Análisis de Tendencias y Clusterización de Canciones

Cesar Moreno, Lucas Calahumana y Gonzalo Saravia

Llibreria Spotipy

Para el analisis, utilizamos la libreria spotipy, que ofrece 9 índices de ”danceability","energy", "loudness", "speechiness", "instrumentalness", "liveness", "tempo", "time_signature",”valence”.

Así como un índice de popularidad porcentual.

album	artist	popularity	danceability	acousticness	energy	instrumentalness	liveness	loudness	speechiness	valence	tempo	time_signature
Bad Habits	Ed Sheeran	87	0.808	0.0469	0.897	0.000031	0.3640	-3.712	0.0348	0.591	126.026	4
Break My Heart	Dua Lipa	74	0.730	0.1670	0.729	0.000001	0.3490	-3.434	0.0886	0.467	113.012	4
Save Your Tears (Remix)	The Weeknd	95	0.650	0.0215	0.825	0.000024	0.0936	-4.645	0.0325	0.593	118.091	4
MONTERO (Call Me By Your Name)	Lil Nas X	100	0.610	0.2970	0.508	0.000000	0.3840	-6.682	0.1520	0.758	178.818	4

- Danceability: Describe qué tan adecuada es una pista para bailar basándose en el tempo, la estabilidad del ritmo, la fuerza del ritmo y la regularidad general.
- Energy: Representa una medida perceptual de intensidad y actividad.
- Instrumentalness/Speechiness: Predice si una pista no contiene voces. Por ejemplo los sonidos “Ooh” y “aah” se tratan como instrumentales en este contexto.
- Liveness: Detecta la presencia de una audiencia en la grabación.
- Loudness: El volumen general de una pista en decibelios (dB).
- Valence: Describe la positividad musical que transmite una pista.
- Tempo: El tempo global estimado de una pista en pulsaciones por minuto
- Time Signature: Una signatura de tiempo total estimada de una pista.

PRIMER PASO: entendiendo tendencias

Lo primero que hicimos fue limitar nuestro scope a un área que consideramos más homogénea, y por tanto predecible. Tal área es la década de los 80, género pop. Por otro lado elegimos también el género pop en la actualidad, en los cuales clasificamos las canciones en hits y no hits en base la popularidad de los temas.

Qué se buscó hacer:

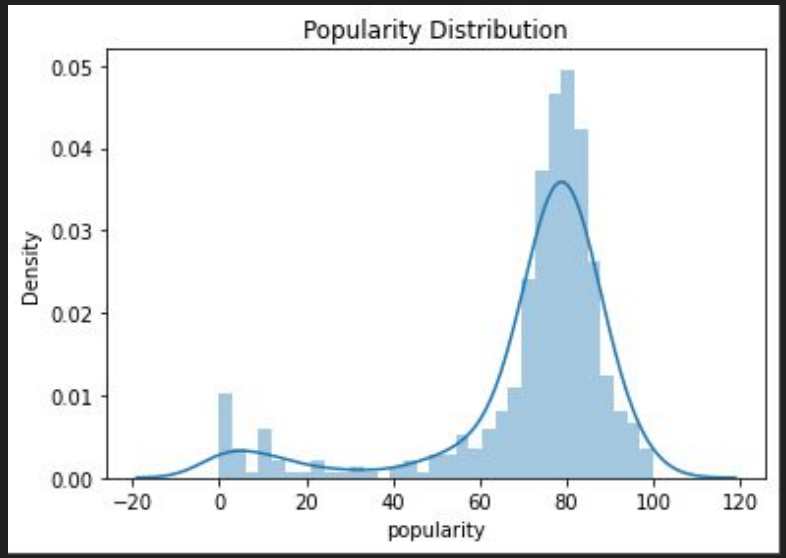
- Aplicar etiquetas binarias a las canciones del dataset de pop actual.
- Entrenar diferentes modelos para hacer un predictor de tendencias:
- Regresión Logistica.
- KNN.
- SVM.

Primeros insights

Para tomar un poco de perspectiva, hicimos una regresión logística sobre los datos. Como sabemos, la regresión logística es una manera simple y con interpretacion probabilistica de clasificar datos. Los resultados nos pincharon la idea de poner una disquera. Con un Accuracy de 0,62 y AUC de 0,57 nos surge la hipótesis de que los hits podrían surgir más por la fama del compositor o por ser anómala, es decir, tema no pertenece a una categoría definida.

Ahora con KNN y SVM

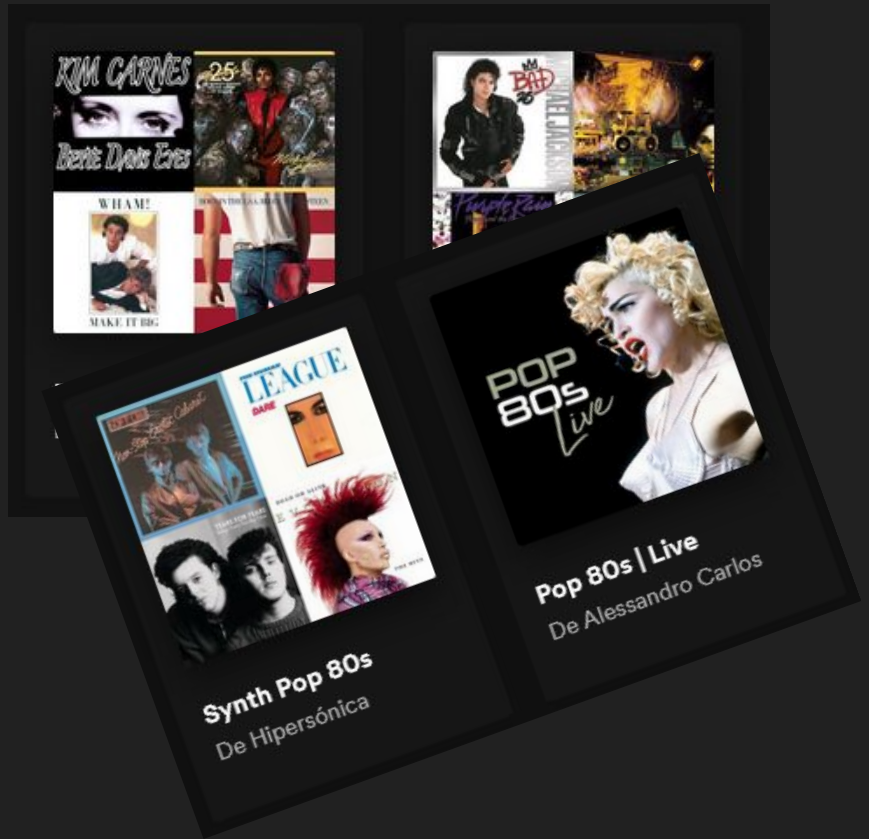
	Model	Accuracy	AUC
0	SVM	0.773723	0.559221
1	K-NNeighbors	0.737226	0.586602



