DS 5110

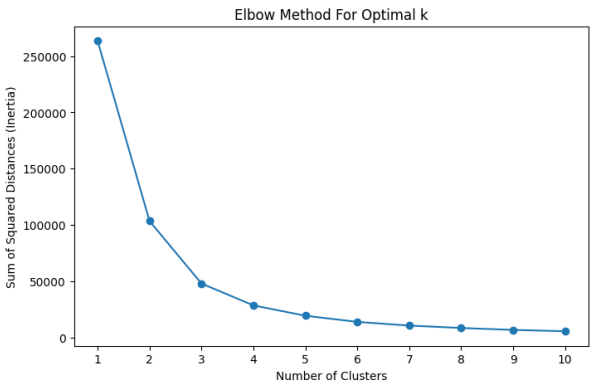
HW 7

Cody Snow

April 1, 2025

K-Means – 3D Data Set

Elbow Graph: K=3

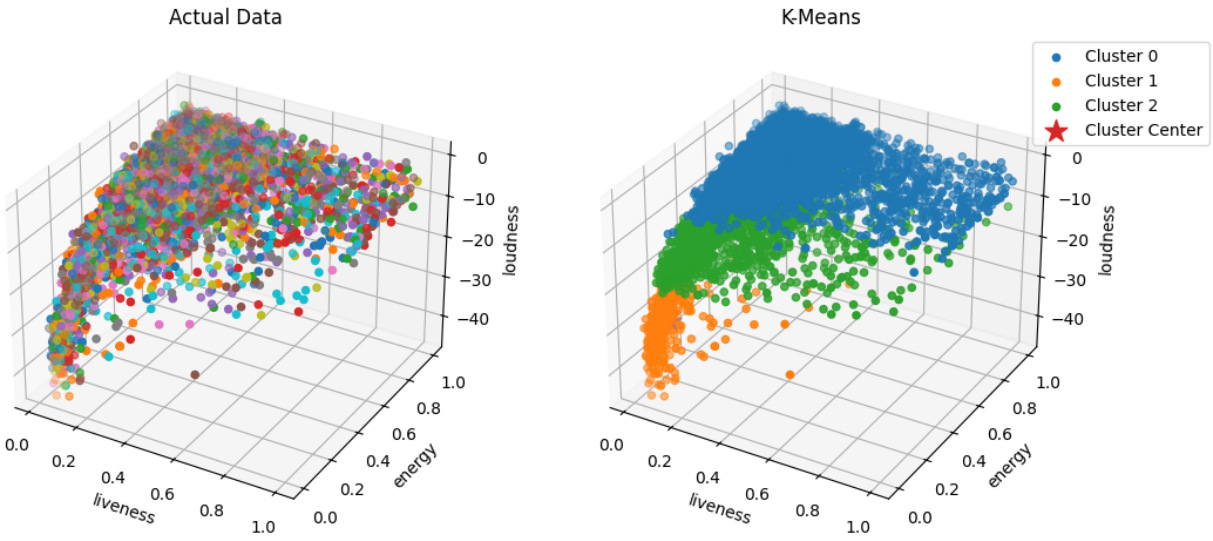


I debated whether the elbow graph is telling me 3 or 4 clusters. I tried 4 but ultimately went with 3 given how the data points plot out in my interactive graph. Also, the “elbow” in the graph does appear to be the 3 vs. the 4.

I modified the code provided in class to include a third axis, using the new dataset provided. This was straightforward – the k_clusters array was increased to contain three lists instead of two. Then within the nested for loop that loops through each datapoint, I appended the z axis to the third list. The biggest change required was in the final loops that generated the legend and the cluster centers.

