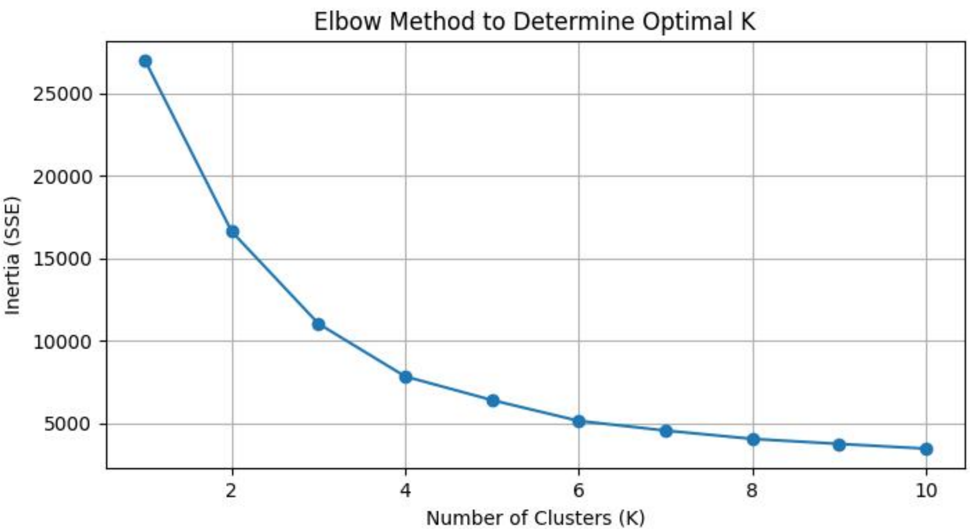
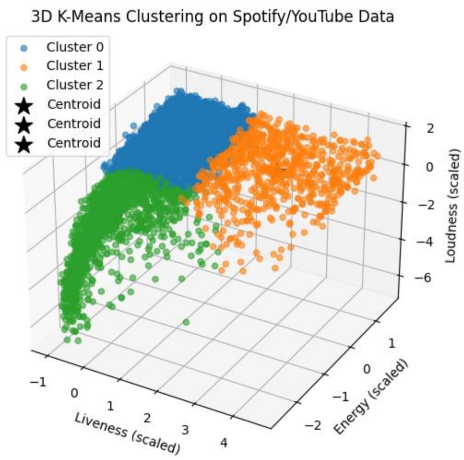
DS5110
HW7
Name: Wenyi Ye
Date: Mar 31.2025

1. To update the original K-means code to work for 3D data, I modified the feature selection to include three dimensions Liveness, Energy, and Loudness from the Spotify_YouTube.csv dataset, instead of just two features like in the original Iris dataset. I used StandardScaler to standardize the values because these features have different units and scales, which can affect clustering performance. I then applied K-means clustering using scikit-learn and used the elbow method to determine the optimal number of clusters by plotting inertia SSE against different values of K. From the elbow graph, K=3 appeared to be the optimal choice, as the decrease in SSE started to level off after that point. For visualization, I created a 3D scatter plot using matplotlib and mpl_toolkits.mplot3d, where each data point was colored based on its assigned cluster, and the centroids were marked with black stars. The resulting plot revealed three distinct groupings of songs. One cluster seemed to represent high-energy, loud tracks, possibly suited for parties or workouts. Another cluster likely contained calmer, quieter tracks, while the third group fell somewhere in between. This clustering could be useful for organizing music by mood or intensity, recommending songs to users based on listening preferences, or even for curating playlists with a consistent vibe.





2. After running hierarchical clustering on the standardized features Liveness, Energy, and Loudness from the Spotify/YouTube dataset, I generated a dendrogram using Ward's linkage method. The dendrogram revealed a clear separation into three major clusters, aligning well with the earlier K-means result. When visualized in a 3D scatter plot, the hierarchical clusters showed distinct groupings of songs based on their audio characteristics. One group clustered around high loudness and energy levels, likely representing upbeat or dance tracks. Another group featured lower loudness and energy, possibly corresponding to mellow or acoustic songs. The third group fell somewhere in between. These findings suggest that hierarchical clustering, like K-means, can effectively uncover natural groupings in musical features. This technique is especially useful when the number of clusters is not known in advance, as the dendrogram provides a visual guide for selecting the optimal cut-off point. Overall, hierarchical clustering supports the presence of meaningful, interpretable groupings within the music data.