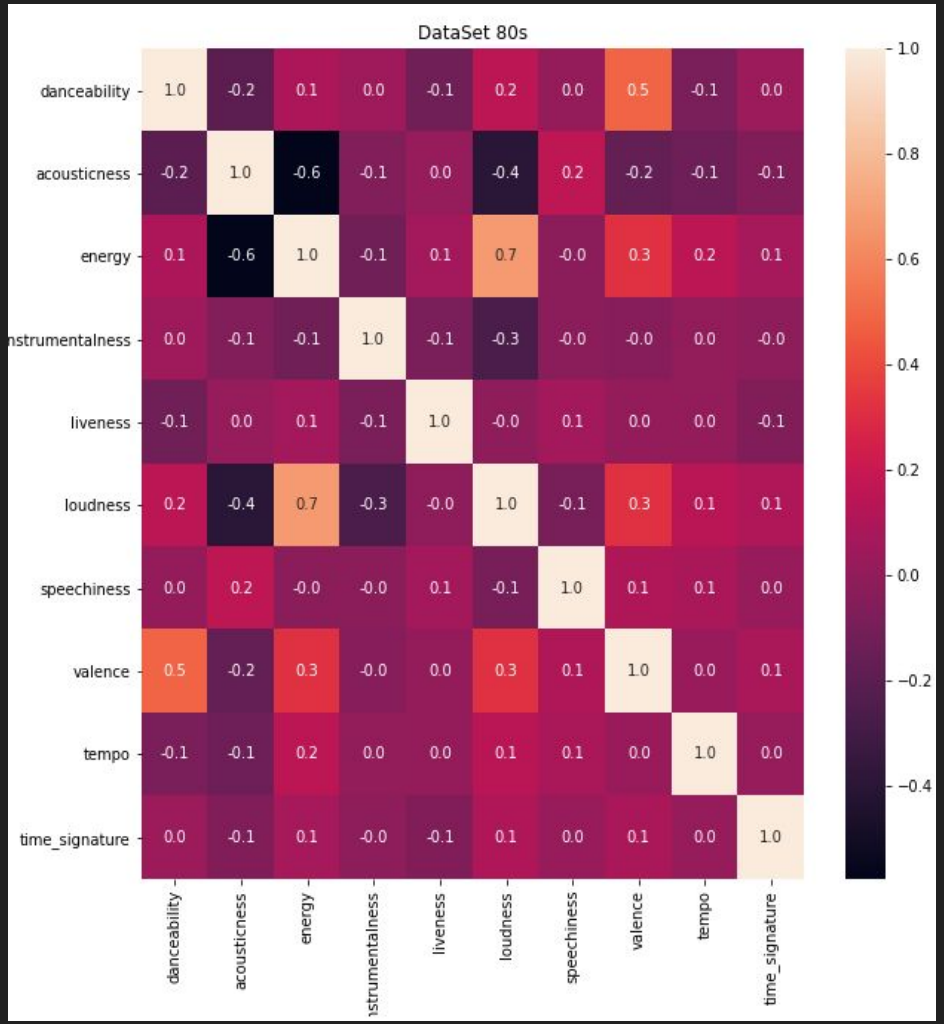
Algunos comentarios sobre los resultados:

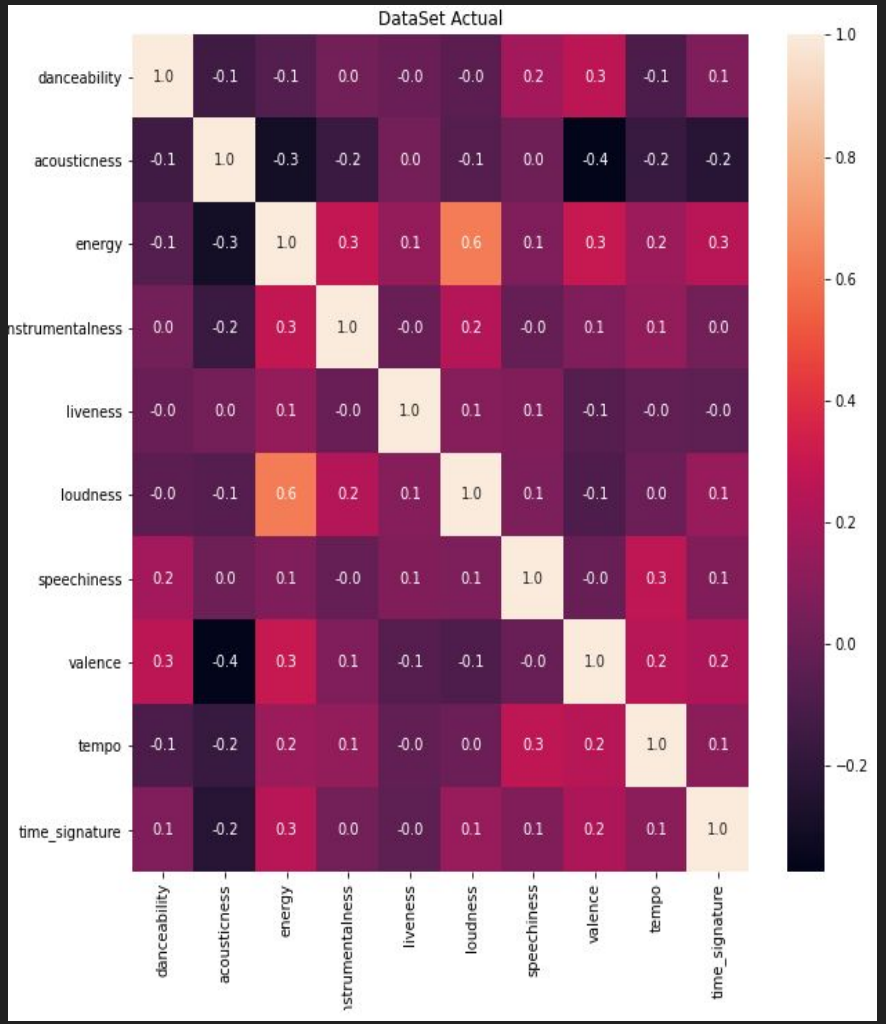
- Influencia del tamaño del dataset.
- Desbalance de las etiquetas. Puede afectar a los clasificadores.
- Cierto grado de homogeneidad en los valores de los features.
- Sensibilidad a disparidades de escala de los features.

