# Amazon Music Clustering Project

## Objective

The goal of this project is to use unsupervised machine learning to automatically group songs that share similar musical characteristics. By analysing each song's rhythmic, acoustic, and emotional features, the project aims to uncover hidden patterns that could be used for playlist generation, mood detection, or music recommendation system

## Dataset Overview

The dataset (single_genre_artists.csv) contained various audio features (e.g., energy, danceability, valence, loudness, tempo) describing the sonic and emotional qualities of songs.

## Methodology

The workflow followed a complete machine-learning pipeline:

- **Data Preprocessing** – Checked for missing values, duplicates, and data types; removed identifiers.
- **Feature Scaling** – Standardized all numeric features using StandardScaler.
- **Clustering Algorithms** – Applied both **K-Means**.
- **Model Evaluation** – Compared Silhouette Score and Davies–Bouldin Index to judge cluster quality.
- **Visualization** – Used PCA 2-D scatter plots, bar charts, heatmaps, and boxplots to interpret cluster structure

## Results & Cluster Analysis

### K-Means (k = 3)

- **Silhouette Score**: 0.242
- **SSE (Elbow Method):** Showed a strong elbow at k = 3, indicating optimal compactness.
- Three well-separated groups captured distinct song moods.

| Cluster | Dominant Traits | Interpretation |
|---|---|---|
| K 0 | High energy, high danceability, positive valence | Upbeat / Pop songs |
| K 1 | High acousticness, lower tempo and energy | Acoustic / Calm songs |
| K 2 | High loudness and energy | Energetic / Rock or EDM songs |

## Key Visual Insights

- PCA Plots: Showed distinct cluster separation.
- Heatmaps: Highlighted contrasting features such as acousticness vs. energy.
- Boxplots: Displayed feature distribution differences for each cluster.
- Dendrogram: Confirmed three major branches with strong distance separation.

## Final Deliverables

- Source Code: Preprocessing, Clustering Implementation, Visualization (clean modular notebook).
- CSV Outputs:
  - clustered_songs_kmeans_k3.csv
  - clustered_songs_final_combined.csv
- Cluster Summary: cluster_summary_kmeans.csv
- Visualizations: PCA, heatmap, and dendrogram images.

## Conclusion

Both K-Means consistently identified three natural musical groups in the dataset:

1. Energetic & danceable tracks (Pop / Electronic)
2. Calm & acoustic songs
3. Loud & high-energy performances (Rock / EDM)