

# WBS School Project

---

Alexandra Cange



# Project overview

This project was made in my bootcamp time in November 2022. This week project was oriented to teach us what was unsupervised Machine Learning.

The story is that we work for a music company that want to create playlists for their consumers. As Data Scientist, we should advice the CEO if the KMean algorithm is a good one to create playlists.



# Project Content

*Is KMean algorithm good for creating playlists?*



## DECISION ON CLUSTERS

Scaling the datas  
Number of Centroid



## CLUSTERING METHODS

All Parameters - Full Clustering  
All Parameters - Subclustering  
Few Parameters - Full Clustering  
Few Parameters - Subclustering



## CONCLUSION

Best Method  
Recommendation

# Decision on Clusters

## Requirements

Requirements were to create playlists from 50 to 250 songs in each.

We got 3 data set: 10 songs DF, 1500 songs DF and 5000 songs DF.

I worked with *Heatmap* on 10 songs to decide which scaling was the best for the rest of the project.

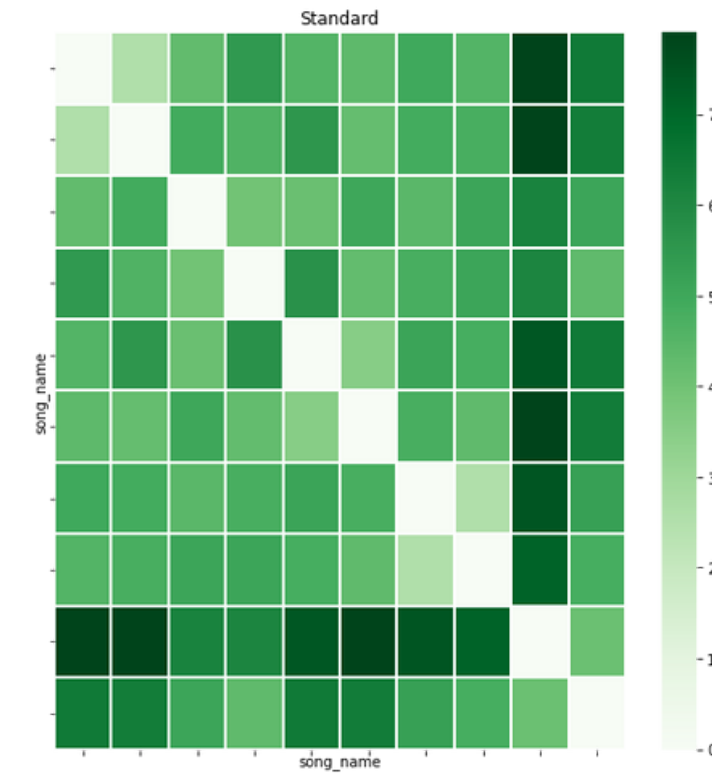
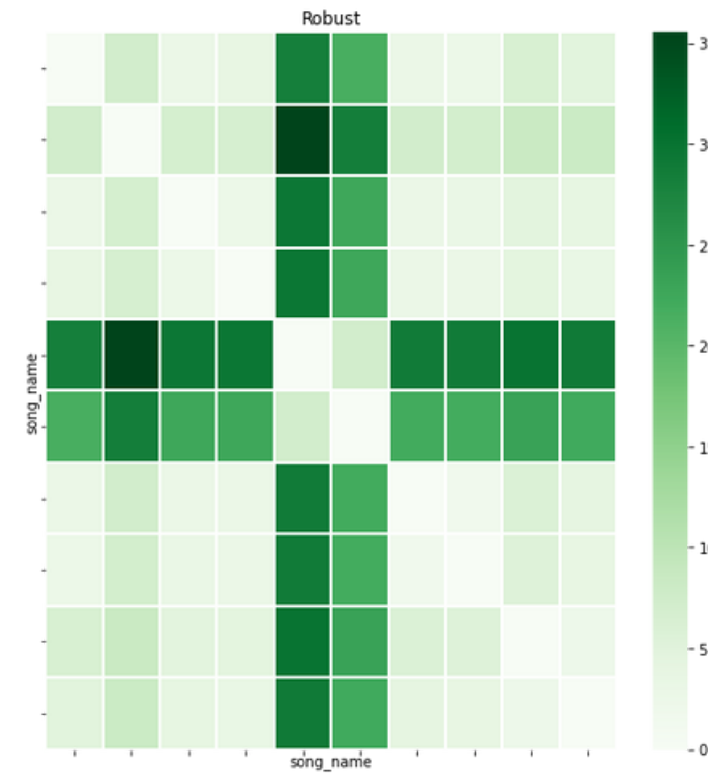
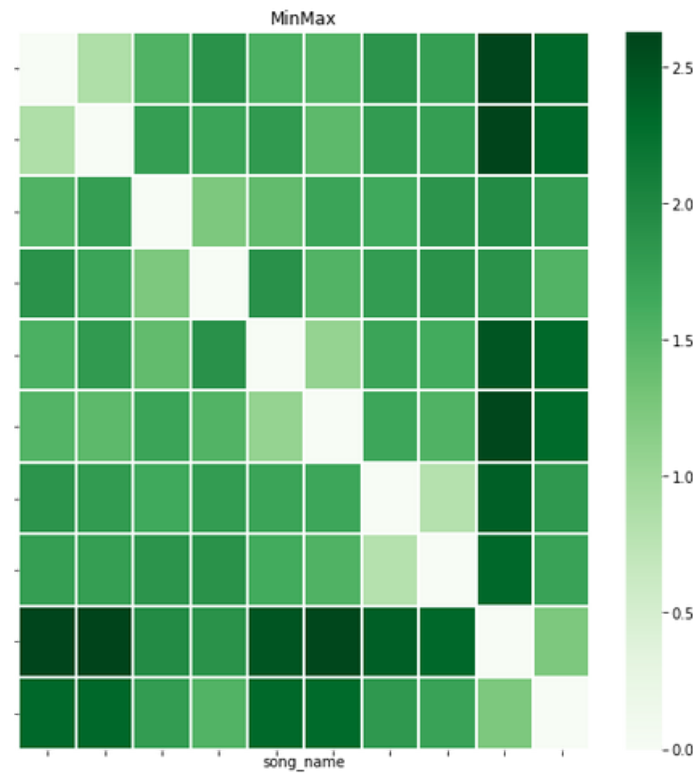
## Inertia

Inertia score is a way to see (if plotting) when the amount of KMean points makes the most of sense. This is of course not an exact science, but it helps to guide. Also call : Elbow Method.

## Silhouette

Silhouette score show how good a data point is in its cluster. The Score vary between -1 to 1 where 1 is the best, and -1 means the data point should be in another cluster.