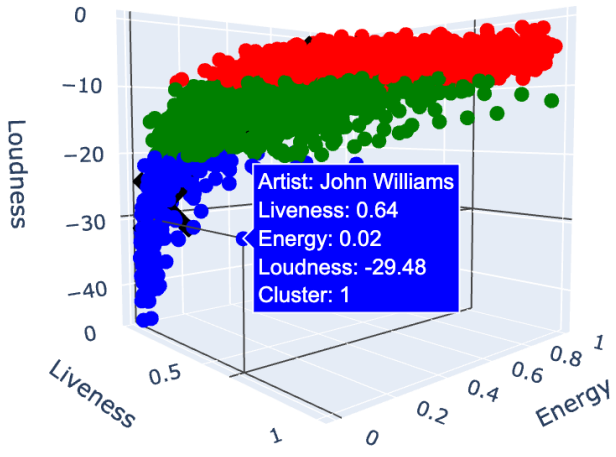
The three clusters are clearly presented in the K-Means graph on the right. Unfortunately, I couldn’t get the cluster centers to show up—they are obscured by the many, many dots in the graph.

The three groupings show an interesting relationship between loudness and liveness/energy. There are no quiet songs that have high energy or liveness. Likewise, there are no songs with higher energy/liveness and a low loudness level. The middle group, shown in green, shows a very wide distribution of liveness, energy, and even a larger span of loudness than the blue group. However, the greatest span of loudness is shown by the yellow group.

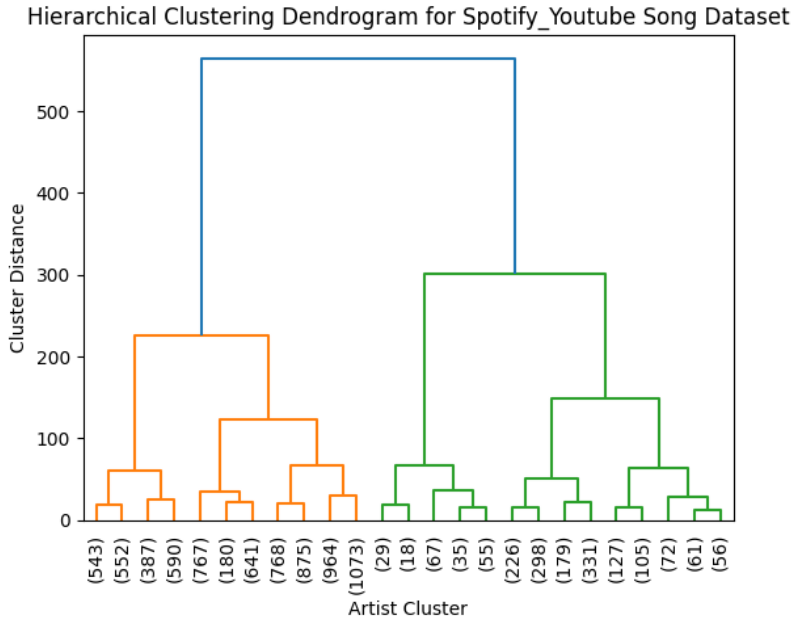


This group contains artists like the great classical composers (e.g. Mozart) and film composers like John Williams. The middle group has an interesting crossover of classic rock and more modern pop, whereas the blue group contains rap and harder rock artists. This is my main observation—that the clusters correspond very roughly to broad musical genres.

For fun, I used ChatGPT to help me create an interactive graph for the K-Means clusters using plotly. This allowed me to rotate and zoom in and out, and examine individual data points, to which I added the artist and the features we're analyzing. (to run this, use Google Colab, as it doesn't appear to run locally on my mac). I was also able to get a better look at the centers.



Hierarchical Clustering



There are distinct groups here as well, but as discussed in class, cutting the dendogram is somewhat subjective. If I cut it at 150, I get four distinct groups. If I cut it at 250, I get three. I went with three. I wanted to see the max/min of each feature for each cluster, so I printed them out:

	Liveness		Energy		Loudness	
	min	max	min	max	min	max
Cluster						
1	0.0145	1.000	0.09640	0.997	-12.159	-0.140
2	0.0418	0.635	0.00194	0.155	-44.761	-25.933
3	0.0190	0.910	0.01200	0.890	-25.773	-11.866

Interestingly, the hierarchical clusters line up with the K-Means clusters. For example, cluster 2 has a loudness min and max that corresponds to the yellow K-Means cluster pictured above, as well as the liveness and energy. So, once again, we have found the relationships between a song's loudness, energy, and liveness that indicates these artists are split across styles of music or rough genres.