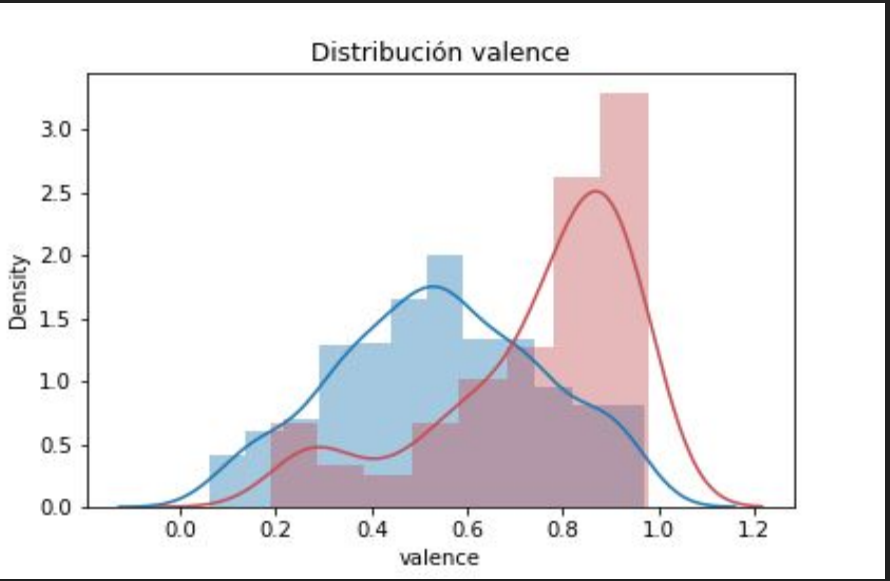
DataSet Pop de los 80’s:

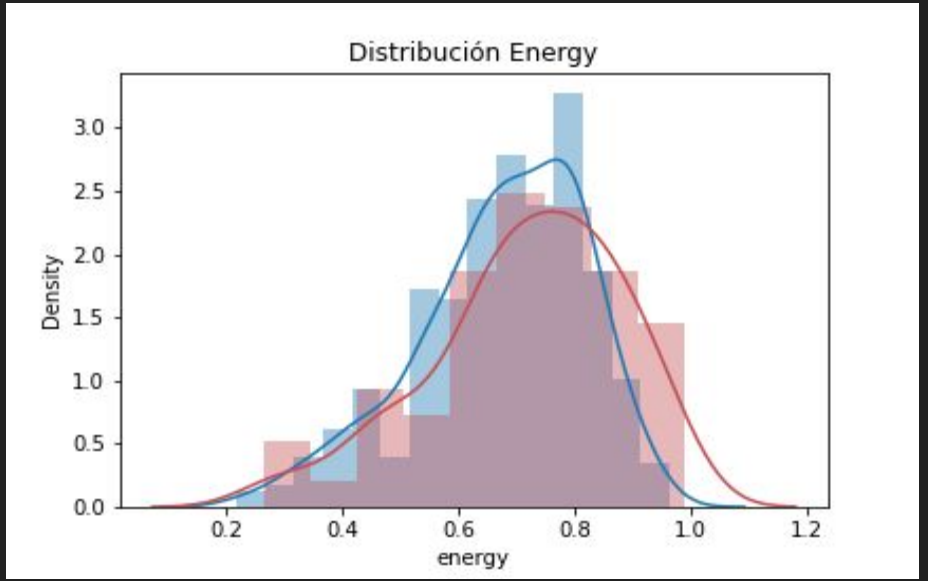


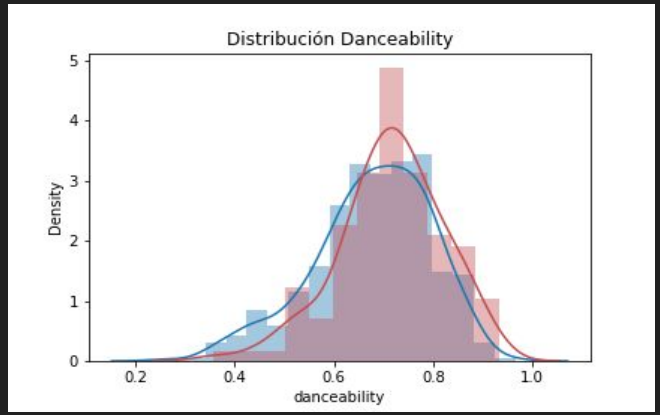
- Playlist de spotify.
- Los discos más vendidos de la década.

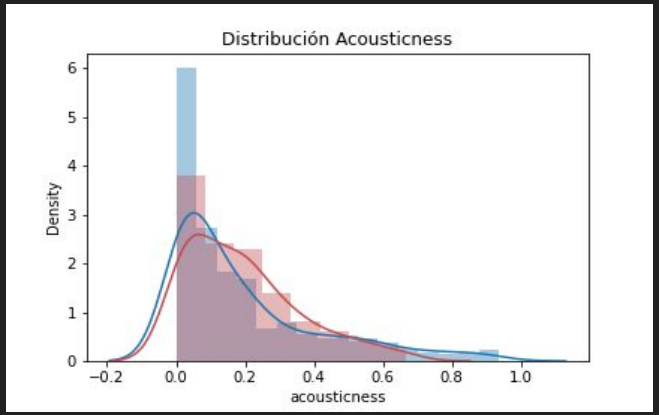

DataSet 80s
danceability 1.0 -0.2 0.1 0.0 -0.1 0.2 0.0 0.5 -0.1 0.0
acousticness -0.2 1.0 -0.6 -0.1 0.0 -0.4 0.2 -0.2 -0.1 -0.1
energy 0.1 -0.6 1.0 -0.1 0.1 0.7 -0.0 0.3 0.2 0.1
nstrumentalness 0.0 -0.1 -0.1 1.0 -0.1 -0.3 -0.0 -0.0 0.0 -0.0
liveness -0.1 0.0 0.1 -0.1 1.0 -0.0 0.1 0.0 0.0 -0.1
loudness 0.2 -0.4 0.7 -0.3 -0.0 1.0 -0.1 0.3 0.1 0.1
speechiness 0.0 0.2 -0.0 -0.0 0.1 -0.1 1.0 0.1 0.1 0.0
valence 0.5 -0.2 0.3 -0.0 0.0 0.3 0.1 1.0 0.0 0.1
tempo -0.1 -0.1 0.2 0.0 0.0 0.1 0.1 0.0 1.0 0.0
time_signature 0.0 -0.1 0.1 -0.0 -0.1 0.1 0.0 0.1 0.0 1.0
danceability
acousticness
energy
nstrumentalness
liveness
loudness
speechiness
valence
tempo
time_signature
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4


