Data Mining Project Report:
Music "Vibe" Segmentation Using Clustering on Spotify Dataset
Author : Justin Matthew T
Student ID : A11.2023.14877

---

1. Main Objective of the Analysis

The main objective of this analysis is to perform **segmentation** (**clustering**) on Spotify's song catalog based on their intrinsic audio characteristics. This analysis combines two Unsupervised Learning techniques: **Clustering** as the primary goal and **Dimension Reduction** as the supporting method.

Specifically, **Dimension Reduction (PCA)** is first applied to simplify 7 complex audio features into 2 main components. Then, Clustering (K-Means) is applied to these simplified components to discover meaningful and visually interpretable "vibe" segments of songs.

The business benefits of this analysis include improving user experience and content strategy through:

1. **Better Music Recommendations**: By understanding the "vibe" of a song (e.g., "High Energy" vs. "Chill Acoustic"), we can recommend musically similar tracks, even across different artists or genres.
2. **Automated Playlist Generation**: Enables smarter and more cohesive playlist creation (e.g., "Focus Playlist" or "Party Playlist") automatically.
3. **Content Insights**: Provides deeper understanding of the "audio recipe" of various music segments on the platform.

---

2. Data Description and Initial Exploration

Dataset Download: https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

Colab Link:
https://colab.research.google.com/drive/1CPE5_yT7bLUTB5aeUlRQv9JjEWo1tO44#scrollTo=_uRg5IyJRrH8

The chosen dataset is **"Spotify Tracks DB" (SpotifyFeatures.csv)**, a comprehensive database containing metadata and audio features extracted by Spotify for thousands of songs.

• **Key Attributes**:
This analysis focuses on 7 primary numerical audio features that define a song's "feel":

- **acousticness**: Confidence measure (0.0–1.0) indicating whether a track is acoustic.
- **danceability**: Measure (0.0–1.0) indicating how suitable a track is for dancing, based on tempo, rhythm stability, and beat strength.
- **energy**: Measure (0.0–1.0) representing intensity and activity. Energetic tracks feel fast, loud, and noisy.

- **instrumentalness**: Measure (0.0–1.0) predicting whether a track lacks vocals.
- **liveness**: Measure (0.0–1.0) detecting the presence of an audience (indicating a live recording).
- **loudness**: Overall sound level in decibels (dB), typically ranging from -60 to 0 dB.
- **speechiness**: Measure (0.0–1.0) detecting the presence of spoken words. High values indicate non-musical content (e.g., podcasts, audiobooks).

• **Objective**:
Use these 7 audio features to cluster songs into sonically homogeneous segments, without using genre or artist information.

• **Data Size**:
The full dataset contains hundreds of thousands of entries. For efficiency and computational performance, a random sample of 20,000 tracks was used for this project (original dataset size: **232,725 tracks**).

---

3. Data Preparation (Data Cleaning & Feature Engineering)

The following data preparation steps were performed before modeling:

1. **Cleaning**: Duplicate entries based on track_id and rows with missing (NaN) audio feature values were removed.
2. **Feature Selection**: Only the 7 relevant audio features (listed above) were selected as model inputs.
3. **Standardization** (**Scaling**):
   This crucial step standardized all 7 audio features using StandardScaler (from Scikit-learn), transforming them to have mean = 0 and standard deviation = 1.
   This ensures that features like *loudness* (range -60 to 0) do not dominate features like *danceability* (range 0 to 1).

