Case Study 2 Documentation: Clustering of Music Tracks

Overview: This report details the process and results of clustering 2,683 music tracks to categorize them into potential genres based on loudness, speechiness, and instrumentalness. The optimal number of clusters was determined to be three, with a silhouette score of 58.81%.

1.Action Summary

Libraries Import: In this analysis, I used key Python libraries for data manipulation and visualization. pandas was used to handle the dataset, chardet to determine the file encoding, and StandardScaler, KMeans, and silhouette_score from scikit-learn were employed to build the clustering model. For visualization, seaborn and matplotlib provided insightful graphs.

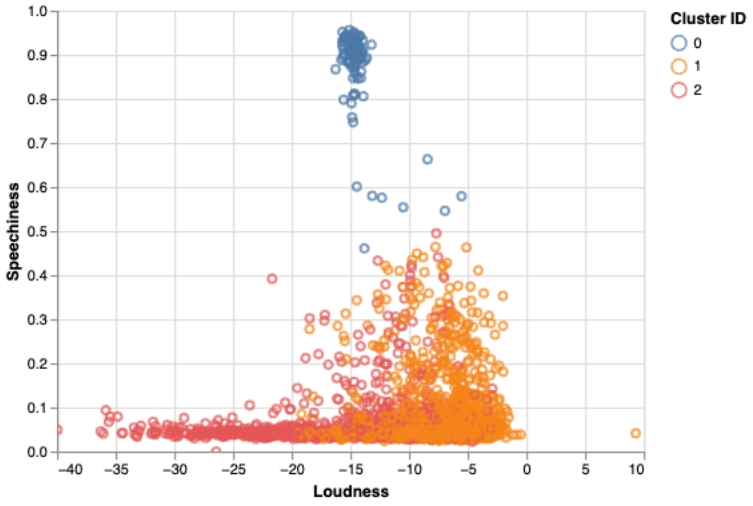
Data Preparation

Upon initial data inspection, I observed and cleaned unique identifiers and dropped duplicate entries based on "ID" while preserving songs with the same "Name" due to potential multiple versions. Missing values were detected across the columns and subsequently removed, ensuring data integrity for our analysis. At this point, our data is ready for further analysis.

Analysis

My primary focus was to cluster these songs into distinct genres using attributes such as loudness, speechiness, and instrumentalness, excluding energy to avoid redundancy with loudness. The dataset was standardized for uniformity, and a KMeans clustering model was implemented to identify three distinct clusters. The Cluster_ID column was then added to the dataset for further examination.

Cluster_ID
1 1574
2 1001
0 108



Visualization with pairplots allowed a better understanding of the distribution and relationships between features. The elbow method confirmed the suitability of three clusters, which was examined by a silhouette score of 58.81%, indicating a fair cluster fit.

2. Result

The clustering resulted in 2,683 songs being categorized into three clusters, which suggest potential genre groupings:

- **Cluster 0:** Characterized by lower loudness and high speechiness, this cluster likely represents genres such as hip-hop, rap, or spoken word. Low instrumentalness suggests a strong vocal presence.
- **Cluster 1:** High loudness and low speechiness hint at louder mastered music such as rock or pop, while the low instrumentalness points to mainstream vocal tracks.
- **Cluster 2:** This cluster's wide loudness range and high instrumentalness may encompass genres with dynamic variations and instrumental focus, such as classical or jazz.

3. Next steps/Difficulties

The insights gained from clustering can guide personalized music recommendations. However, the issue of unconverted rows remains. It is necessary to validate the uniqueness of each "ID" to ensure accurate clustering and recommendations.