# Scaling Datas

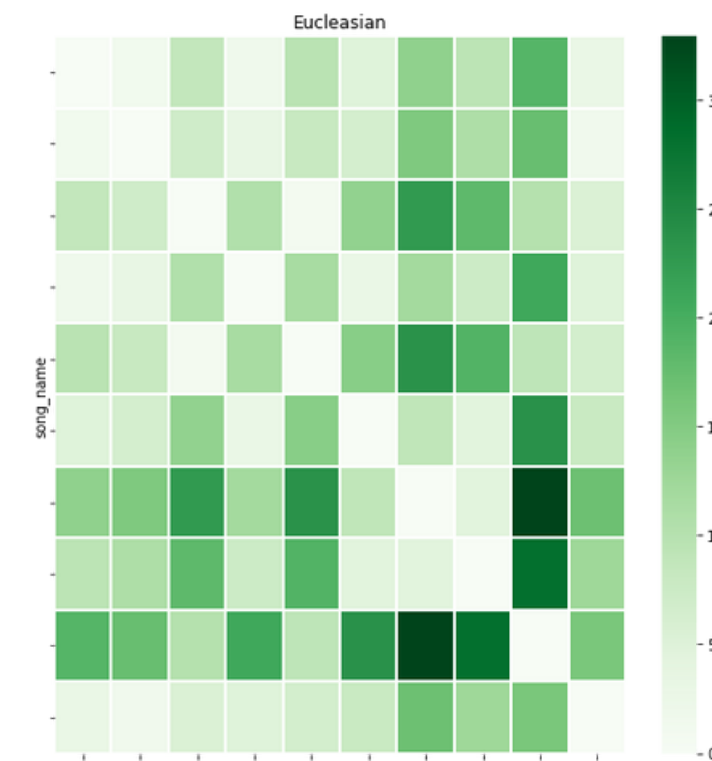
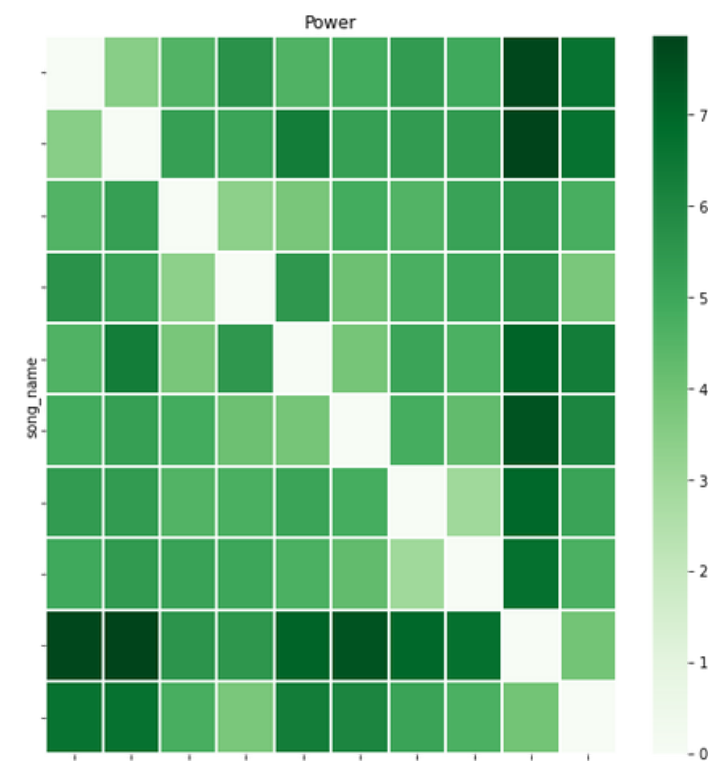
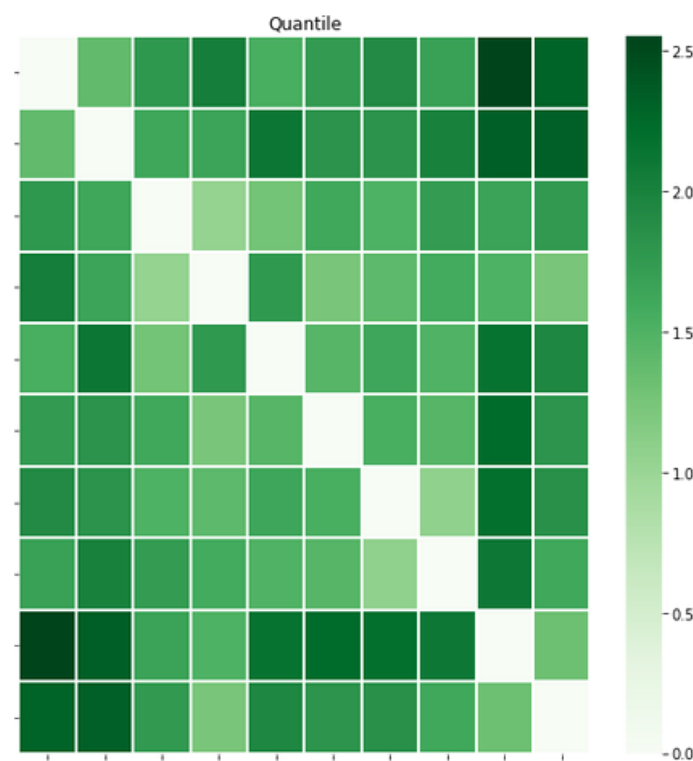


These are :

- Minmax scaling method
- Robust scaling method
- Standard scaling method
- Quantile scaling method
- Power scaling method
- Euclesian scaling method

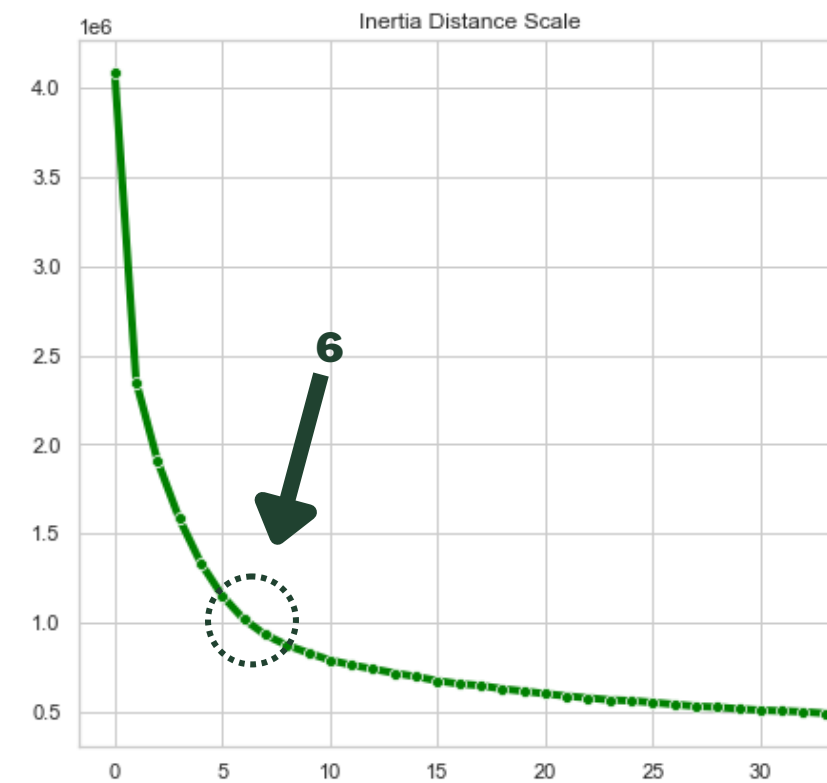
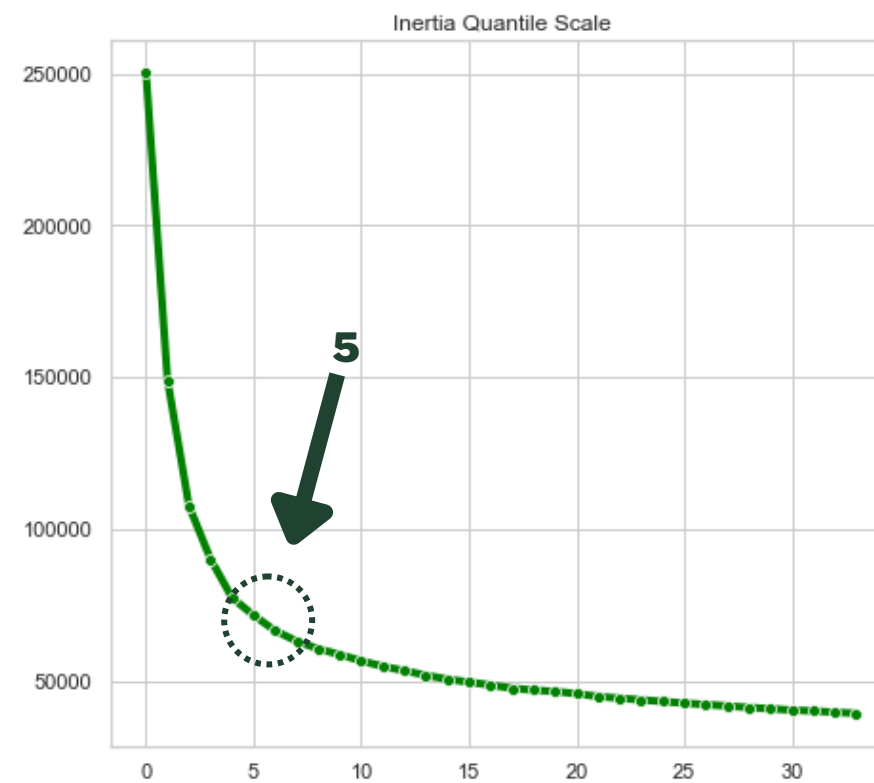
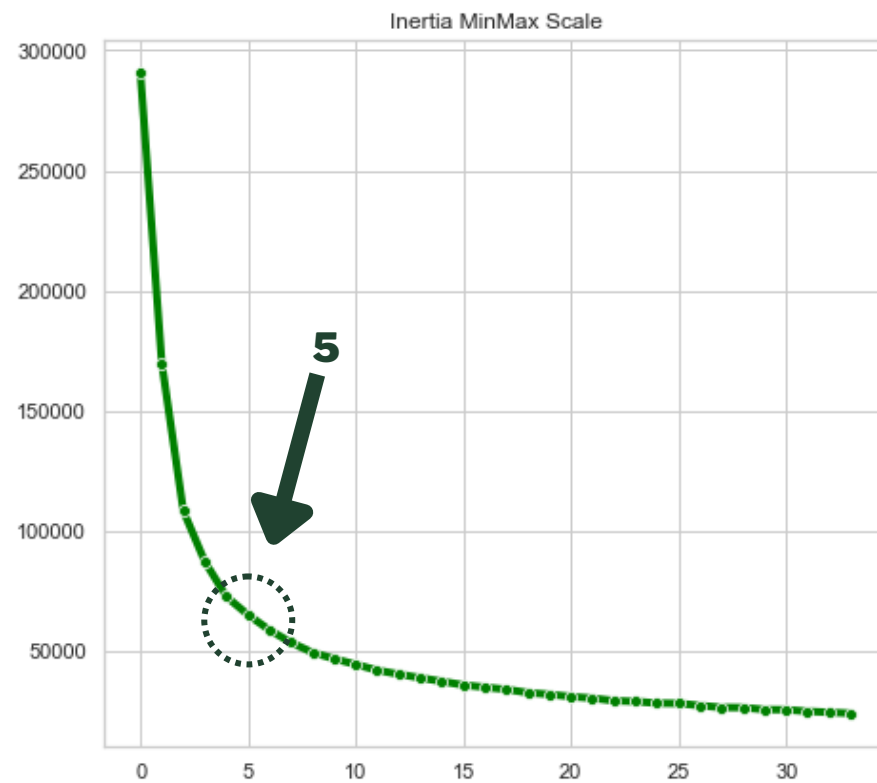
3 of them give a good differentiation:

