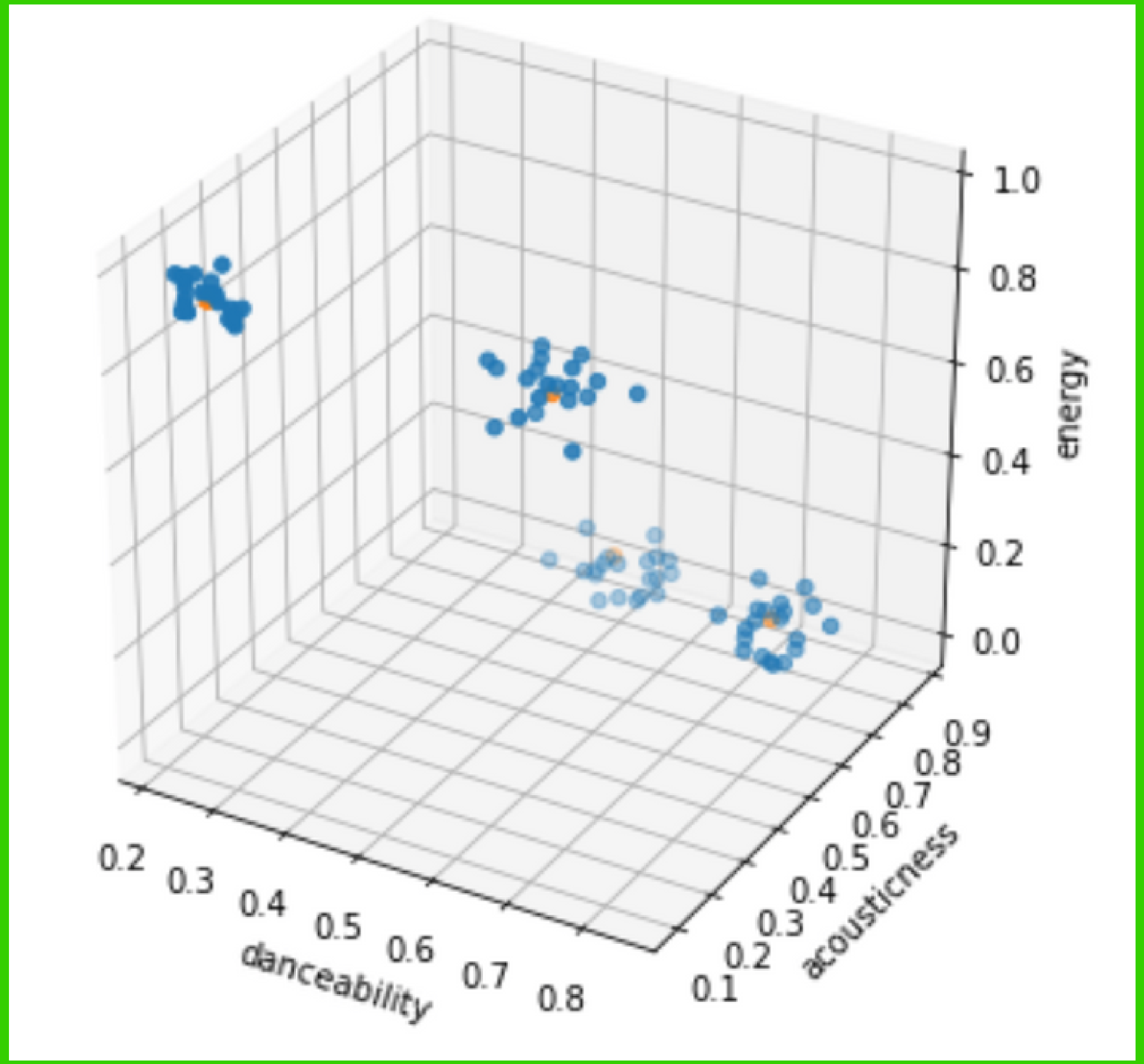
We recommend a human supervision to confirm if the playlist actually make sense

https://open.spotify.com/playlist/0ZKP4fcCUlvlKgEZTK2sZR?si=1d48e3e26fe4484b

Function to choose the songs

songs, cl_pos = **selective_songs**(n_cluster, moosic, ['danceability', 'acousticness', 'energy'], 30, 'euclidean')





One can choose different **features** and **number of clusters** to run the function, "selective_songs" which provides the best songs.