---

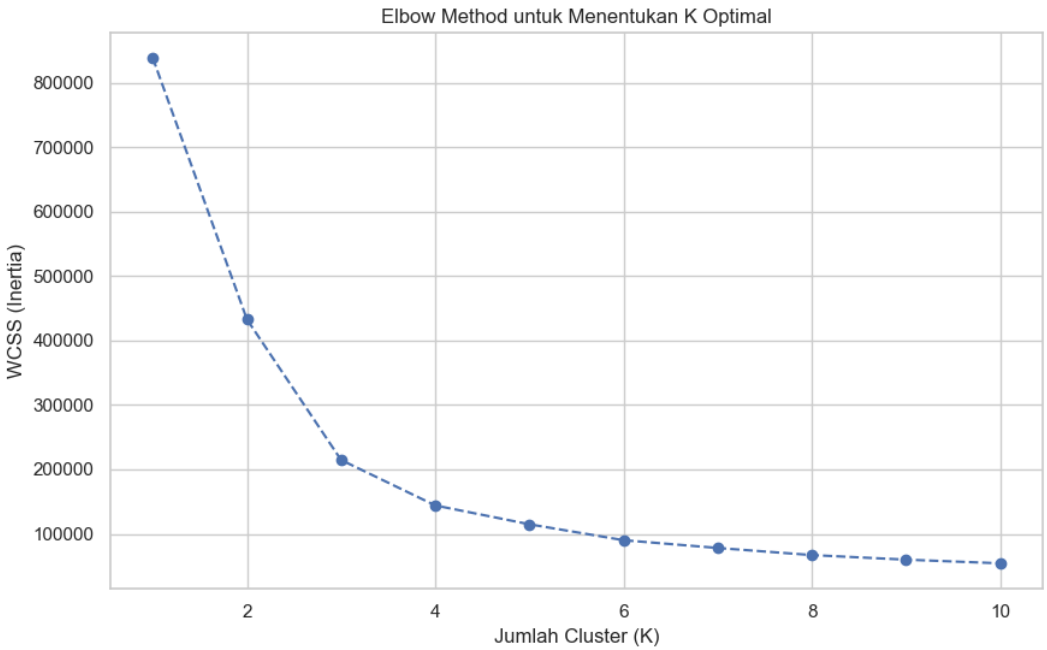
4. Model Training Summary

To achieve optimal and visualizable clustering results, a two-step approach was used:

1. **Dimension Reduction (PCA):**
   The 7 standardized features were reduced into 2 Principal Components (PCA) for 2D visualization.
   These two components captured 66.30% of the total data variance.
2. **Clustering (K-Means):**
   The K-Means model was applied on the 2 PCA components.

Three model variations were trained to determine the optimal number of clusters (K), supported by two evaluation methods:

1. **Experiment 1 – Elbow Method:**
   The *Elbow Plot* showed that WCSS (inertia) started flattening significantly after K=4 or K=5, indicating 4, 5, or 6 as potential cluster counts.



2. Experiments 2, 3, 4 – Silhouette Score Comparison:
K-Means models with K=4, K=5, and K=6 were compared using their Silhouette Scores (higher is better).

Model (Variation)	Silhouette Score
K-Means with K=4	0.4798 (BEST)
K-Means with K=5	0.4507
K-Means with K=6	0.4287

---

5. Final Model Recommendation

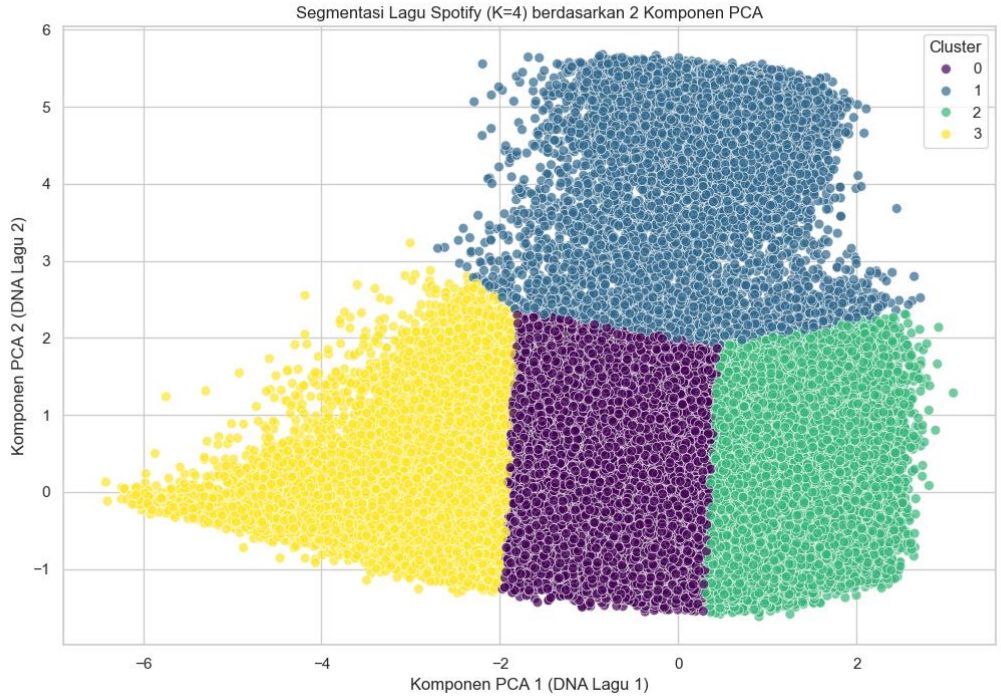
Based on the comparison above, **K-Means with K=4 is recommended as the final model**.

