# Rhythms Through Time:
# A Data-Driven Analysis of Lyrics and Audio Features

An Nguyen, Segun Adewola, Andrew Min. Advisor Prof. Yifan Hu

Northeastern University

## 1. Background

This study analyzes the relationship between audio features and lyrical content using extensive Spotify datasets. By integrating machine learning we aim to make a user tailored recommender system. Additionally, we explore semantic meanings in song lyrics and develop an improved recommender system. This research aims to prototype a recommender system that is more personalized than those in music apps.

## 2. Motivations & Objectives

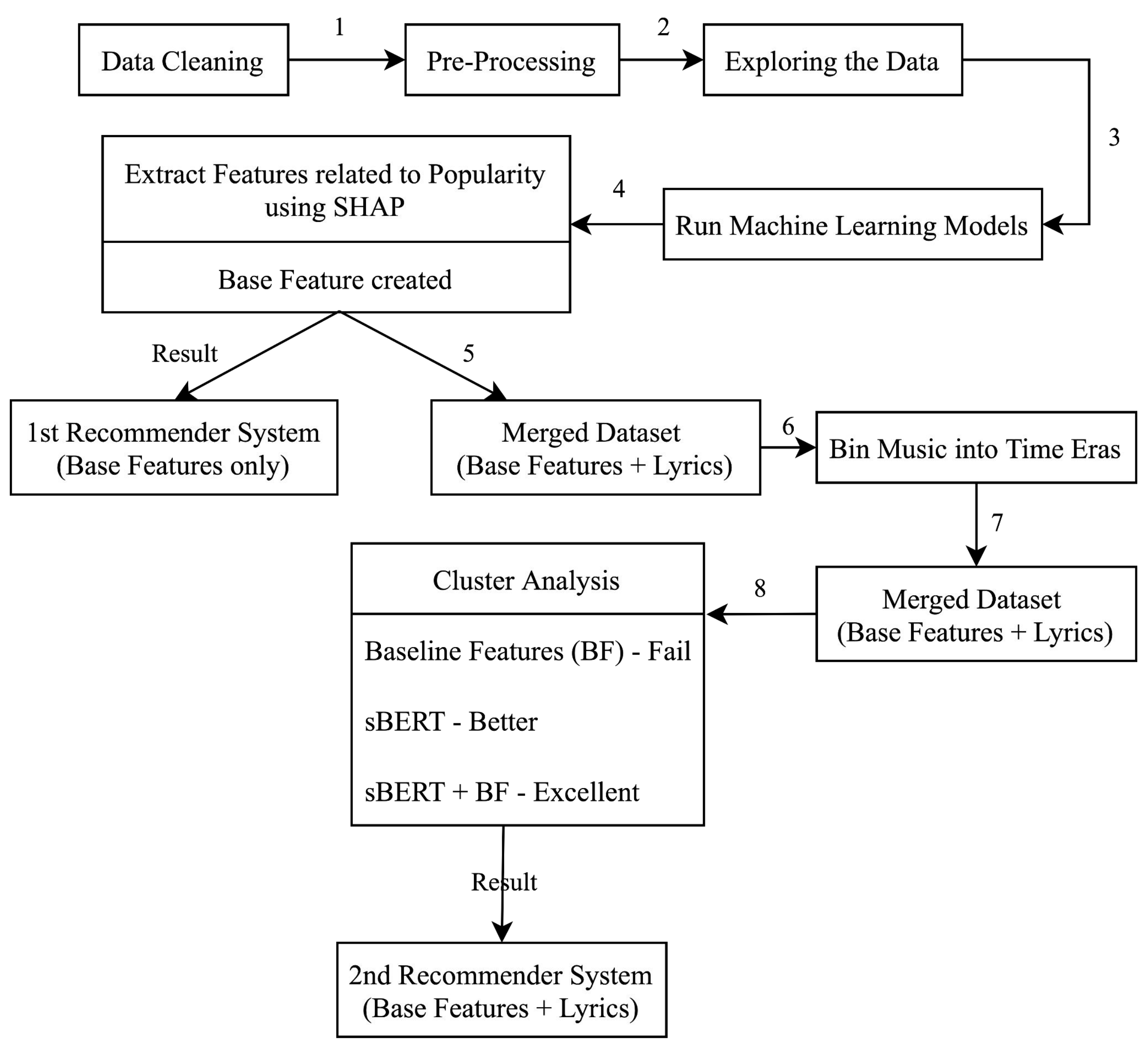
This research has two core objectives: (1) Creating a recommender system tailored to user preferences and (2) analyzing lyrical meanings within lyrics using the sBERT deep learning model. The study motivations include: (a) enhancing user engagement and revenue by increasing user satisfaction (b) create monetization opportunities through targeted advertising from user preferences

## 3. Methods

The study utilized two comprehensive Spotify datasets: one comprising 10,000 songs with audio features and another featuring 57,651 songs with lyrics. Models are ran on data here to predict popularity and the related features are captured using SHAP analysis. These based features are used for the 1st recommender.