- Minmax
- Standard
- Quantile



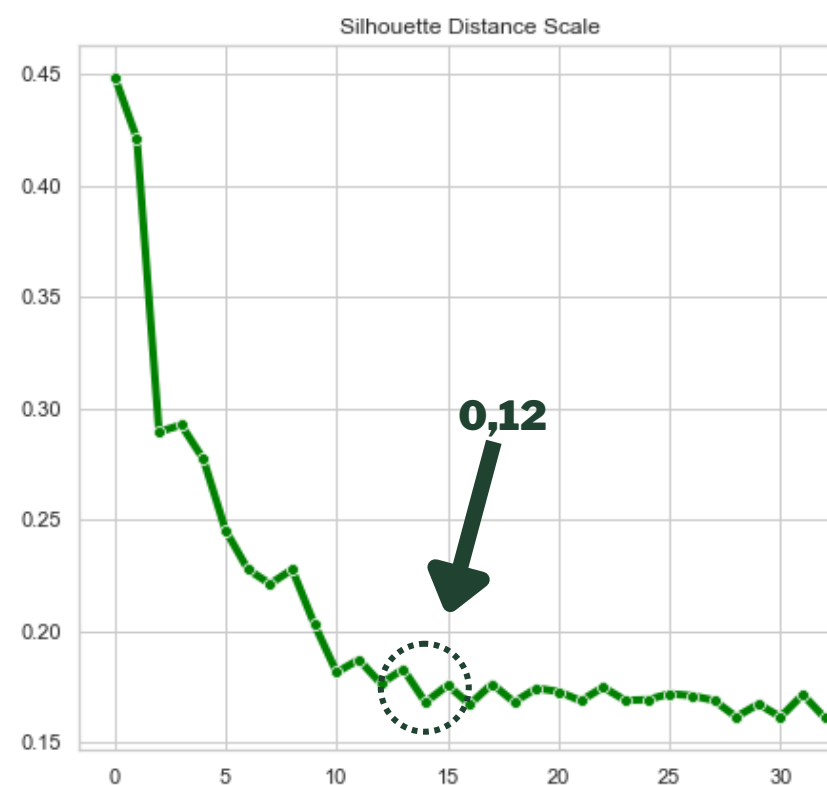
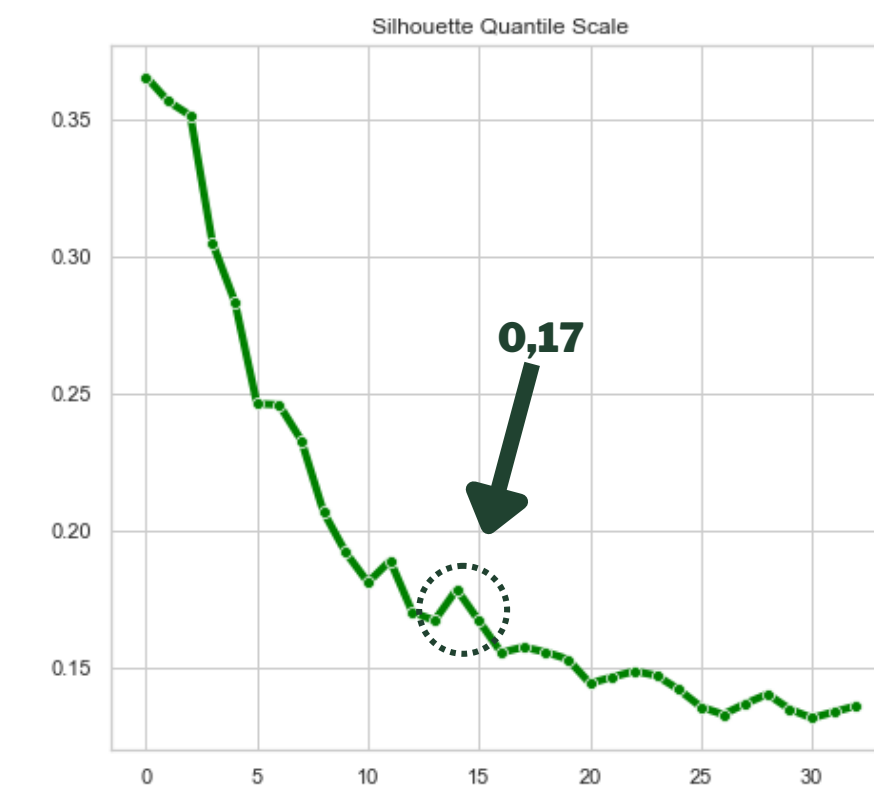
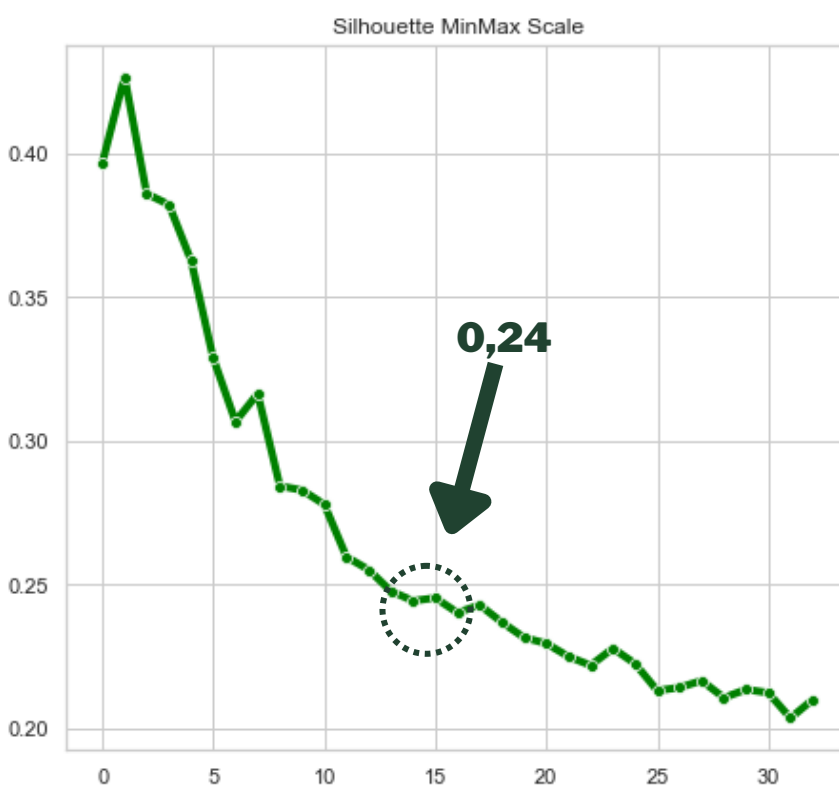


# Inertia and Silhouette Scores (1440 songs)



Inertia Score up to 30 clusters

We can see that 5 or 6 clusters would be a good number, but as said before, this is not an exact science, and I would not separate 1400 songs within 5 playlists.