DataSet Actual
danceability 1.0 -0.1 -0.1 0.0 -0.0 -0.0 0.2 0.3 -0.1 0.1
acousticness -0.1 1.0 -0.3 -0.2 0.0 -0.1 0.0 -0.4 -0.2 -0.2
energy -0.1 -0.3 1.0 0.3 0.1 0.6 0.1 0.3 0.2 0.3
nstrumentalness 0.0 -0.2 0.3 1.0 -0.0 0.2 -0.0 0.1 0.1 0.0
liveness -0.0 0.0 0.1 -0.0 1.0 0.1 0.1 -0.1 -0.0 -0.0
loudness -0.0 -0.1 0.6 0.2 0.1 1.0 0.1 -0.1 0.0 0.1
speechiness 0.2 0.0 0.1 -0.0 0.1 0.1 1.0 -0.0 0.3 0.1
valence 0.3 -0.4 0.3 0.1 -0.1 -0.1 -0.0 1.0 0.2 0.2
tempo -0.1 -0.2 0.2 0.1 -0.0 0.0 0.3 0.2 1.0 0.1
time_signature 0.1 -0.2 0.3 0.0 -0.0 0.1 0.1 0.2 0.1 1.0
danceability
acousticness
energy
nstrumentalness
liveness
loudness
speechiness
valence
tempo
time_signature
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Distribución valence
Density
valence
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Distribución Energy
Density
energy
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.2
0.4
0.6
0.8
1.0
1.2


Distribución Danceability
Density
danceability
0
1
2
3
4
5
0.2
0.4
0.6
0.8
1.0


Distribución Acousticness
Density
acousticness
0
1
2
3
4
5
6
−0.2
0.0
0.2
0.4
0.6
0.8
1.0

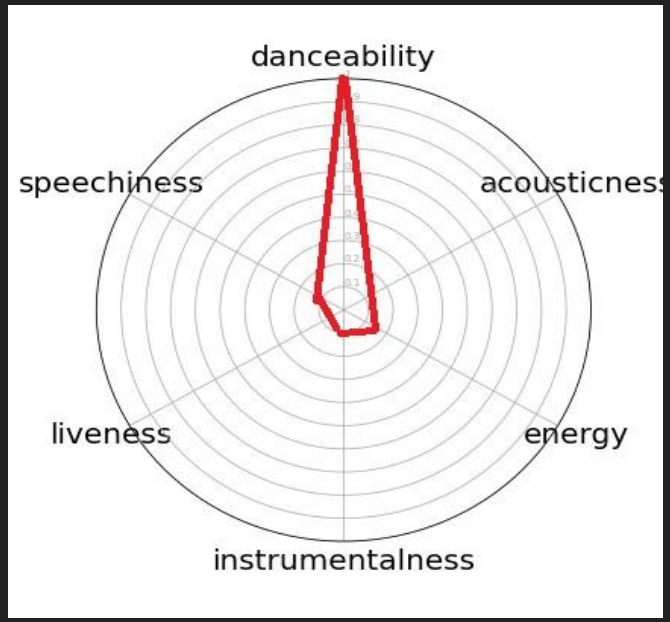
Segunda parte : Clusterización de canciones

Utilizando los dos set de datos presentados previamente y habiendo visto las similitudes que estos dos guardan.

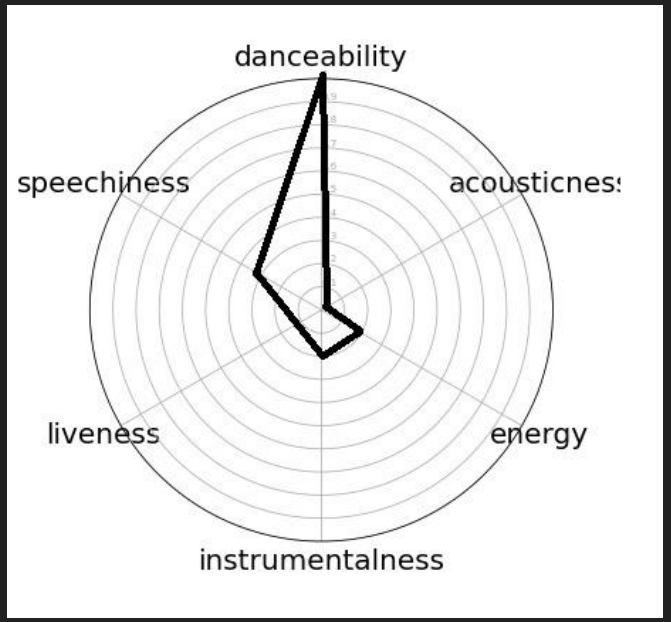
Que se buscó hacer

- Partir de dos conjuntos iniciales
- Hacer clusterización sobre estos dos
- Caracterizar cada Cluster
- Emparejar cada uno de estos de acuerdo a caracteristicas en comun

Es decir:



Cluster de muestra 1 (A)