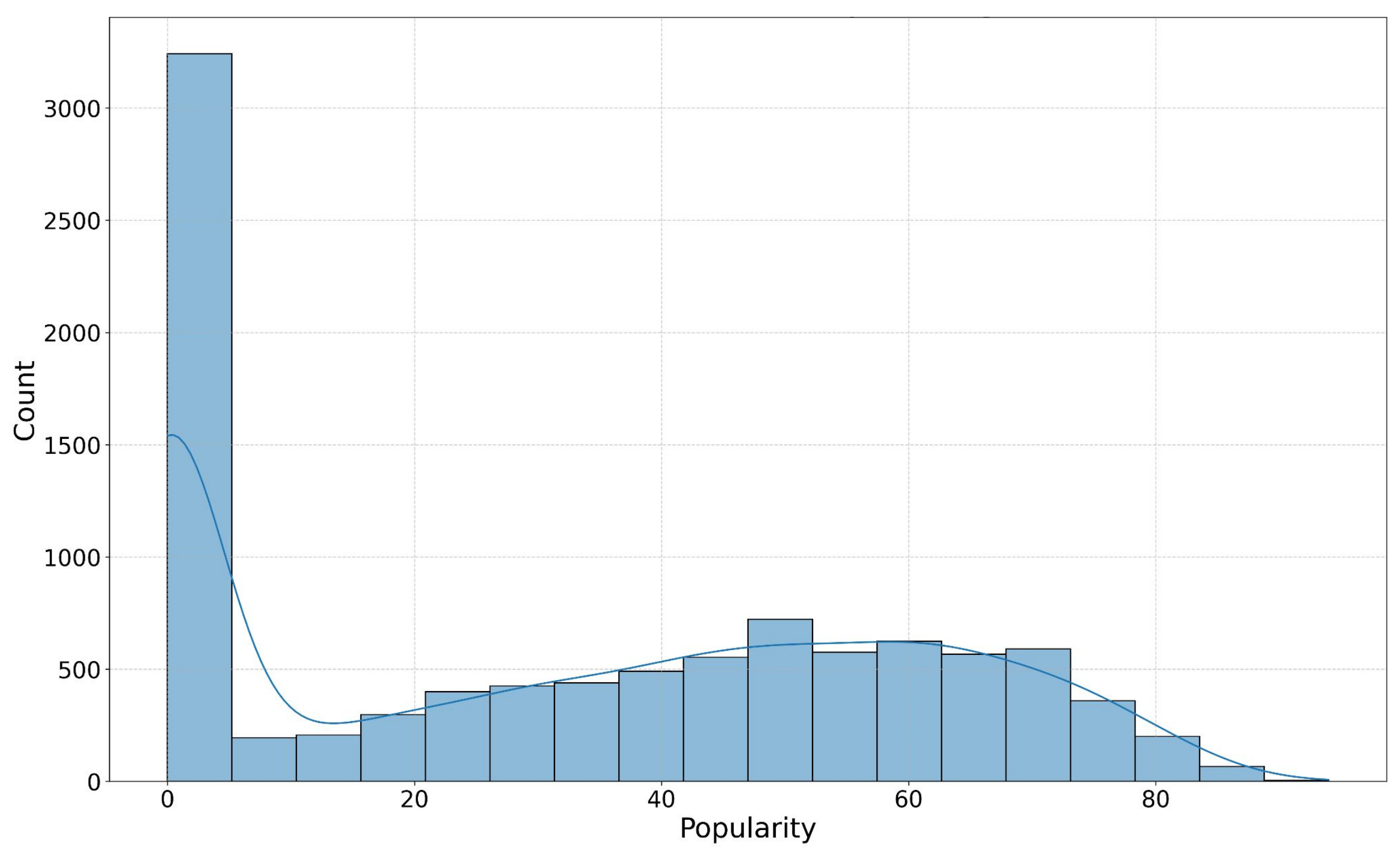
The merged dataset required extensive preprocessing and after cleaning, the final dataset contained 1,634 entries, blending audio features and lyrics. sBERT is ran on this dataset and clustering analysis is used to determined the best combinations of features. These features are used for the 2nd recommender.

