

AMAZON MUSIC CLUSTERING

Unsupervised ML | K-Means | PCA | Feature Engineering



BY ..PRAVEEN....

PROBLEM STATEMENT:

- Automatically group Amazon music tracks based on audio features such as tempo, energy, danceability, etc.
- Useful for playlist generation, recommendations, and genre inference.

INTRODUCTION:

Objective:

To group similar songs based on their audio characteristics using unsupervised clustering.

- *Understand natural groupings in music*
- *Help in playlist creation*
- *Support recommendation engines*

“Instead of manually tagging songs, this project allows the system to automatically discover groups of similar-sounding tracks.”



WORKFLOW OF THE PROJECT:



Made with  Napkin

This workflow ensures the project flows from raw data → clean data → meaningful clusters.

DATA EXPLORATION & PREPROCESSING:

Steps Followed:

- *Loaded dataset into Pandas*
- *Checked column names, dtypes, shape*
- *Handled missing values and duplicates*
- *Converted followers and release_year columns*
- *Dropped non-useful columns*

- **track_id**
- **track_name**
- **artist_name**

“We clean unnecessary columns because clustering only depends on numerical audio features.”



FEATURE SELECTION:

Selected Audio Features:

- Danceability
- Energy
- Loudness
- Speechiness
- Acousticness
- Instrumentalness
- Liveness
- Valence
- Tempo
- Duration_ms

Reason:

“These features describe rhythm, mood, energy, instrumentation, and musical character.”

DATA SCALING:

Method Used:

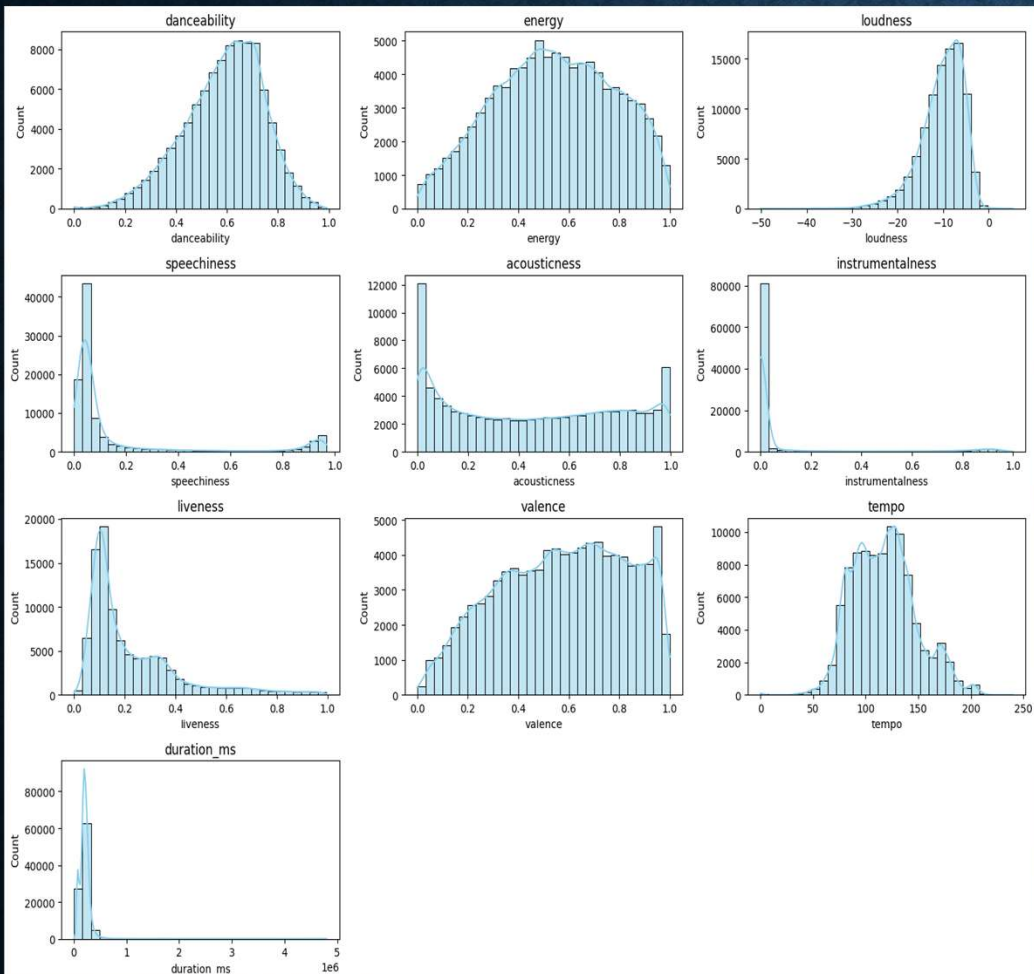
StandardScaler()