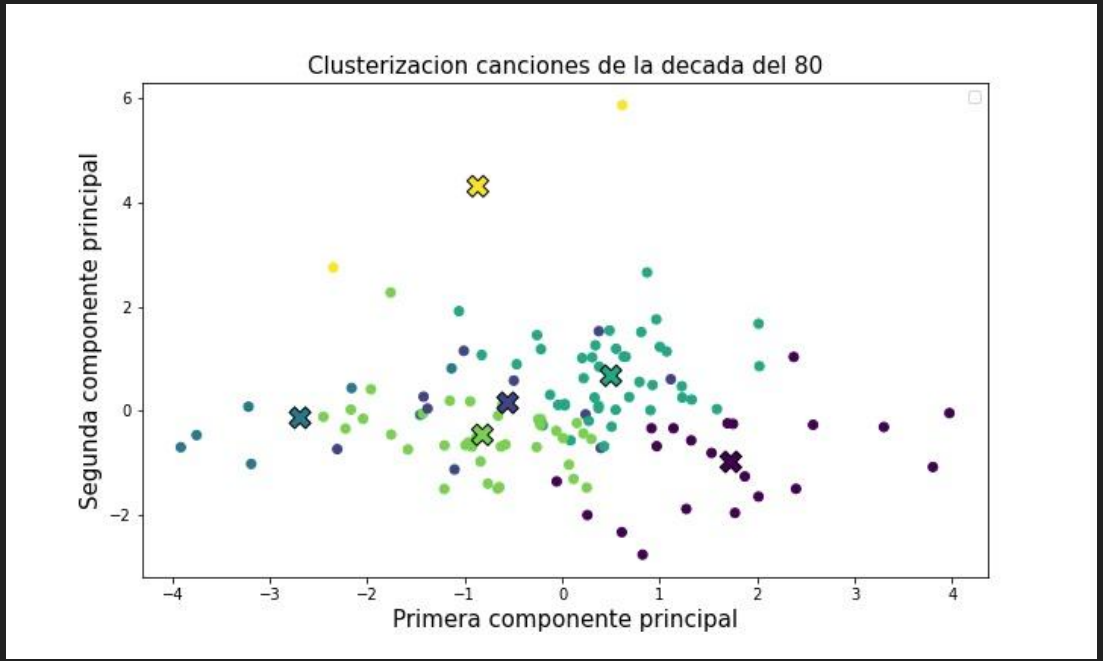


Cluster de muestra 2 (B)

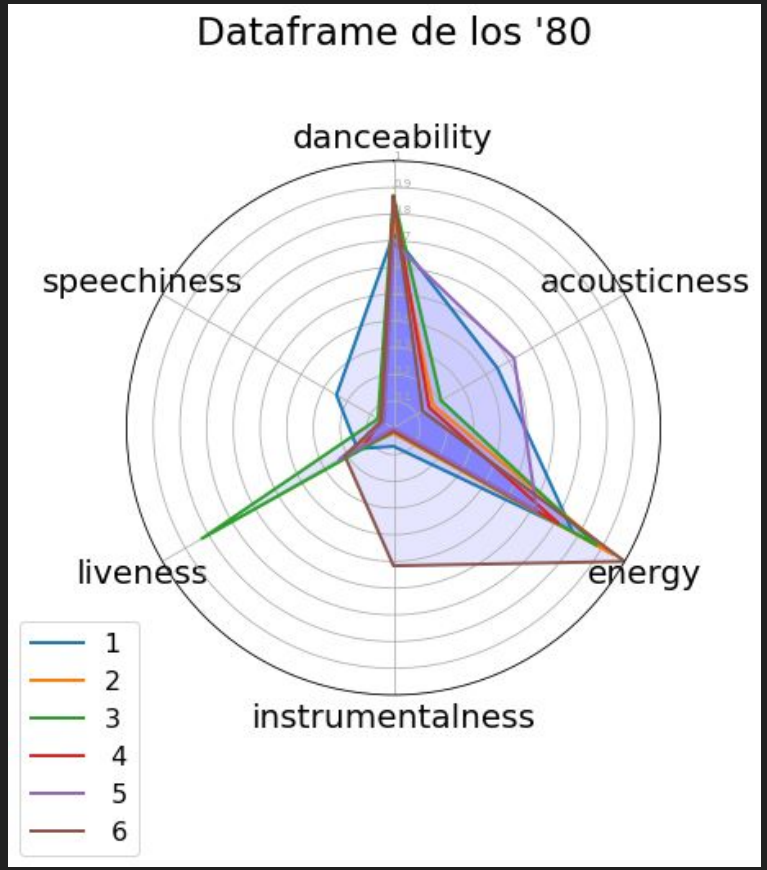
Empezando por el set de canciones de los ‘80

Dataset de música de los ochenta

Visualización en el espacio de las dos primeras componentes de PCA

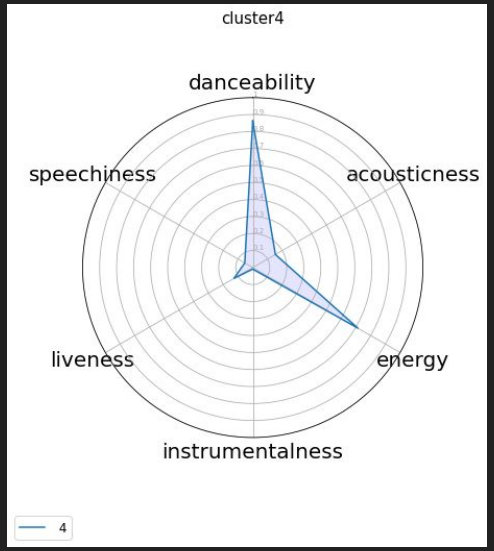


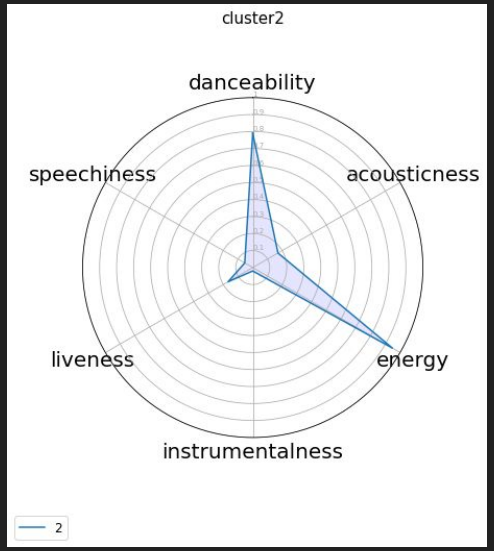
Vemos los clusters



Clusters

Vemos que de por si hay clusters que se separan bastante bien, pero hay ciertos clusters “conflictivos” que no son tan faciles de caracterizar, tales como los cluster 4 y 2

