Moosic

Creating Playlist in spotify using KMeans Clustering

By
Kalpana, Ben, Mohamed, Maath

Objective: Business question

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?

Steps Involved:

- Preprocessing the data
- Identify the optimal number of clusters
- Clustering using Kmeans
- Identify the patterns in our clusters
- Using Patterns label the Cluster.

Data cleaning and preprocessing

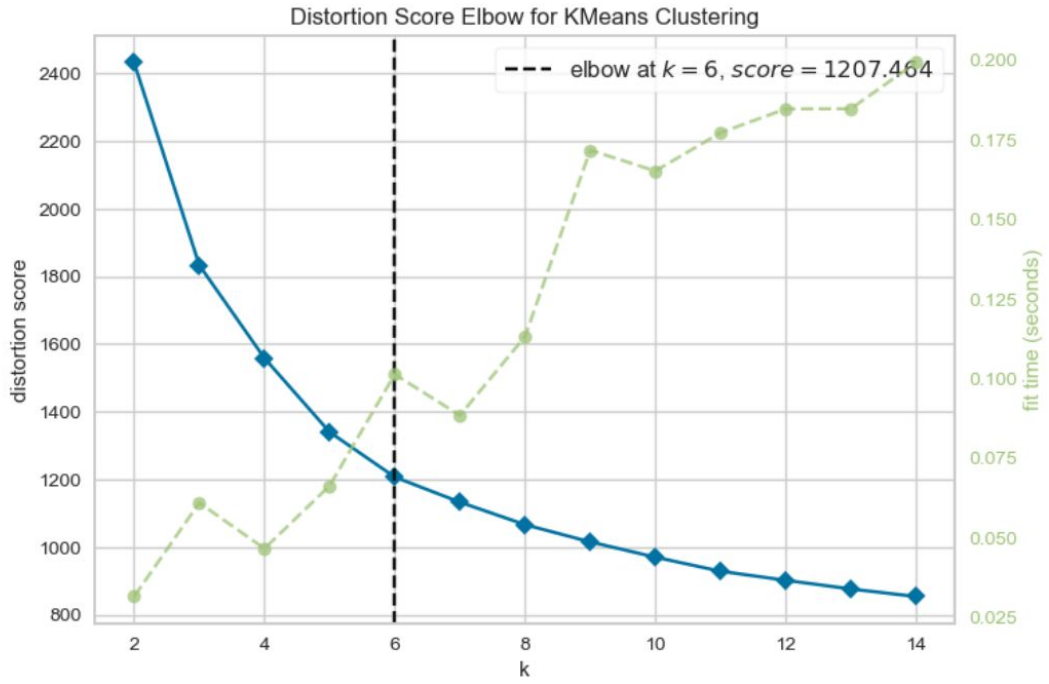
- Checked and Removed duplicates
- Drop NAN values
- Cleaned empty spaces in column-names
- Deleted attributes like id, html, mode, type,duration_ms
- **Size of Dataframe:** 5235 Songs

name	artist	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	time_signature
Se Eu Quiser Falar Com Deus	Gilberto Gil	0.658	0.2590	11	-13.141	0	0.0705	0.694	0.000059	0.975	0.306	110.376	4
Saudade De Bahia	Antônio Carlos Jobim	0.742	0.3990	2	-12.646	1	0.0346	0.217	0.000002	0.107	0.693	125.039	4
Canta Canta, Minha Gente	Martinho Da Vila	0.851	0.7300	2	-11.048	1	0.3470	0.453	0.000063	0.124	0.905	93.698	4
Mulher Eu Sei	Chico César	0.705	0.0502	4	-18.115	1	0.0471	0.879	0.000041	0.386	0.524	106.802	4
Rosa Morena	Kurt Elling	0.651	0.1190	6	-19.807	1	0.0380	0.916	0.000343	0.104	0.402	120.941	4