- Centers all features
- Ensures fair comparison in distance-based clustering

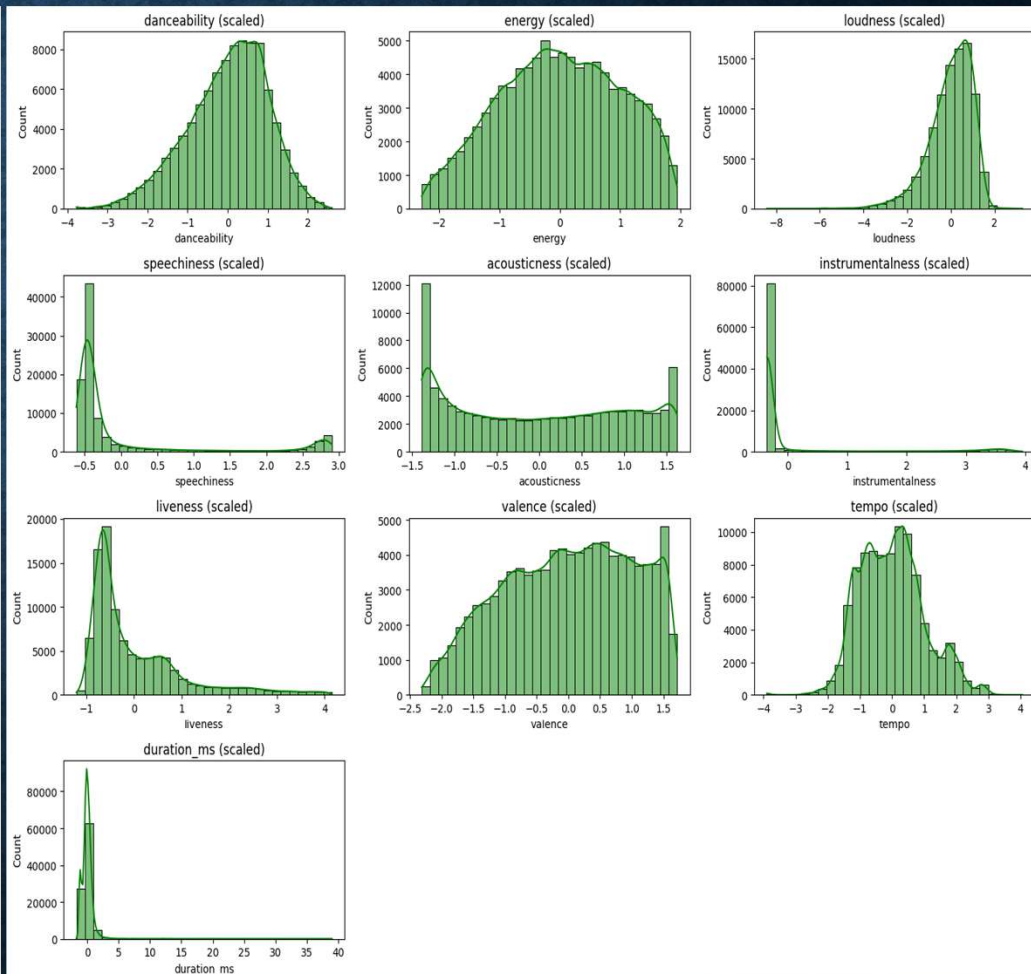
Presenter Note:

“K-Means uses Euclidean distance. So scaling is mandatory.”

BEFORE SCALING



AFTER SCALING



DIMENSIONALITY REDUCTION (PCA)