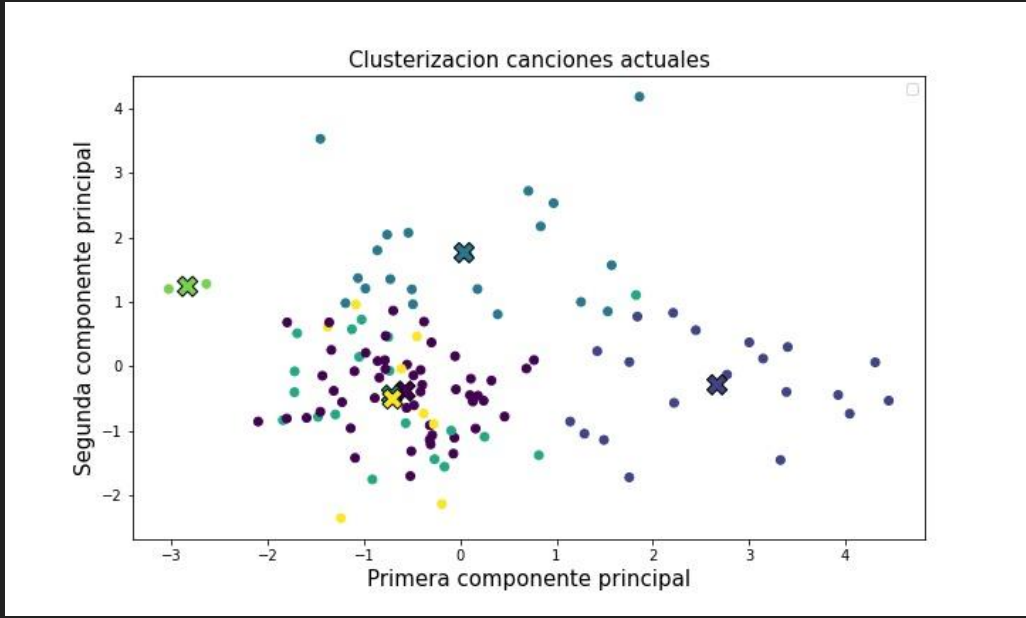




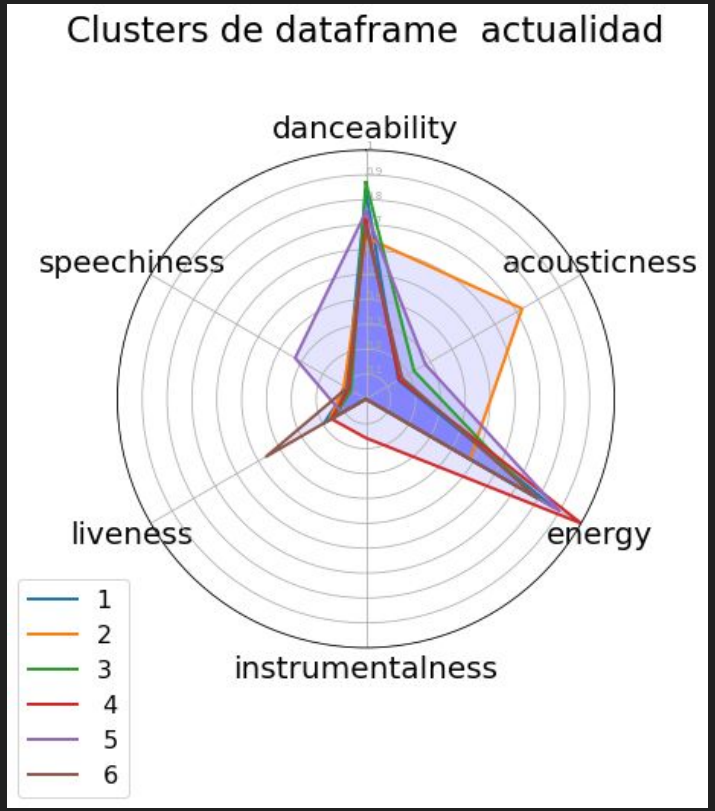
Yendo al set de canciones actuales

Clusterización del set de canciones actuales

Representamos los datos en el espacio de las primeras componentes de PCA

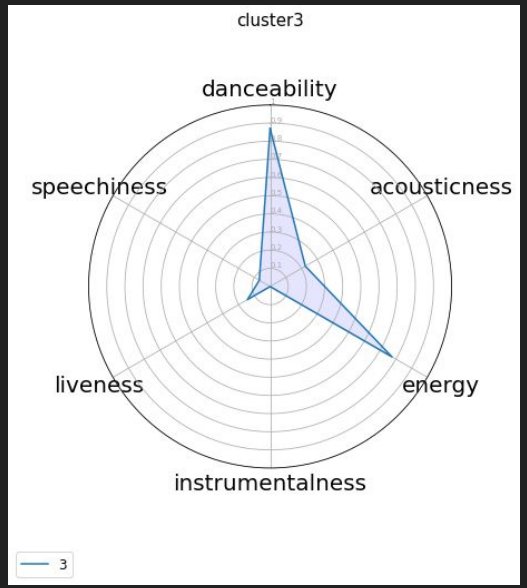


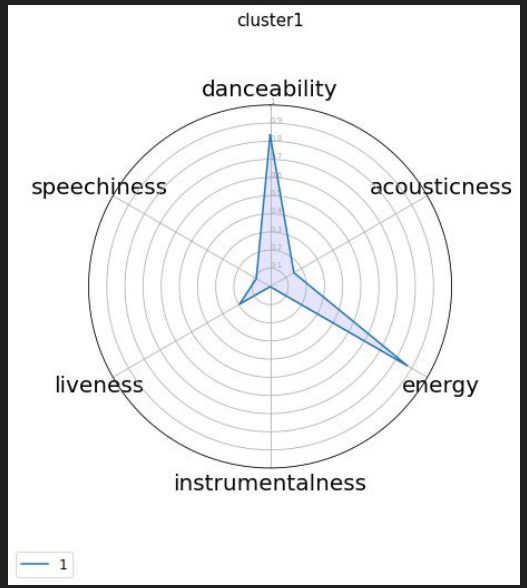
Vemos los clusters en este caso :



Desmenuzando los clusters

Vemos que en este caso hay separaciones marcadas, pero aun así tenemos que todos los clusters tienen una forma levemente similar. Dónde podemos señalar a los cluster 3 y 1 como más “conflictivos”