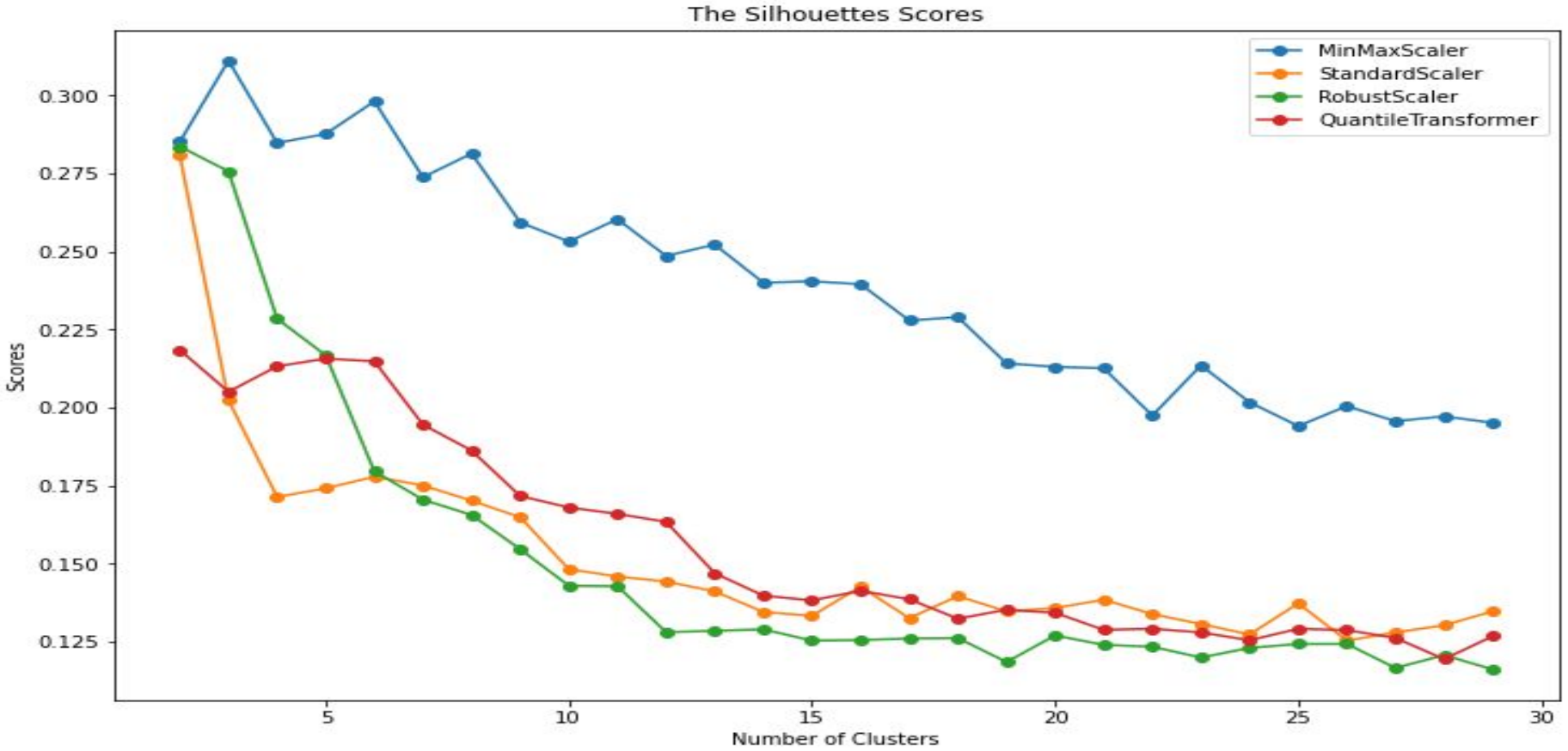
Identify the optimal number of clusters

Identified the elbow point at 6 from other metrics like inertia and silhouette score too.

Distortion score which is the sum of square distances of each point to its associated center.