Silhouette Score up to 30 clusters

We can see that 15 clusters would be a good number, using the Minmax scaling method.

**Decision:**  
Minmax Scaling method and 15 Clusters

# CLUSTERING METHODS

## All Parameters

### Full Clustering

Taking all audio features in consideration to create 15 clusters from the dataframe.

### Subclustering

Taking all audio features in consideration to create 4 'mother' clusters from the dataframe and clusters from these mother clusters.

## Few Parameters

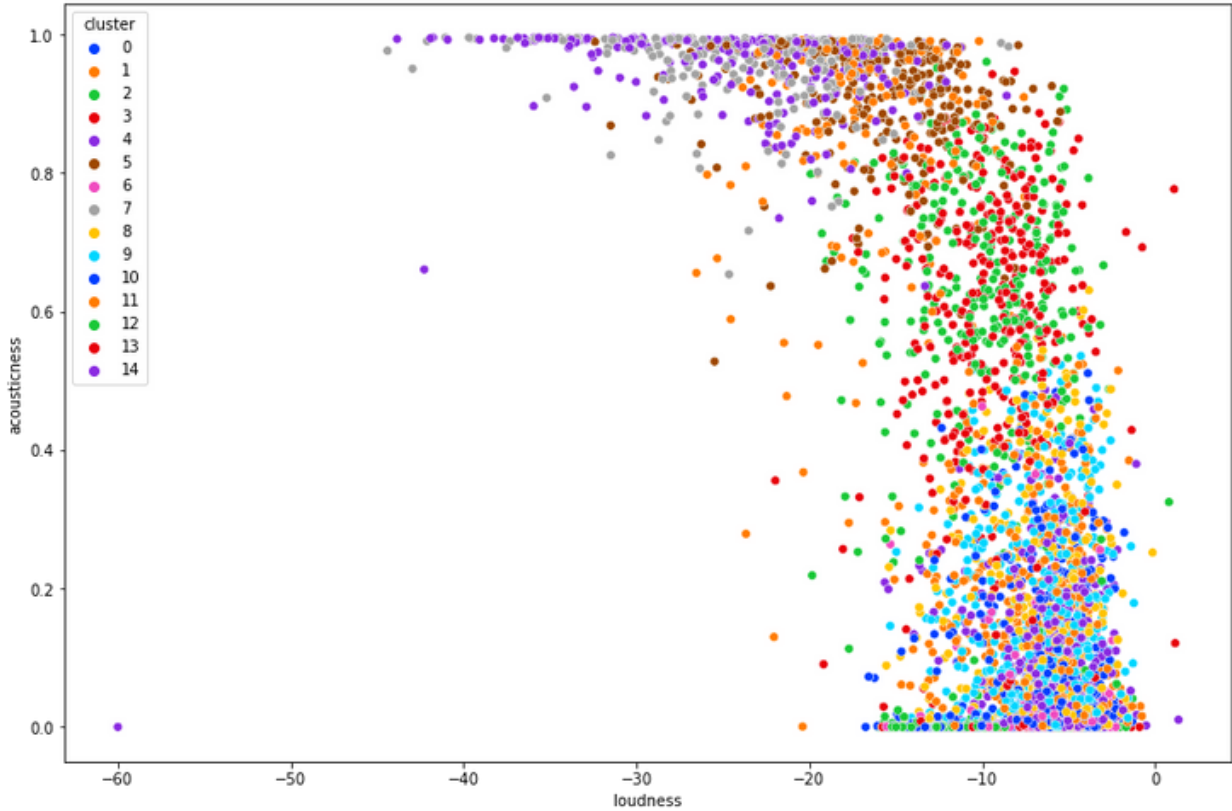
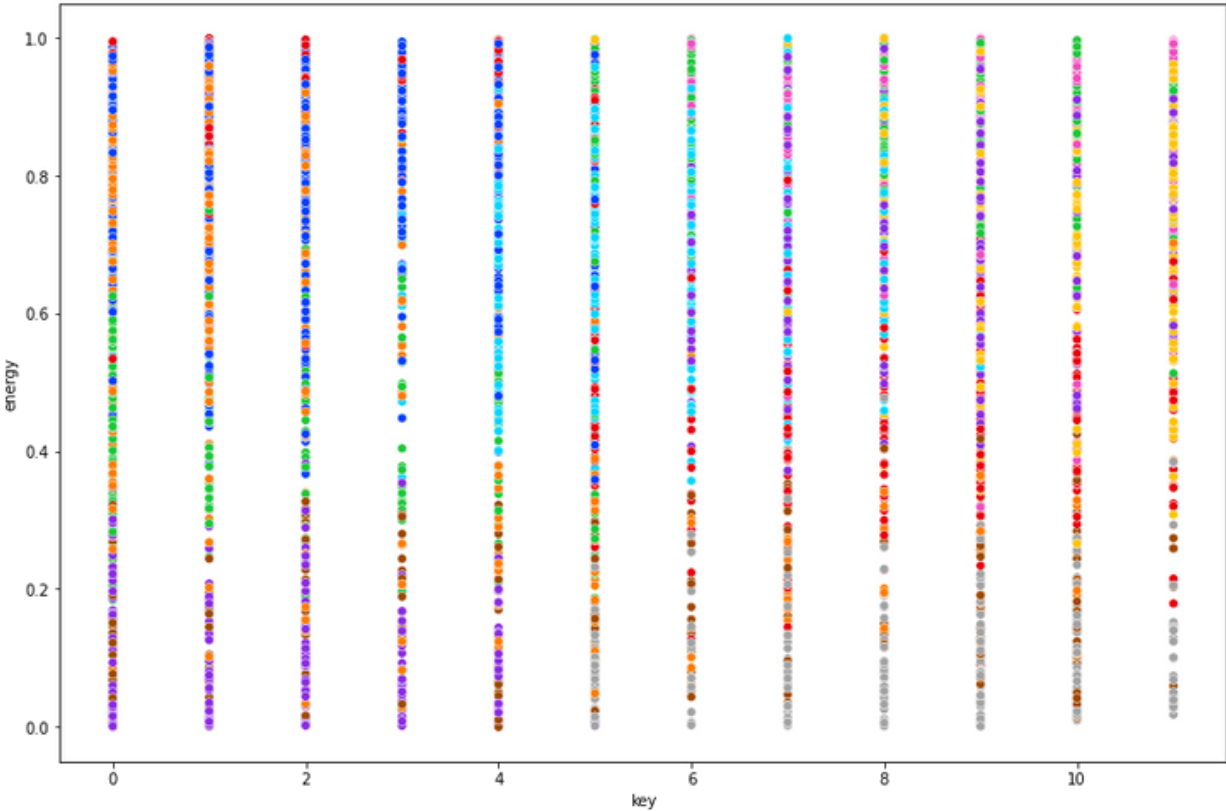
### Full Clustering

Taking three audio features in consideration to create 15 clusters from the dataframe.

### Subclustering

Taking three audio features in consideration to create 4 'mother' clusters from the dataframe and clusters from these mother clusters.