Reason:
This model achieved the **highest Silhouette Score (0.4798),** indicating the most statistically optimal cluster separation. Additionally, these 4 clusters yielded four distinct, meaningful, and actionable “song personas,” described below.

---

6. Key Findings and Insights

The analysis successfully identified 4 distinct "vibe" segments from 20,000 tracks.



cluster_label	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness
0	0.124195531	0.623630605	0.73836009	0.038587227	0.200991002	-6.024179298	0.103229606
1	0.539906433	0.542638344	0.427977089	0.180438039	0.164529901	-10.70484926	0.066576771
2	0.895569432	0.306712781	0.143565869	0.580703384	0.155902428	-20.73804106	0.050170689
3	0.728573632	0.551012948	0.666997211	0.003048721	0.741231574	-11.43844522	0.782036653

• Cluster 0: "High Energy / Mainstream Party Tracks"

- Profile: Very high energy (0.74) and loudness (-6.02), with high danceability (0.62). Very low acousticness (0.12).
- Insight: Represents commercial pop, EDM, and dance tracks that dominate party playlists.

• **Cluster 1: "Chill Acoustic / Relaxed Vibes"**

- Profile: Moderately high acousticness (0.54) with low-to-medium energy (0.43) and loudness (-10.70).
- Insight: Represents easy-listening, pop-acoustic, and folk music. Ideal for chill playlists.

• **Cluster 2: "Instrumental Calm / Classical & Ambient"**

- Profile: Extremely high acousticness (0.90) and instrumentalness (0.58), with very low energy (0.14) and loudness (-20.74).
- Insight: Represents calm, quiet instrumental music — perfect for focus, study, or sleep playlists.

• **Cluster 3: "Spoken Word / Live Recordings"**

- Profile: Exceptionally high speechiness (0.78) and liveness (0.74), much higher than other clusters.
- Insight: A fascinating finding — the model isolated non-musical content such as podcasts, stand-up comedy, or live concert recordings into its own segment.

---

7. Limitations and Future Work

While the analysis was successful, several limitations and improvement opportunities were identified:

• **Model Limitations:**

1. **PCA Information Loss**:
   Using PCA for visualization (reducing 7 features to 2) causes information loss (~33.70% variance).
   Clusters that appear overlapping in 2D may actually be well-separated in the full 7-dimensional space.
2. **“Spoken Word” Cluster Diversity**:
   Cluster 3 may be too broad, mixing podcasts, comedy, and live concert recordings.

• **Recommended Next Steps:**

1. **Genre Analysis:**
   Conduct deeper analysis within Clusters 0, 1, and 2 to see genre distribution. For example, is the “High Energy” segment dominated by Pop or Hip-Hop?
2. **Sub-Clustering:**
   Take Cluster 3 (“Spoken Word”) and perform sub-clustering on its subset to further separate podcasts from live concerts.
3. **Full-Dimension Clustering:**
   Apply clustering on all 7 original audio features (without PCA) and compare results, even if not directly visualizable.