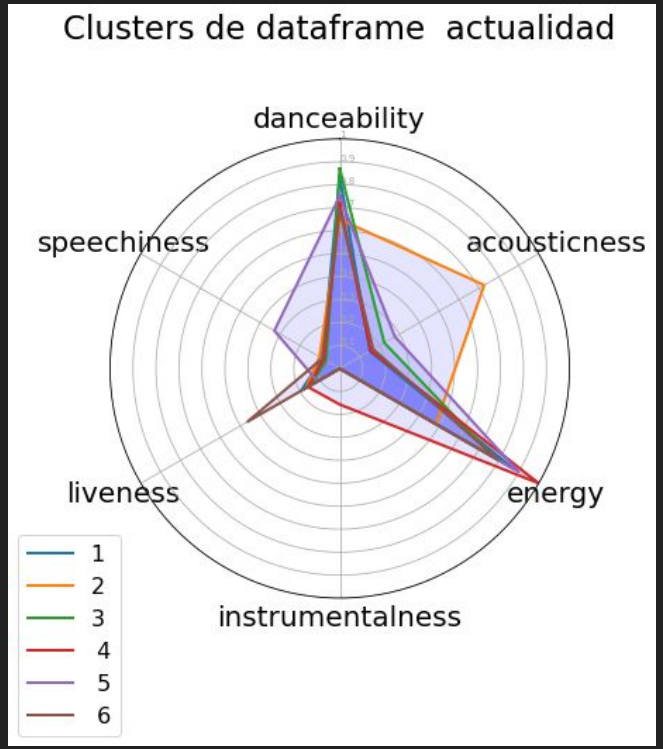


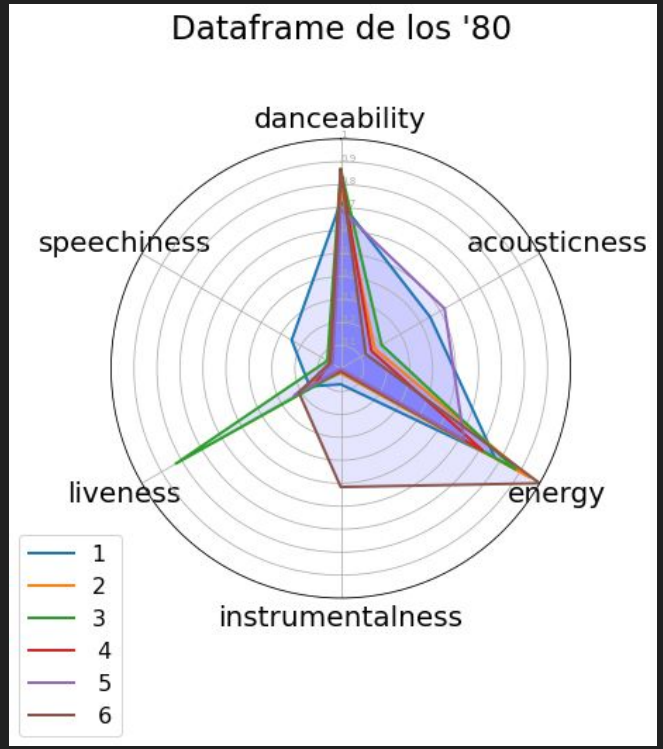


Con esto, podemos empezar a agrupar...

Emparejamiento de clusters

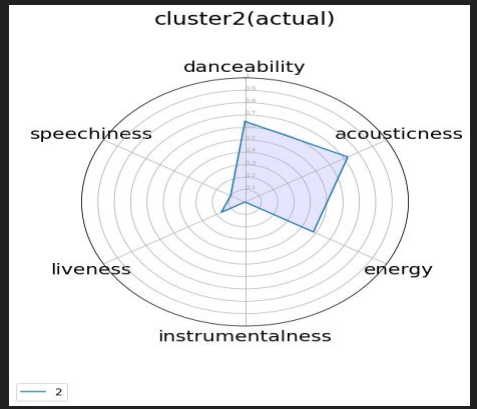
Tenemos que los dos grupos tienen los siguientes clusters:



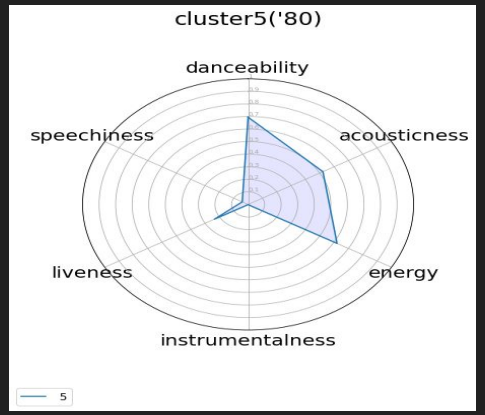


Emparejamiento 1

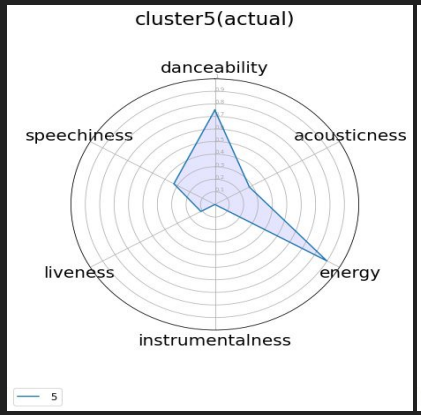
Vemos que hay agrupaciones más fáciles que otras, por ejemplo:



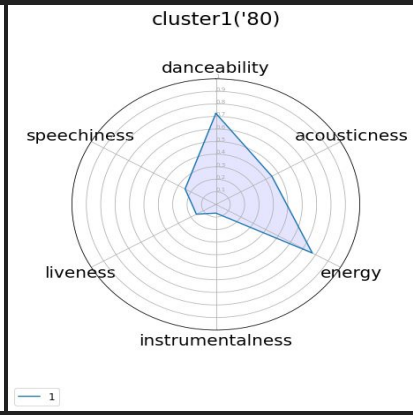
(A)



(B)



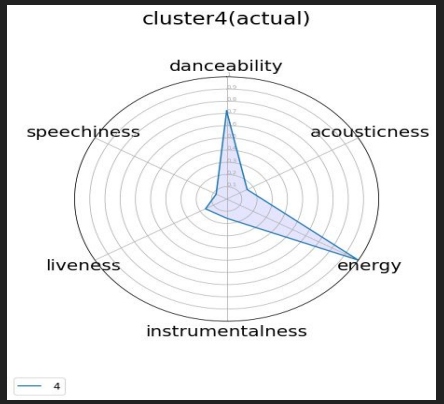
(C)



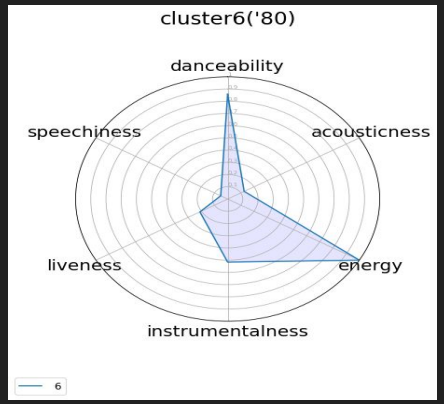
(D)