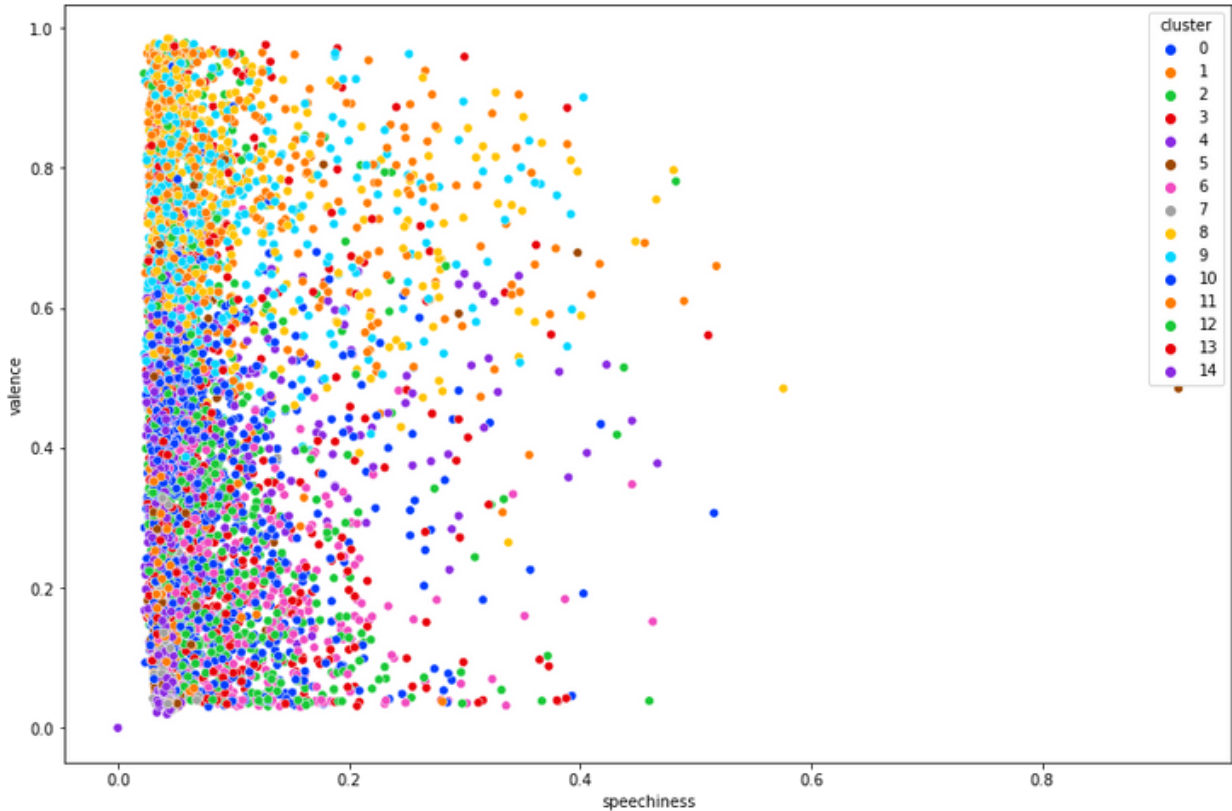
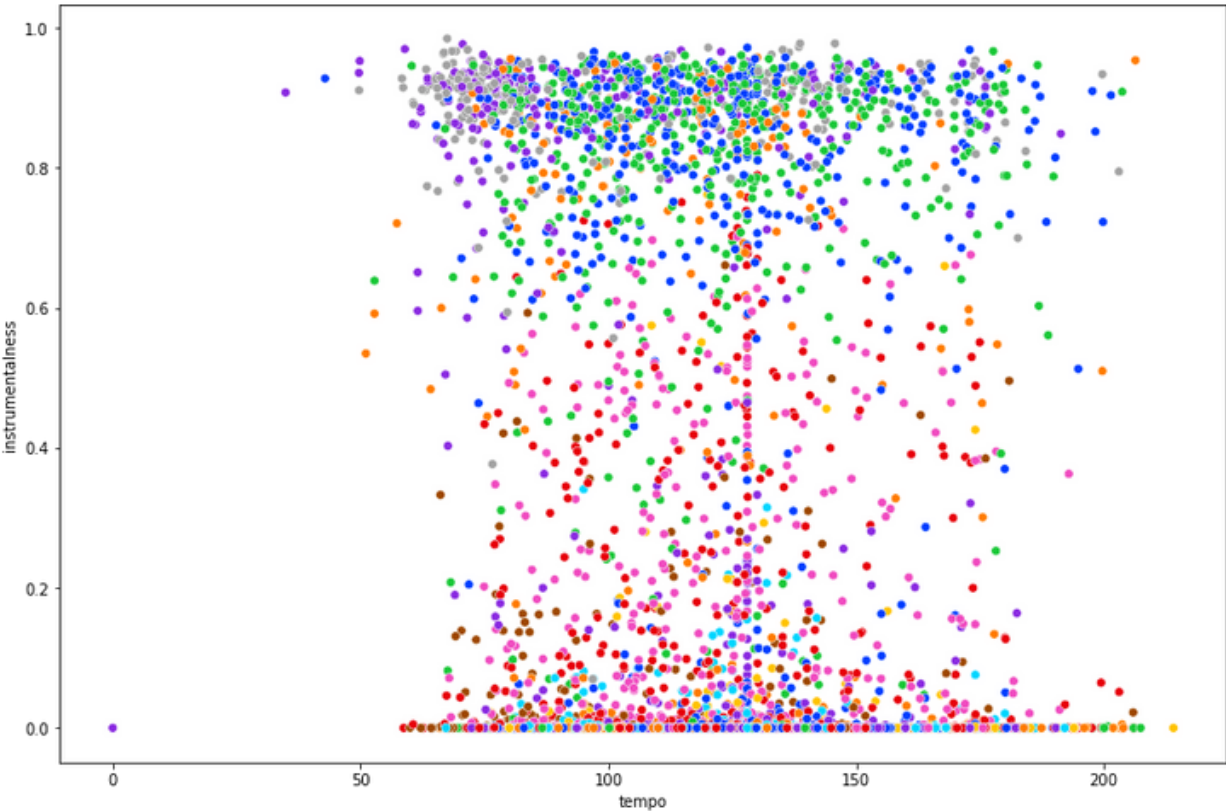
# All Parameters / 15 Clusters (5000 songs)



Due to repeating color, distinguishing clusters is a bit difficult. We can anyway see that some clusters has more tendency according to which parameters we look.

You Can Get It If You Really Want	Desmond Dekker	6
He's a Doll	The Honeys	6
Under Your Spell	Desire	6
San Francisco	Foxygen	6
Manada	Bagunço	6