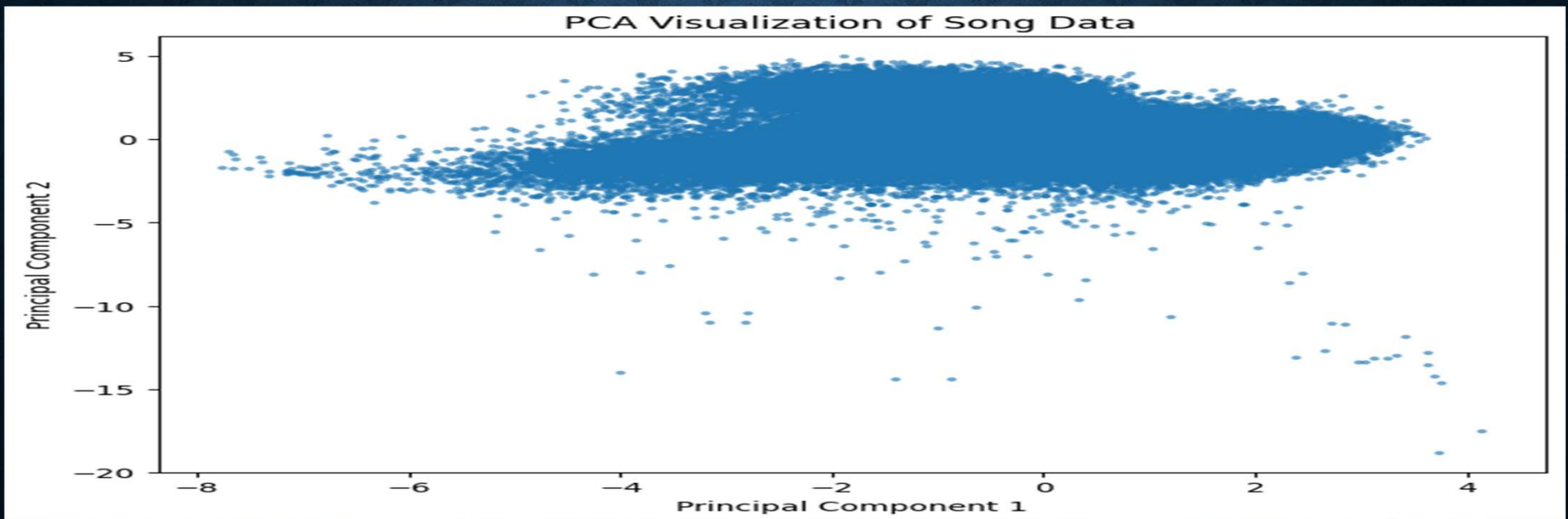
Purpose:

To visualize high-dimensional audio data in 2D.

Technique Used: PCA

Presenter Note:

“PCA is not used for clustering input; only for visualization on graphs.”



CLUSTERING APPROACH

Algorithm Used → K-Means

Reason:

- Simple
- Fast
- Works well on numeric data

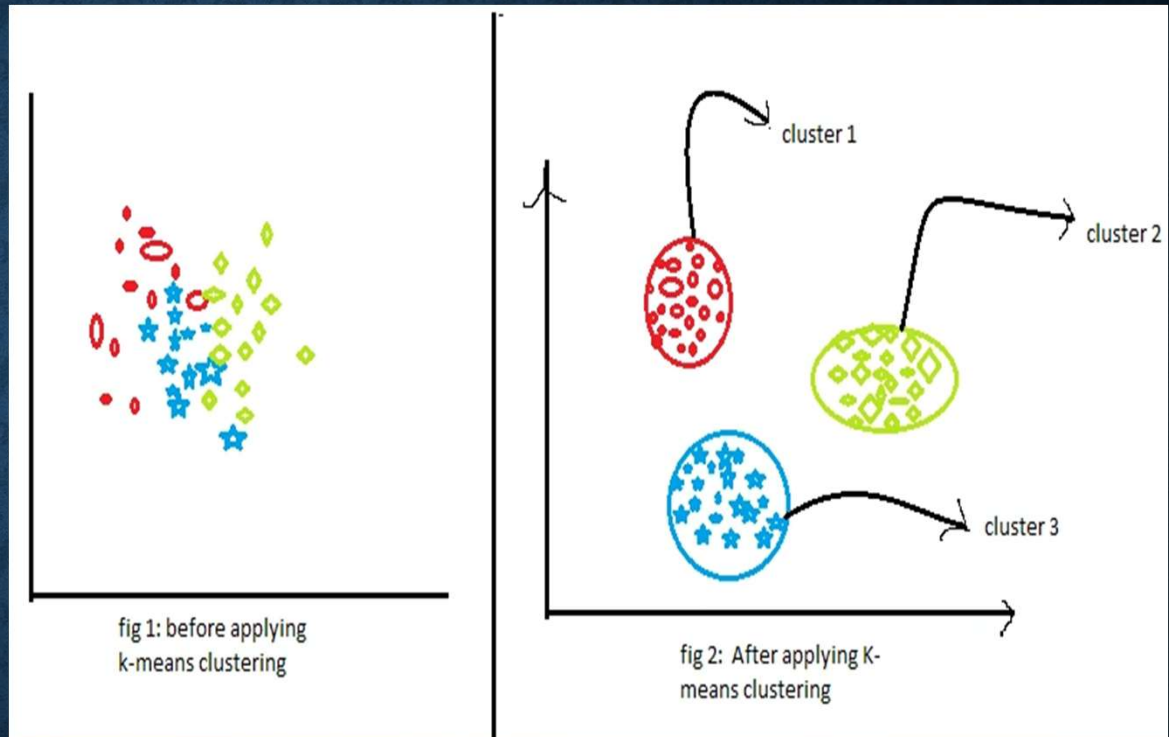
Finding Best K

- Used Elbow Method
- Used Silhouette Scores
- Evaluated DB Index + Inertia

Final Chosen K = 3 Clusters

Presenter Note:

“I tested multiple k values, and based on silhouette score and elbow method, 3 clusters performed the best.”

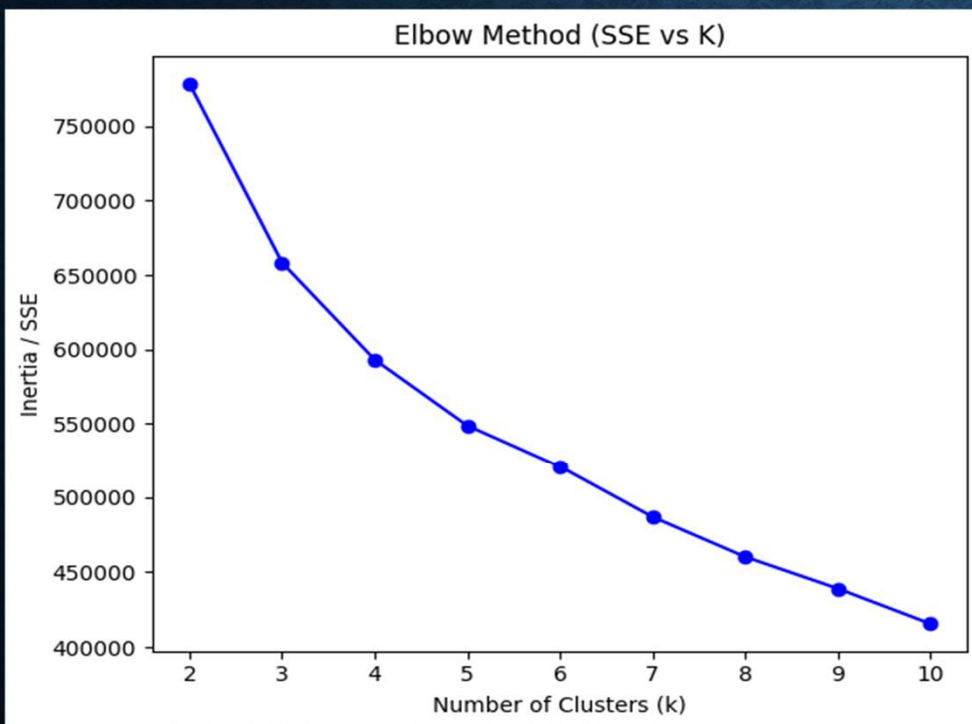


ELBOW METHOD GRAPH:

Explanation:

Inertia decreases as k increases

The “bend” or “elbow” around $k = 3$ suggests the optimal cluster count



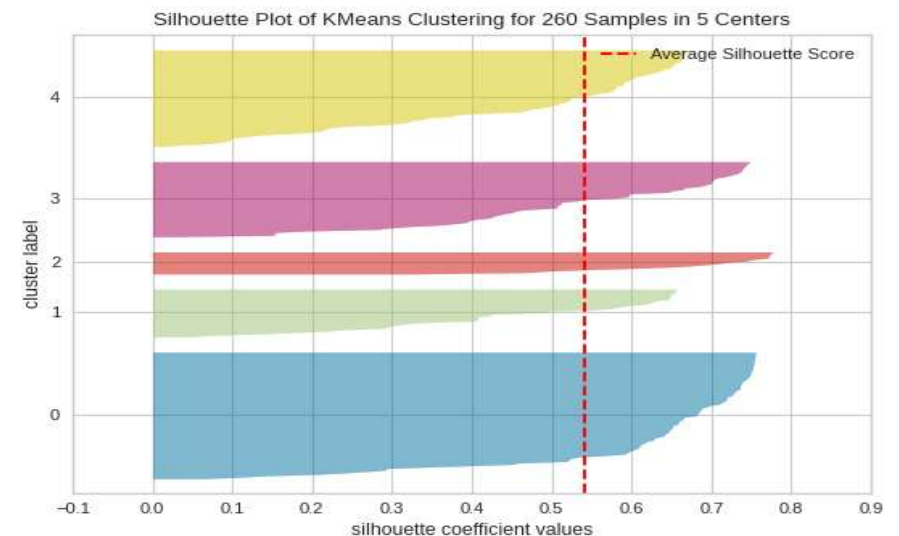
EVALUATION METRICS:

Metrics Calculated:

- **Silhouette Score** → cluster compactness
- **Davies-Bouldin Index** → separation
- **Inertia** → distance within clusters

Presenter Note:

“These metrics help validate whether clusters are meaningful.”



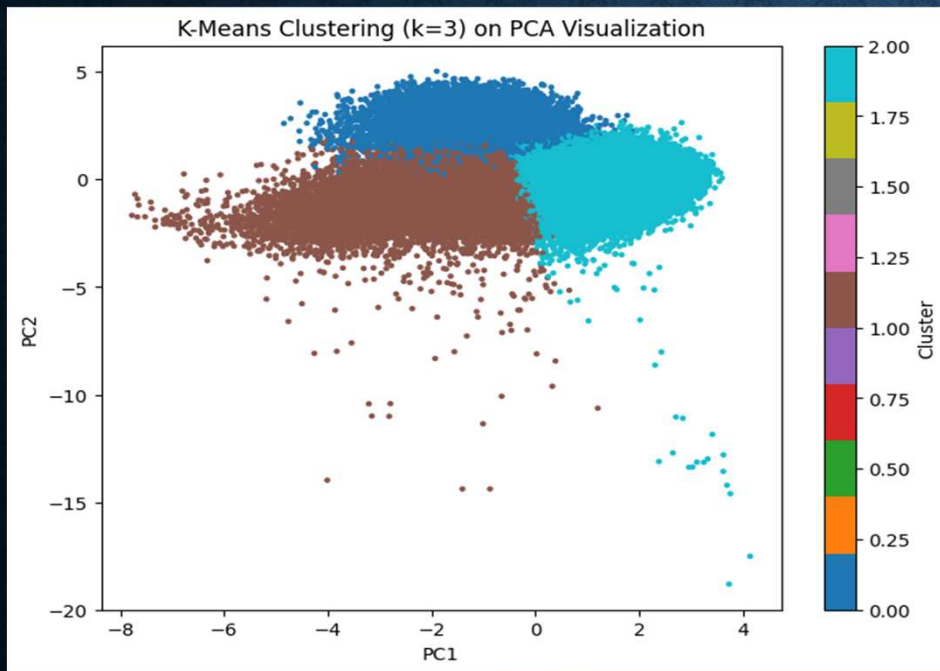
PCA CLUSTER VISUALIZATION

What it shows:

- Songs grouped into 3 clear clusters
- Distinct regions in 2D feature space

Presenter Note:

“This visual proof shows our clustering has succeeded.”



CLUSTER INTERPRETATION

Cluster 0:

High danceability

High energy

→ Party / Dance Tracks

Cluster 1:

High acousticness

Low energy

→ Chill / Acoustic Songs

Cluster 2:

Balanced energy

Mid valence

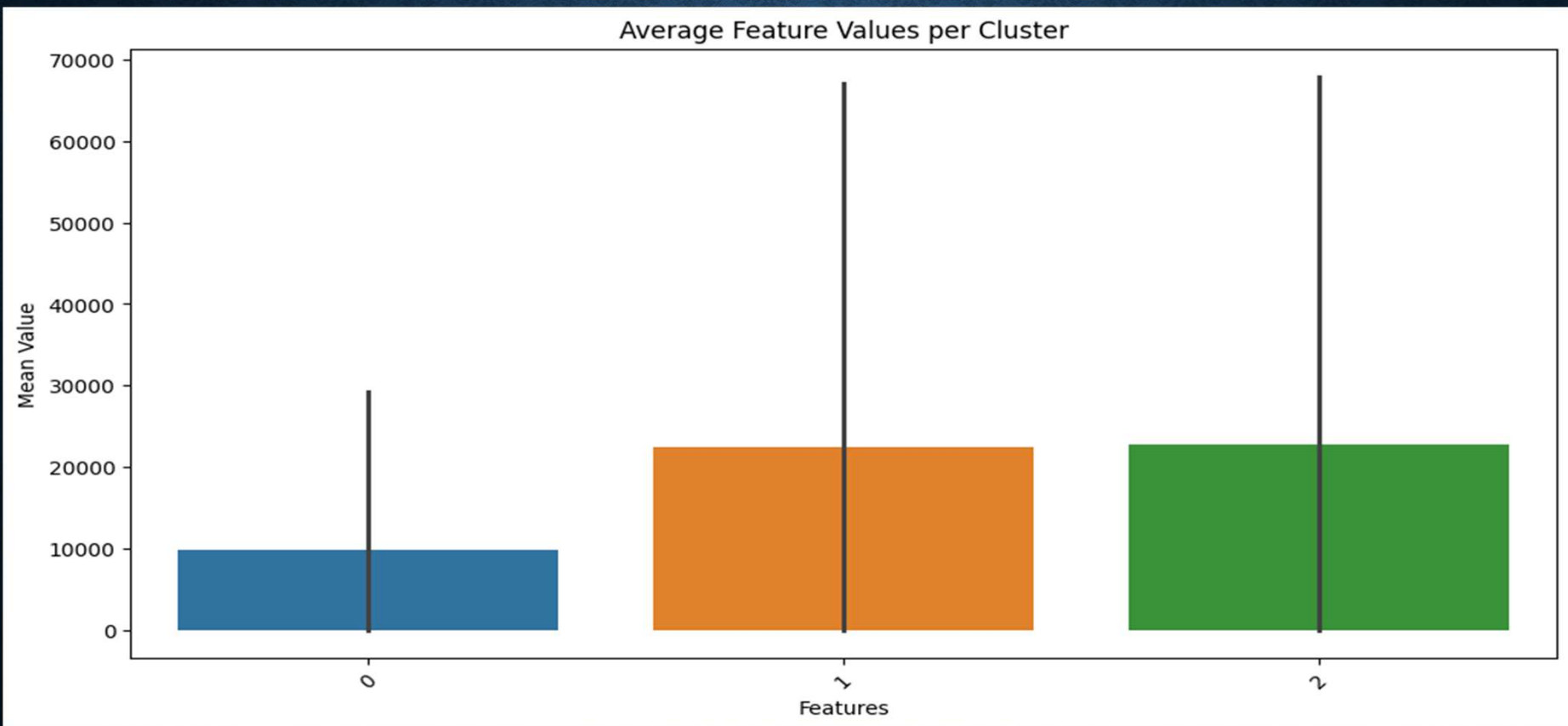
→ General Pop / Mixed Mood Tracks

Presenter Note:

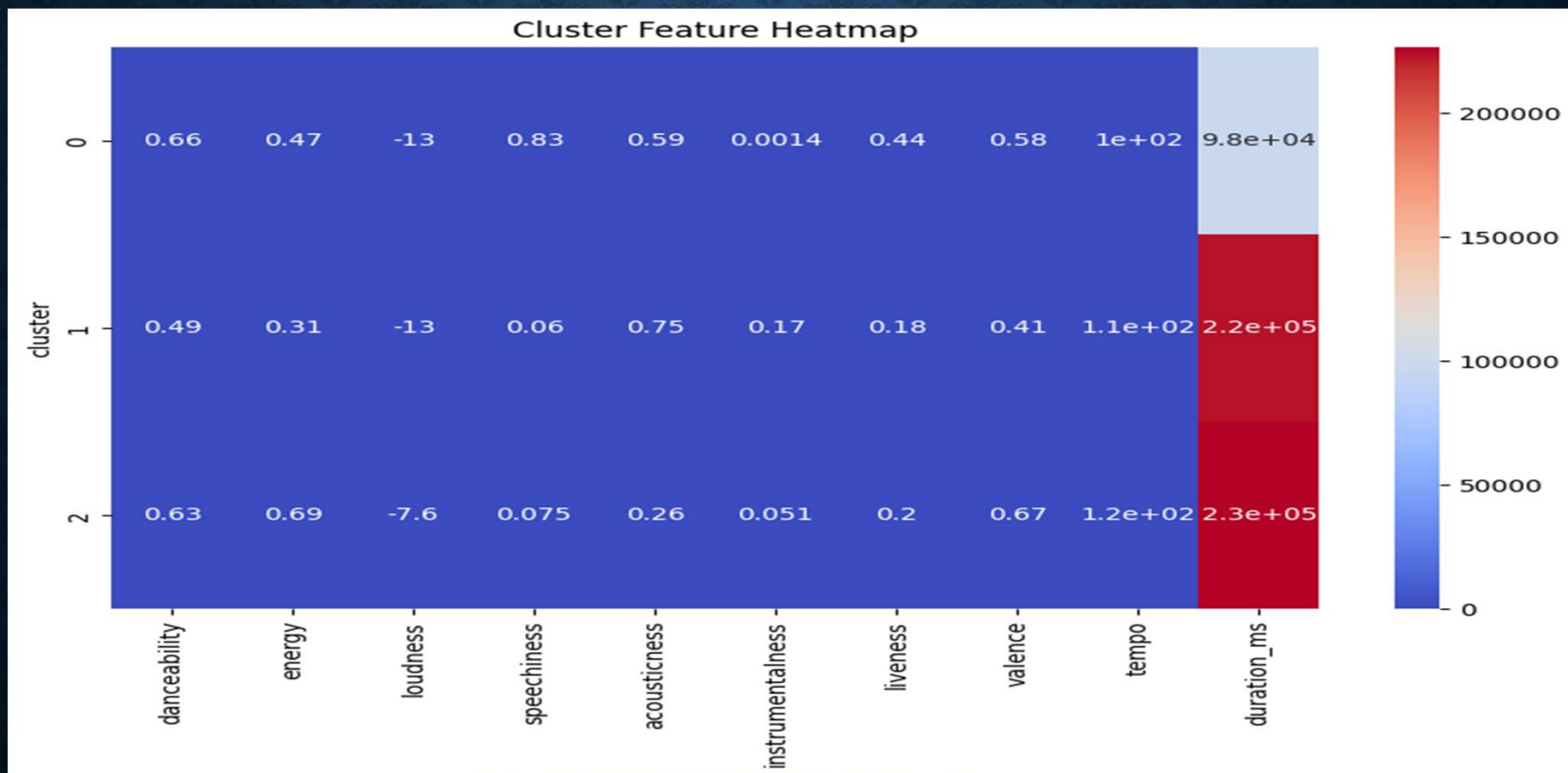
“These descriptions help turn mathematical clusters into real music categories.”