Emparejamiento 2

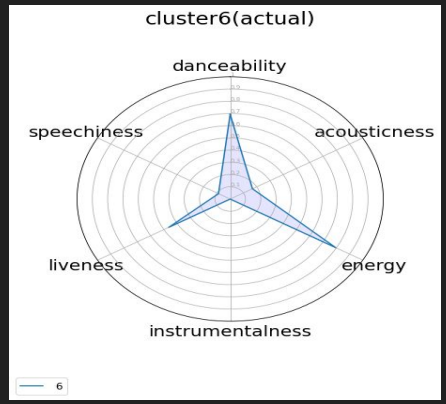
Algunas más desafiantes. Se agrupan los clusters con valor máximo en alguna categoría pero estos máximos son de distinta magnitud en cada uno de los cluster



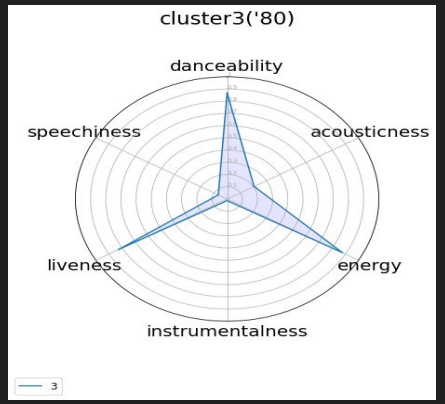
(A)



(B)



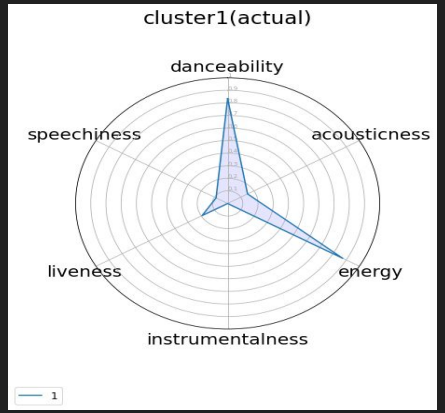
(C)



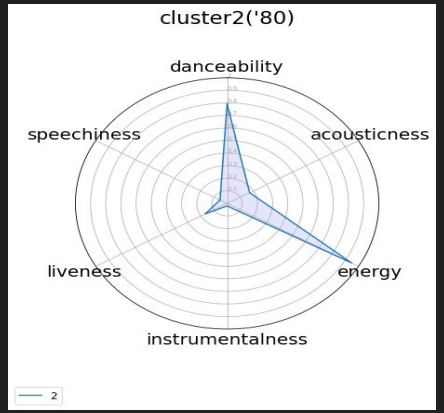
(D)

Emparejamiento 3

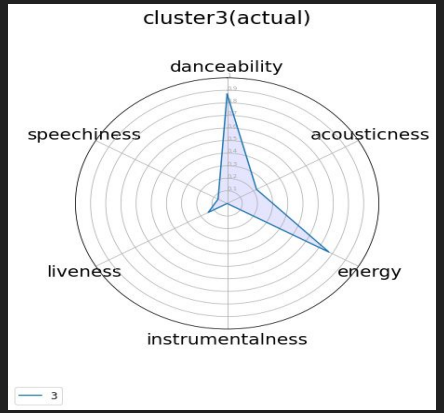
Y los conflictivos … o no tanto.



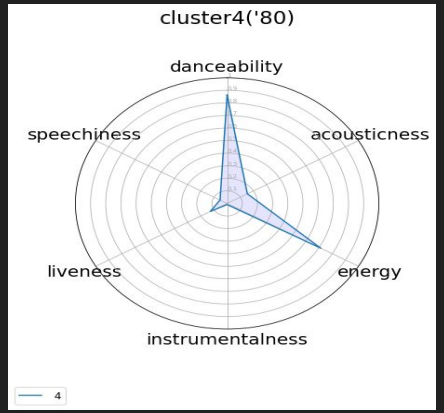
(A)



(B)



(C)



(D)

Propuesta interactiva

Conclusiones de este estudio:

- Depende mucho de la “calificación” previa provista por spotify y la subjetividad
- Es sensible a outliers como se puede ver en los gráficos
- Al ser muestras bastante homogéneas en ciertos casos no hay diferencias muy marcadas entre clusters

Apendice de Playlists

Acusticness alta

Cluster 2 , musica actual

- https://open.spotify.com/playlist/0fKa1Pd0KuUYEhisJE11K1?si=c801127ba0dd49a1

*Cluster 5 musica de los ‘80

- https://open.spotify.com/playlist/6IeGFiHKiZ2ElkJSEDgmNH?si=05aa4abd95bb42c7

Spechness alta

Musica actual:

- https://open.spotify.com/playlist/0VPle2JOoHn38YyfwLarQS?si=a78ea388b39f4bf1

Ochentosa

- https://open.spotify.com/playlist/6yCMOk6jtcW41fKqRR2ynY?si=a174fb231de942c8

Instrumentalness alto

Musica actual:

- https://open.spotify.com/playlist/4qobaxgyOA88sHSDTyPUDi?si=c23dd0475d7e4c0c

‘80:

- https://open.spotify.com/playlist/4I5VzfdN0SS8GMUjHuFGjw?si=12dcf1322311437e

Liveness alto

actual:

- https://open.spotify.com/playlist/39GBkJ53INXNdBO4Uiefxs?si=d80253a48b6f4e43

‘80:

- https://open.spotify.com/playlist/1nDwomDqv4FdiDhJLIgVKn?si=a42ccf2324f34f27

Conflictivas 1 (Energy y danceability altos)

Musica actual:

- https://open.spotify.com/playlist/5ZxM8jrPhT4x64wmdF0D9S?si=053eb57575434b20

‘80:

- https://open.spotify.com/playlist/6AZKTE4jLcUolS5H2HFDBz?si=bfbe1648e1ef477f

Conflictivas 2 (Energy y danceability altos)

Actual:

- https://open.spotify.com/playlist/7MPGLfO6V8dLh5w9EC9RaP?si=9beeaadbde76497c

‘80:

- https://open.spotify.com/playlist/69FPmX5WWAdYfXKKNvePE2?si=465246de4aec4752

¡Muchas gracias!