Jazz in a pop cluster

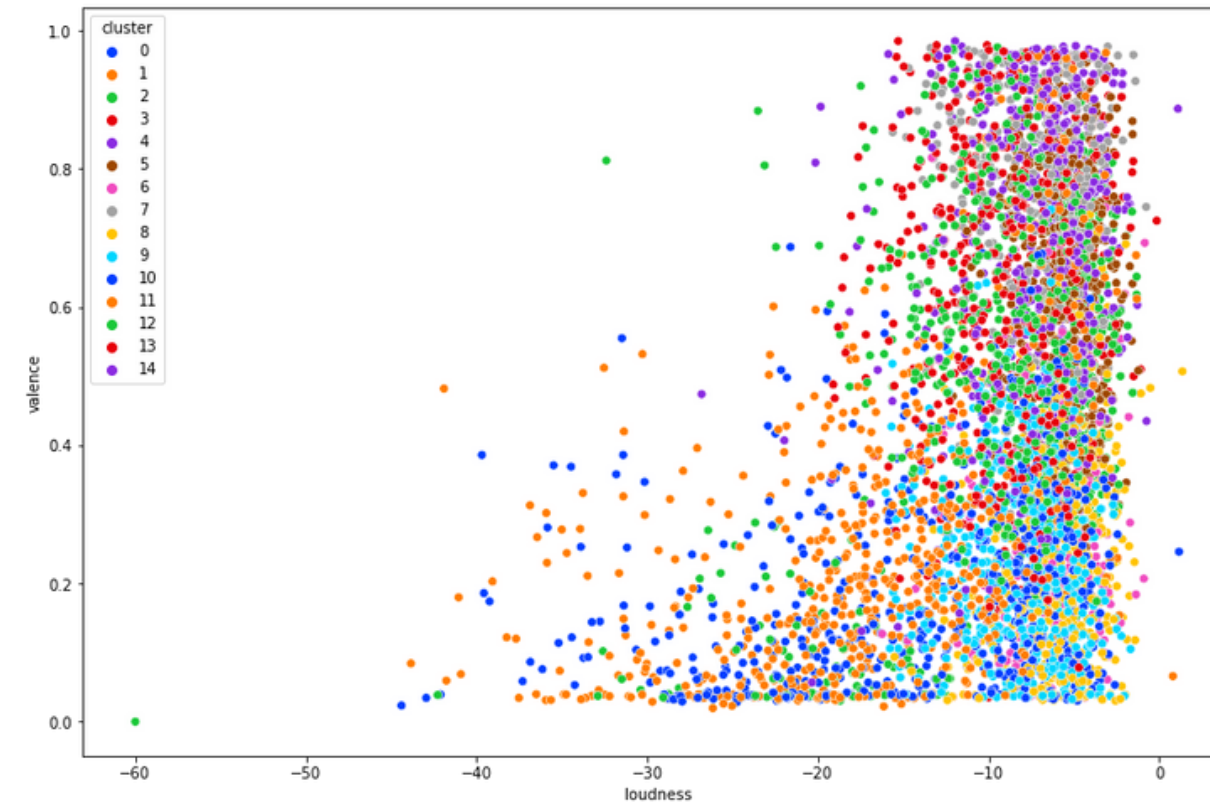
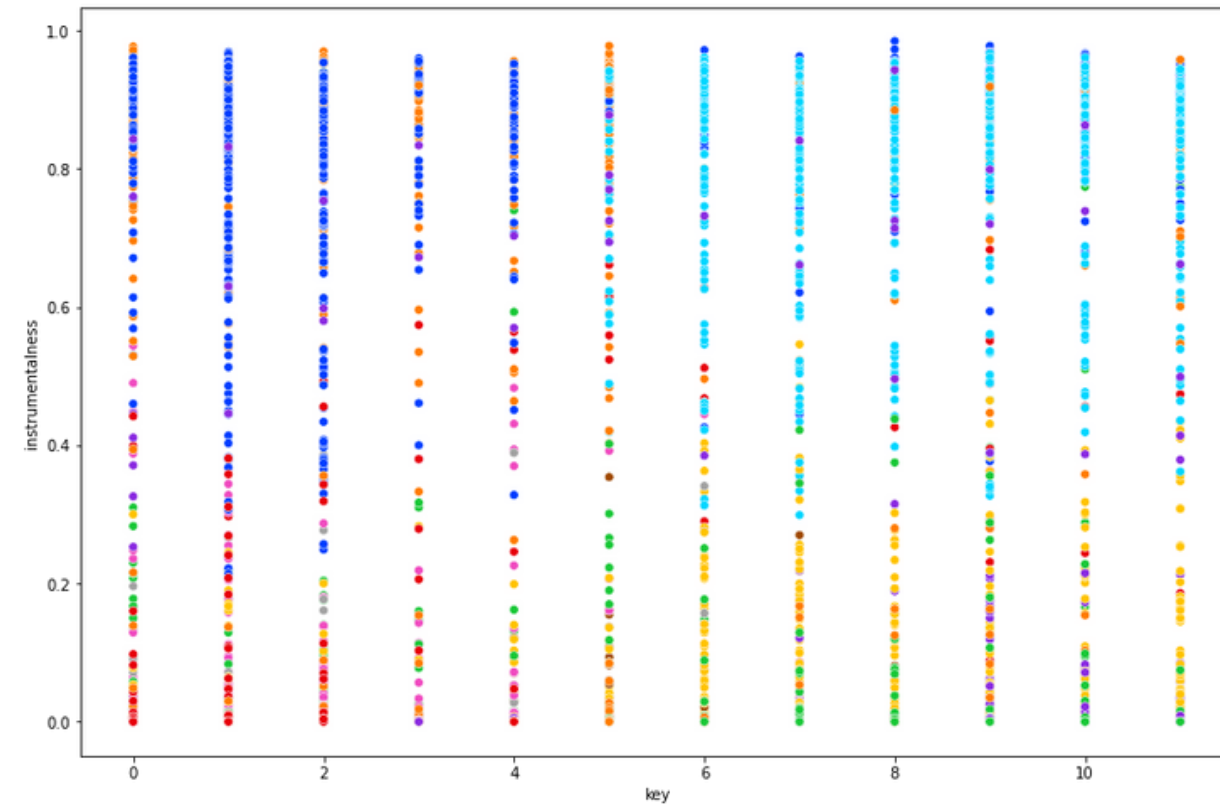


Brigas Nunca Mais	Brazilian Jazz	14
O Pato (The Duck) - Live At Carnegie Hall/1964	João Gilberto	14
Little River	The Tallest Man On Earth	14
Shake It Off	Taylor Swift	14
Two Skies	Dirk Maassen	14

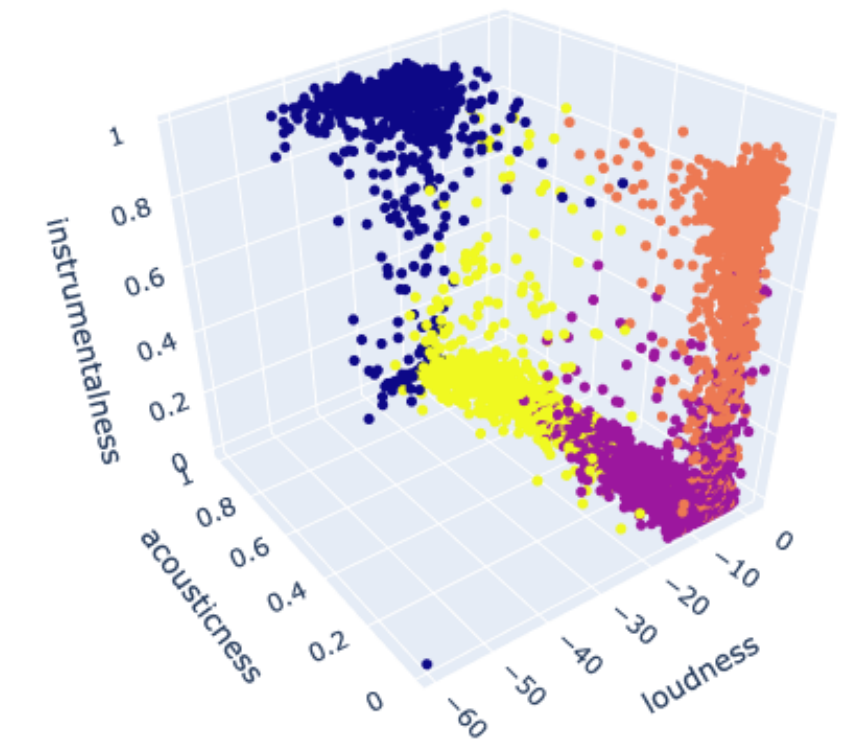
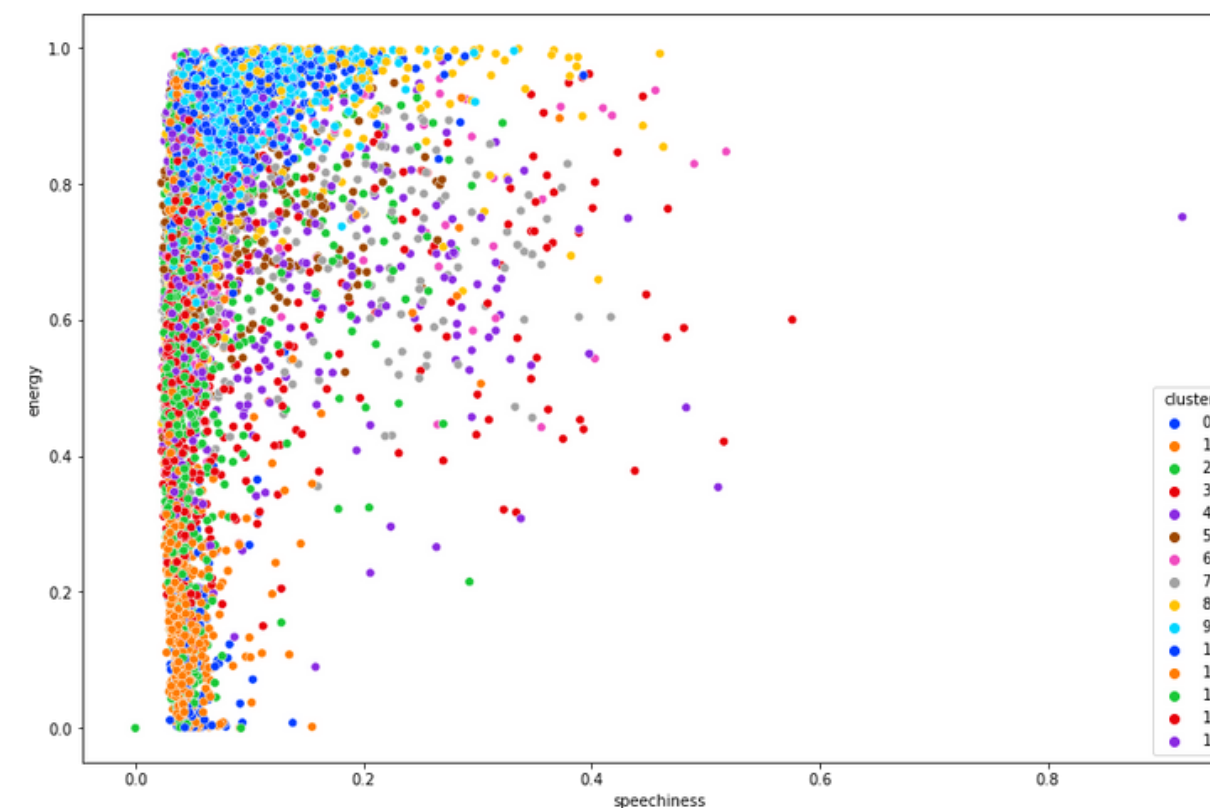
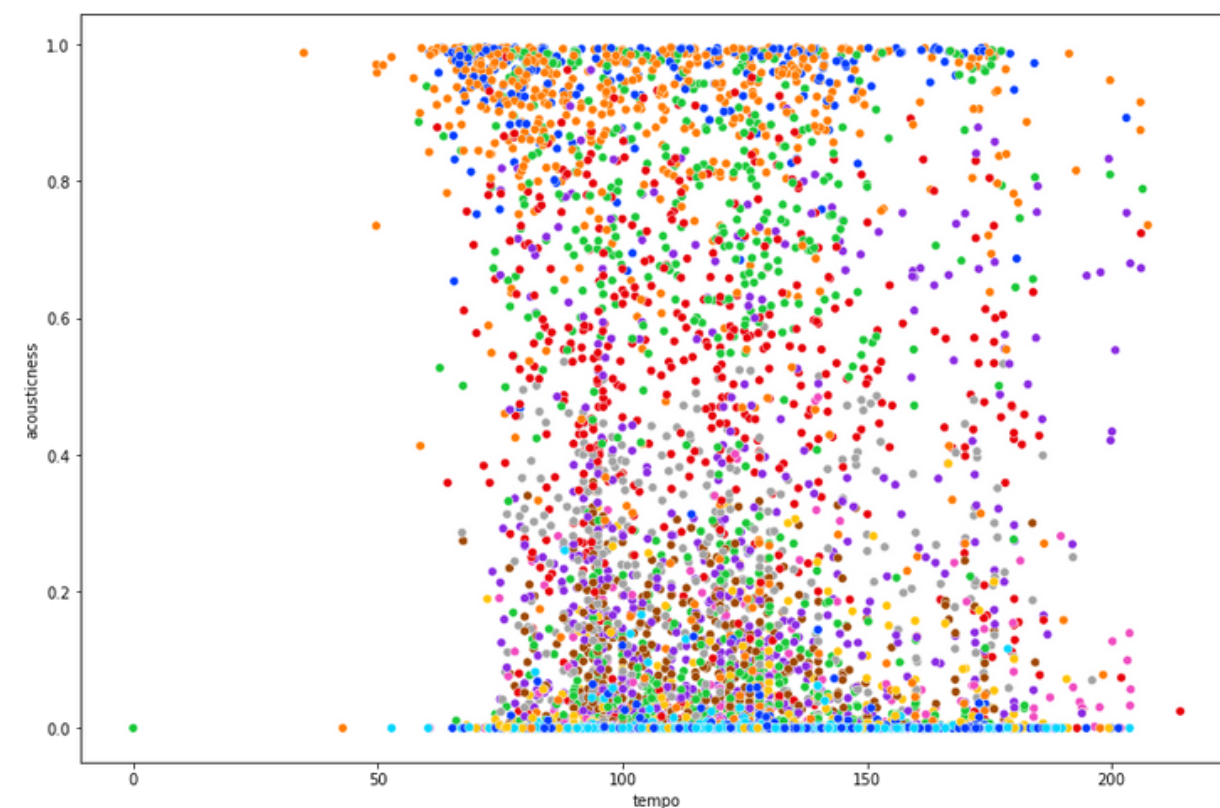
Recent Pop or Romantic Piano in Latin Jazz