FEATURE COMPARISON

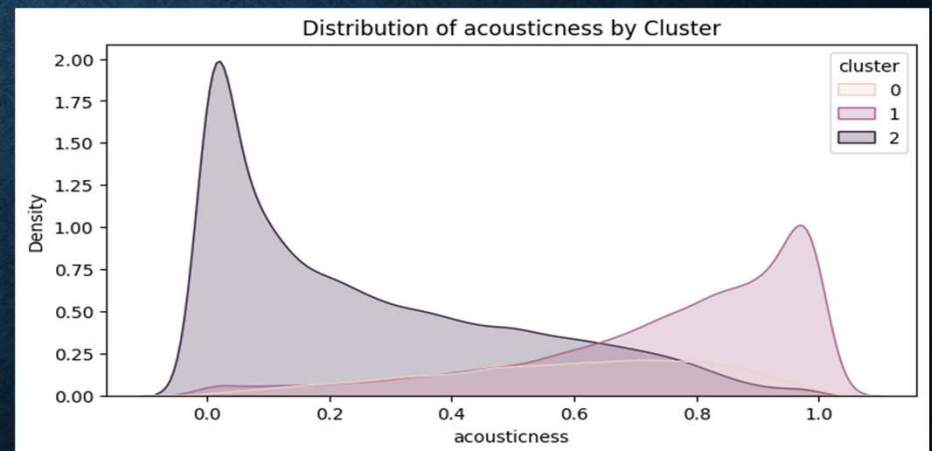
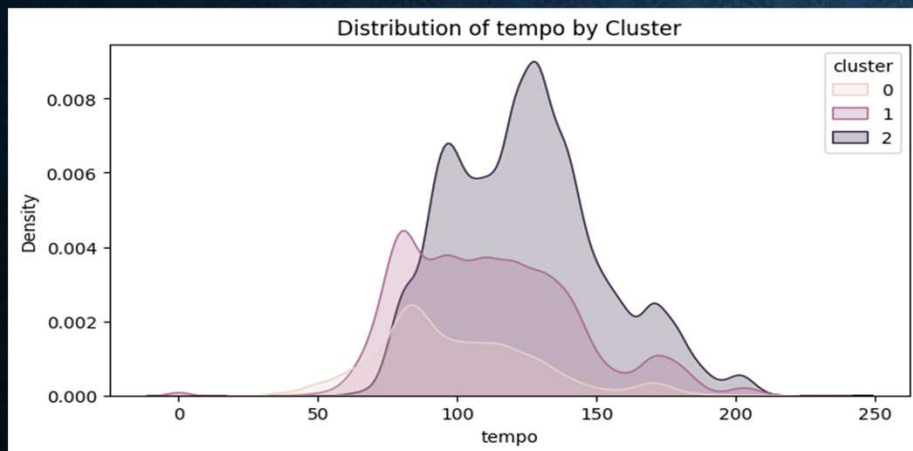
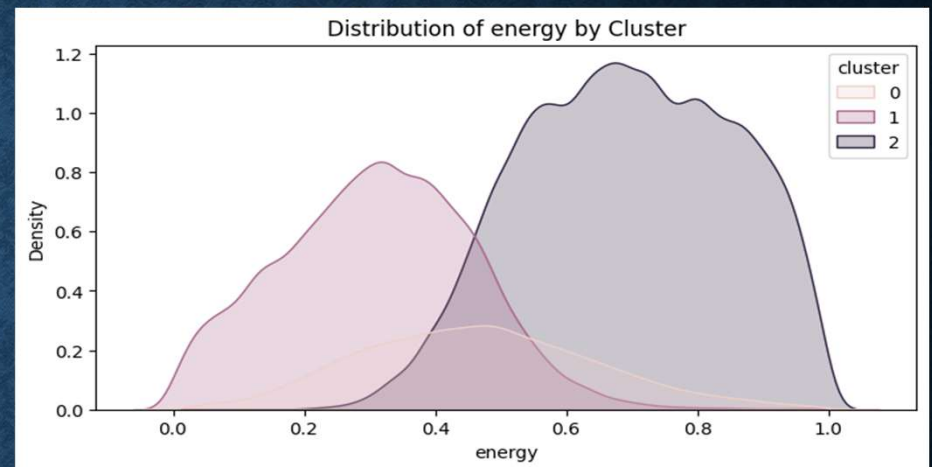
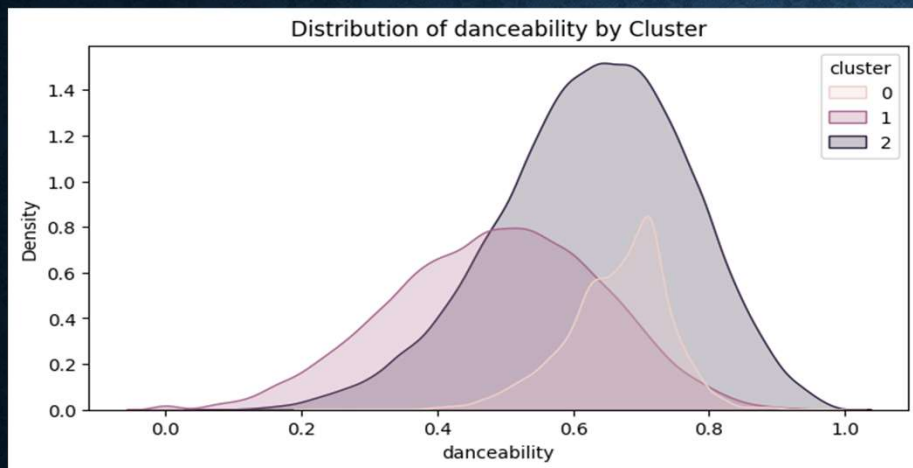
Bar Chart



HEATMAP — FEATURE COMPARISON ACROSS CLUSTERS:



DISTRIBUTION PLOTS — FEATURE DISTRIBUTION PER CLUSTER:



OVERALL SUMMARY

Model successfully grouped songs into three meaningful categories:

Party / Upbeat Music

Chill / Emotional Acoustic Songs

Instrumental / Ambient Tracks

Perfect for:

- **Personalized playlist creation**
- **Music recommendations**
- **Mood-based song discovery**

FINAL RESULT :

The project successfully demonstrates how unsupervised learning can organize large music libraries into meaningful groups, improving music discovery and enhancing user experience on streaming platforms.



THANK YOU