Amazon Music Clustering – Project Report

1. Introduction

With millions of songs being added to streaming platforms, manually categorizing music into genres is inefficient. The goal of this project is to automatically group Amazon Music tracks into meaningful clusters using **unsupervised machine learning**, purely based on audio features such as tempo, danceability, energy, loudness, and more.
This helps identify similar songs, support playlist generation, and improve music discovery.

---

2. Objectives

- Understand and preprocess the audio-features dataset.
- Select relevant features that describe the musical characteristics.
- Normalize feature values before clustering.
- Use **K-Means** to cluster songs based on similarity.
- Determine optimal number of clusters using **silhouette score**.
- Visualize clusters using **PCA**.
- Interpret the characteristics of each cluster.

---

3. Dataset Description

The dataset contains audio characteristics of Amazon Music tracks.
Key columns used for clustering:

- danceability
- energy

- loudness
- speechiness
- acousticness
- instrumentalness
- liveness
- valence
- tempo
- duration_ms

These features describe rhythm, mood, acoustic properties, and intensity of a song.

(If metadata exists: also includes track_id, track_name, artist_name.)

---

4. Data Preprocessing

Steps performed:

- Removed unnecessary metadata columns for clustering (keeping them only for reference).
- Checked and removed missing values in audio features.
- Verified data types and dataset size.
- Performed feature distribution analysis (EDA).
- Scaled the dataset using **StandardScaler** to normalize value ranges.

---

5. Feature Selection

Only features related to the **musical characteristics** were selected.
These features represent:

- Dance feel → `danceability`, `energy`, `tempo`
- Mood → `valence`, `loudness`
- Acoustic behavior → `acousticness`, `instrumentalness`
- Vocal/genre intensity → `speechiness`, `liveness`

These 10 features formed the input for clustering.

---

6. Determining Optimal Number of Clusters

To determine the best number of clusters (`k`):

Fast Method Used

- Reduced data to 5 PCA components temporarily for faster computation.
- Used **MiniBatchKMeans** for quick inertia + silhouette score testing.
- Performed a coarse search (k = 2,4,6,8,10).
- Refined the search around the best k (±2).
- Selected final `k` based on highest silhouette score.

Final Selected k:

k = *YOUR BEST_K HERE*
(Replace with the value printed by your fast search cell.)

---

7. Final Clustering Using K-Means

After deciding the optimal k, the final **KMeans** algorithm (with `n_init=20`) was trained on the full scaled dataset.

The resulting cluster labels were appended to the dataframe as:

`cluster`

Each row now belongs to one cluster.

---

8. Cluster Interpretation

For each cluster, mean feature values were computed.

Typical interpretations:

- **Cluster X:** High energy, high danceability → Party / EDM
- **Cluster X:** High acousticness, low loudness → Chill / Acoustic
- **Cluster X:** High speechiness → Rap / Hip-hop
- **Cluster X:** High instrumentalness → Instrumental / Ambient

(Replace these based on your actual cluster_profile table.)

This helps understand what each group of songs represents.

---

9. Visualization

PCA 2D Scatter Plot

A 2-component PCA transformation was applied for visualization.
The scatter plot showed clear separation among clusters, indicating:

- Groups of similar-sounding songs form naturally
- The cluster decision boundary is meaningful
- Musical characteristics influence grouping patterns

---

10. Final Export

The final dataset exported:

`amazon_music_clusters_final.csv`

This CSV contains:

- All selected features
- `cluster`
- PCA components (`pca1`, `pca2`)
- Any metadata columns (if present)

This file can be used for:

- Playlist generation
- Building recommendation engines
- Further genre inference tasks

---

11. Conclusion

This project successfully grouped Amazon Music tracks into distinct clusters based on audio characteristics.
By applying KMeans clustering and PCA visualization, we identified meaningful patterns representing different moods, genres, and musical styles.

The clustering results can support:

- Personalized playlist creation
- Music recommendation systems
- Market segmentation
- Artist similarity analysis

This workflow also establishes a reusable, scalable approach for future music analytics tasks.

📌 End of Report