# All Parameters / 4 Clusters + Subclusters

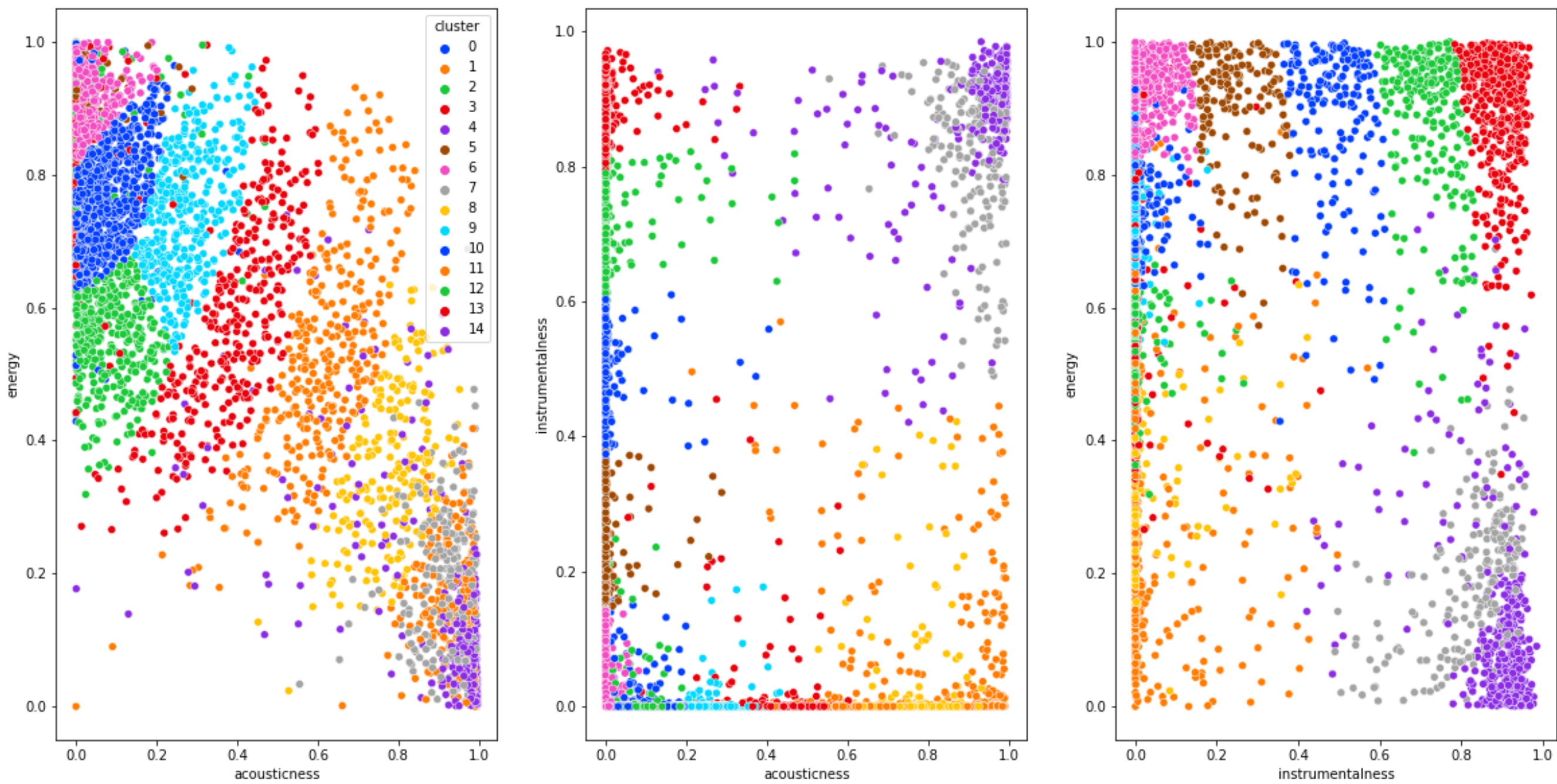


It looks even better with subclustering.  
On the left, all 15 clusters, above, the 4 mother clusters:





# Few Parameters / 15 Clusters



After inspecting the previous graphs, I decided to go only with 3 parameters:

- Instrumentalness
- Energy
- Acousticity

The Clusters are clearly defined.