Silhouette scores from scaled data



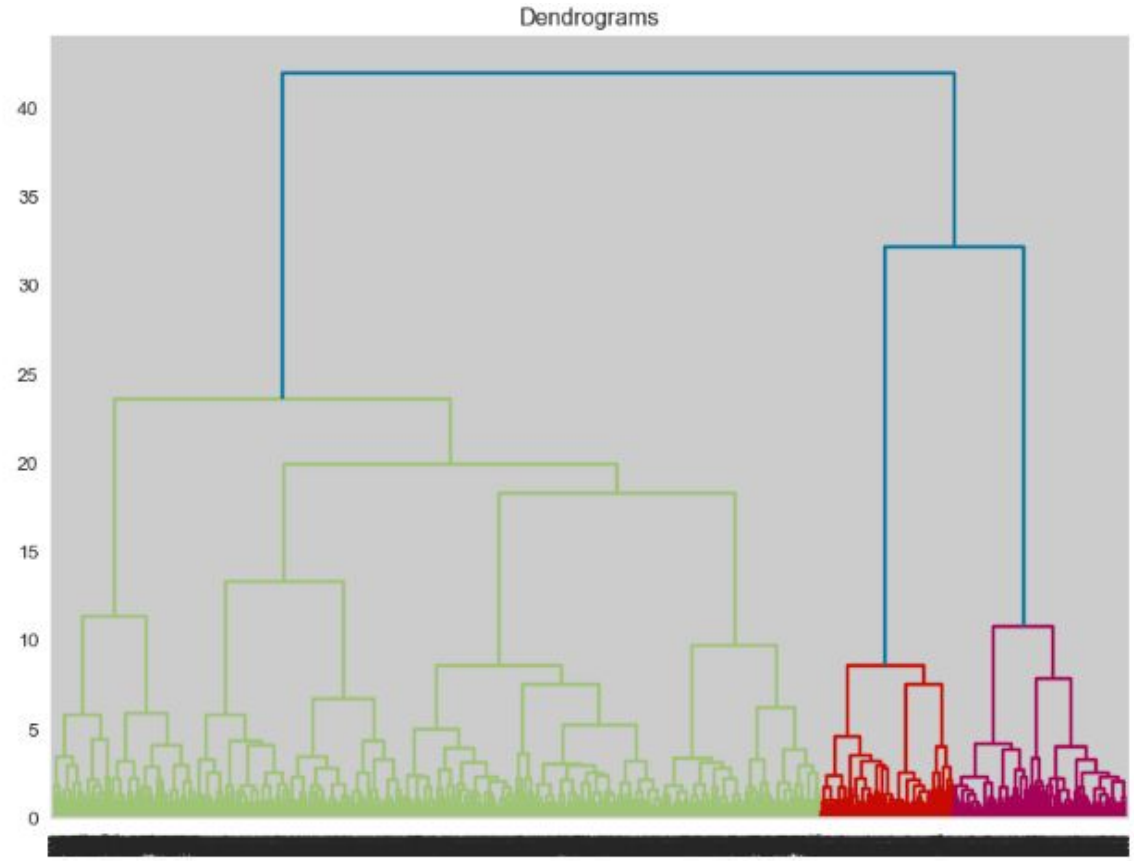
Manual Exploration of Clusters

S.No	Cluster no 6	Cluster no 7	Cluster Size
1	0	2,3,6	438,3,429
2	1	2,3,4	7, 4,1037
3	2	5	686
4	3	2,3,6	8,753,24
5	4	0,2	599,2
6	5	1,3,5,6	1225,12,3,6

Granularity of the Clusters

Dendrogram show the cluster as 3

The cluster at the lowest level shows high density and cohesiveness.

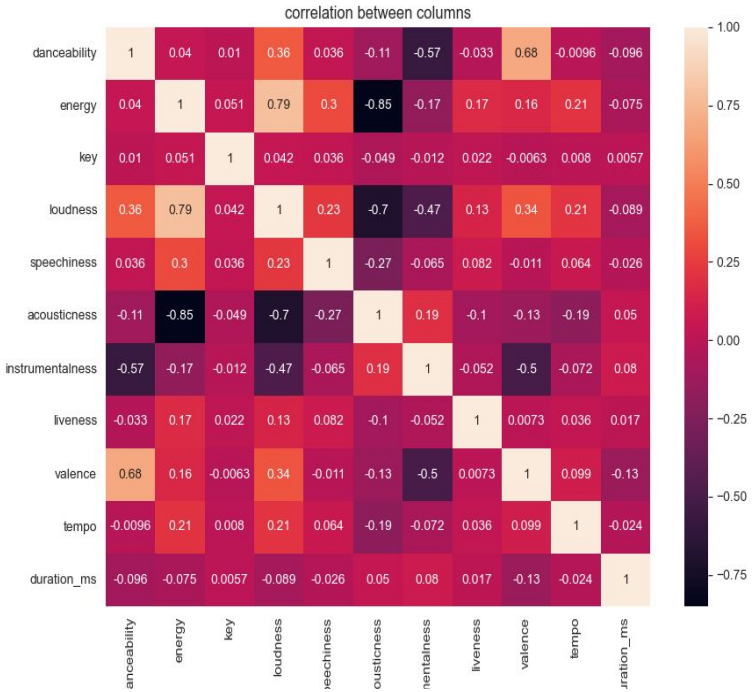


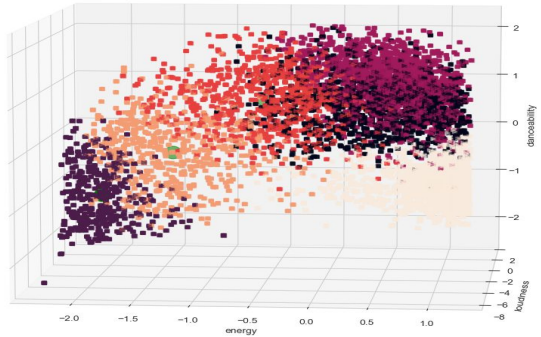
Playlist:

cluster	danceability	energy	key	loudness	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms
0	-1.063905	0.880440	0.028670	0.251819	0.306583	-0.781182	0.799941	0.006432	-0.804515	0.093549	0.014071
1	0.801250	0.275484	0.024655	0.522588	2.699168	-0.318343	-0.646848	-0.076617	0.498247	0.154048	-0.053515
2	-0.950268	-1.859547	-0.061326	-2.065933	-0.537021	1.807852	1.382665	-0.336327	-1.075446	-0.520486	0.353336
3	0.219782	-0.940866	-0.179411	-0.339459	-0.521489	1.100341	-0.489710	-0.263666	0.181153	-0.139482	-0.084709
4	0.711005	0.320738	0.045105	0.500503	-0.316380	-0.435199	-0.580609	-0.227703	0.670797	0.118330	-0.100671
5	0.218403	0.337699	0.136079	0.364434	0.029375	-0.235933	-0.426026	2.834075	0.156976	0.126873	0.111632

- Cluster 0:**Playlist 1[Energy Music]**
- Cluster 1**:Playlist 2[Dance Music]**
- Cluster 2:**Playlist 3[Acoustic Music]**
- Cluster 3:**Playlist 4[Chillout Music]**
- Cluster 4**:Playlist 5[Valence Music]**
- Cluster 5**:Playlist 6[Live Music]**

Correlation between columns:





Limitations

Choosing k manually.

Lack of domain specific knowledge.

Requires more information like genre, styles.

Limited by time constraint.

Thank you for your attention