Les Dones Macabres	Nico Roig	0
Aquele Abraço	Gilberto Gil	0
Island (Nôze Remix)	Dapayk & Padberg	0
Maroca	Mundo Livre	0
A New Error	Moderat	0

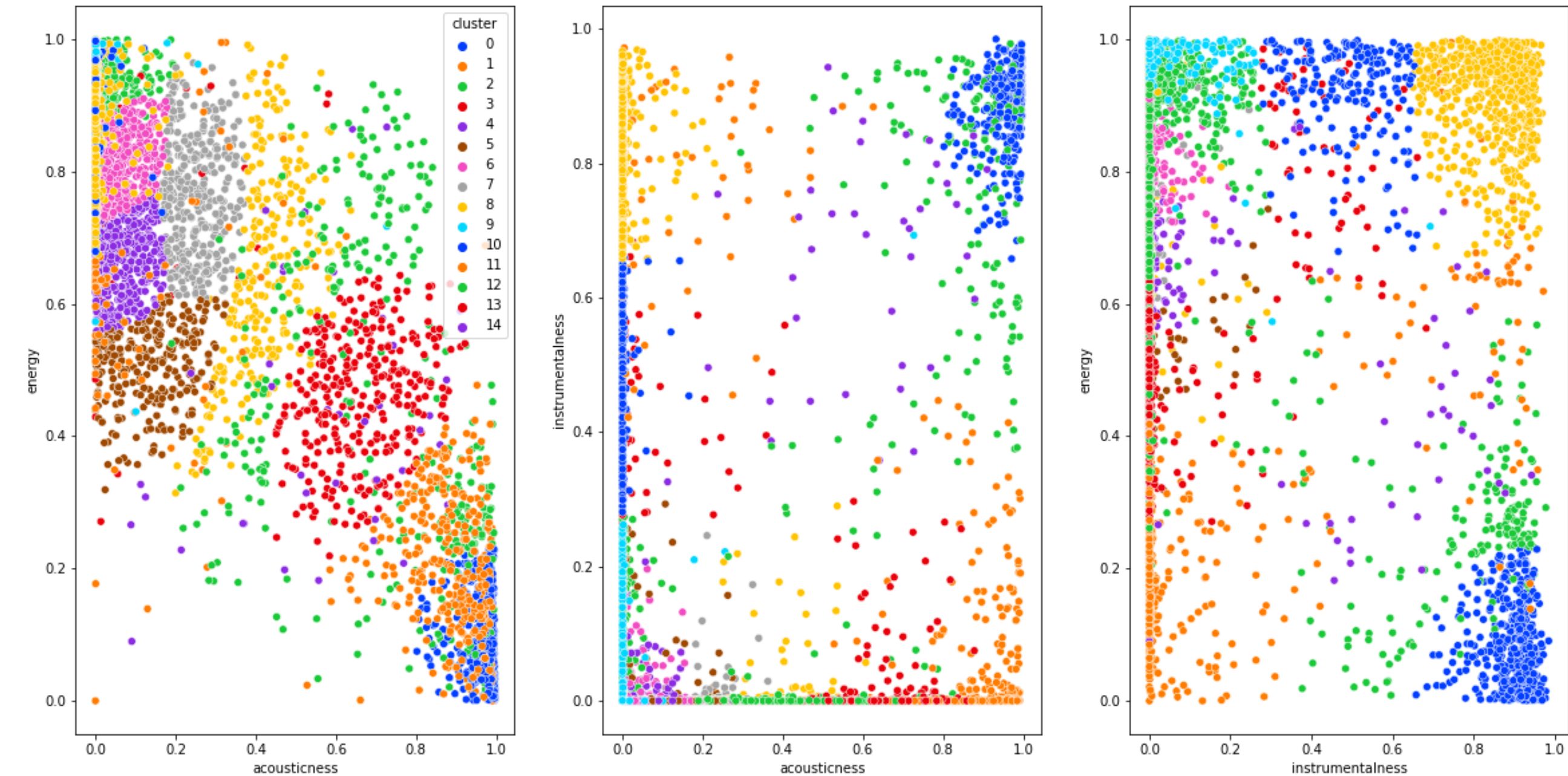
*Techno/House in Latino Jazz*

Tirei o Chapéu	Batida	6
Devil Or Angel	Lou Doillon	6
It's Real	Real Estate	6
Odessa	Caribou	6
Game Of Pricks	Guided By Voices	6

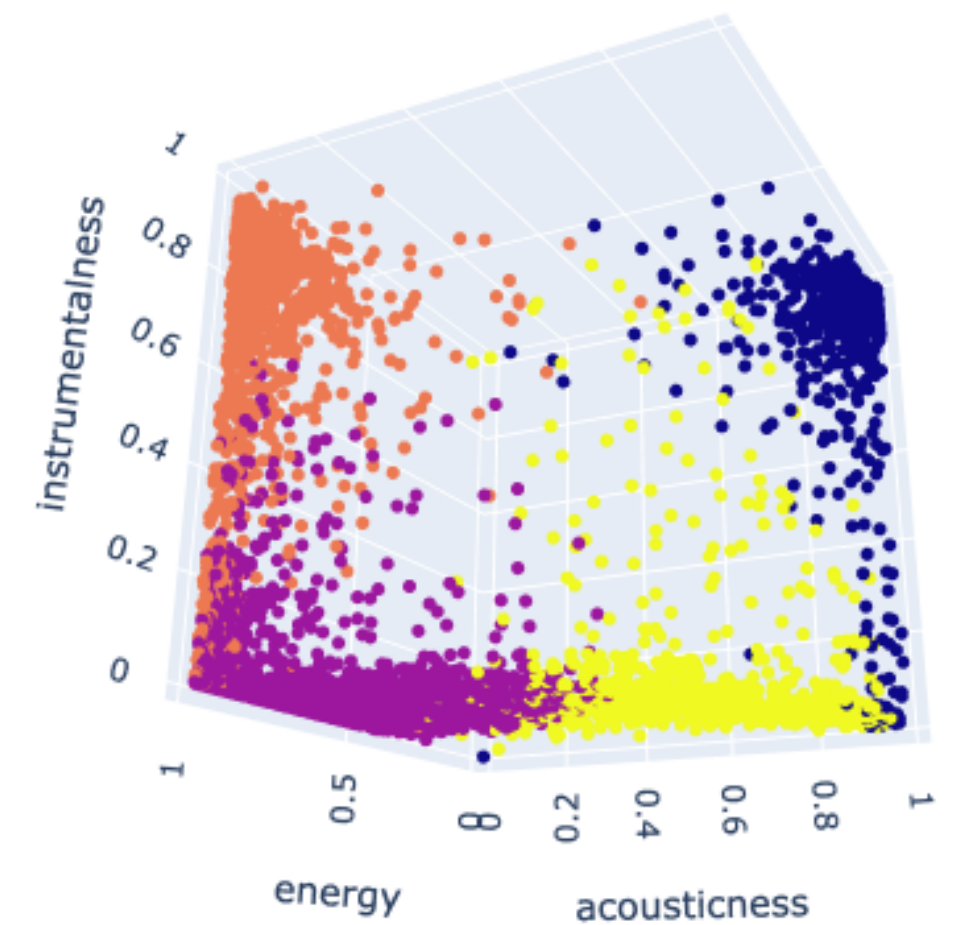
*A mix of pop, Jazz vibes, Electronics*



# Few Parameters / 4 Clusters + Subclusters



On the left, all 15 clusters, above, the 4 mother clusters:



# CONCLUSION

## **Best Method on the charts**

### **Subclustering with all parameters**

Subclustering and keeping all audio features worked better than we thought while listening the playlists. It is far to be good playlists, but at least we started to get wide groups.

### **Only 3 parameters**

Keeping only few features helped a lot and gave a medium result. I would go further on it, and take some time to explore this. The main point is to find the right features

## **Recommendation**

### **KMean works**

I believe that KMean can works. I have no experience on other model so can't recommend a maybe better one.

### **Knowledges missing**

My knowledges are pretty limited here since the point of this week project was to understand the concept of unsupervised ML and Kmean.

### **Time restrictions**

Within one working week, we have learned about concept of ML and KMean. At the same time, we needed to prepare and explore these datas. We got only 2 days to really explore it.



# Thank You!

Do not hesitate to look at my codes:

[Github](#)