**Below is a flow chart of our processes and data flow.**

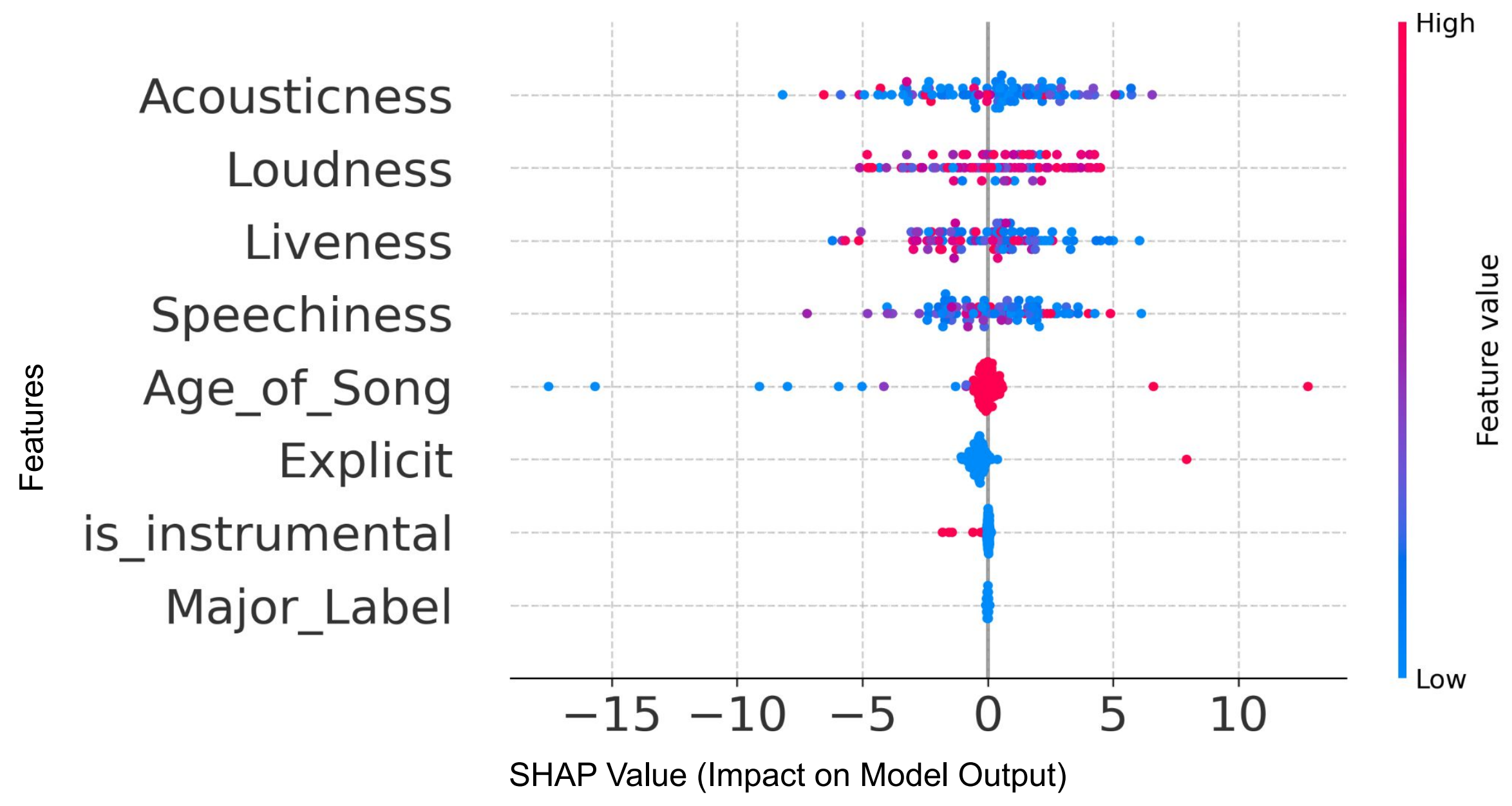

Data Cleaning
1
Pre-Processing
2
Exploring the Data
3
Run Machine Learning Models
4
Extract Features related to Popularity using SHAP
Base Feature created
Result
1st Recommender System (Base Features only)
5
Merged Dataset (Base Features + Lyrics)
6
Bin Music into Time Eras
7
Merged Dataset (Base Features + Lyrics)
8
Cluster Analysis
Baseline Features (BF) - Fail
sBERT - Better
sBERT + BF - Excellent
Result
2nd Recommender System (Base Features + Lyrics)

## 4. Data

Left skewness of 10000 songs Database when looking at popularity label



Utilizing SHAP Analysis to define Selected Features



## 4. Experimental Result

Regression models on average achieved an R² of ~0.5, explaining about 50% of popularity variance. Clustering improved in selected features enriched with sBERT embeddings.

Table 1: Regression Model Performance Metrics

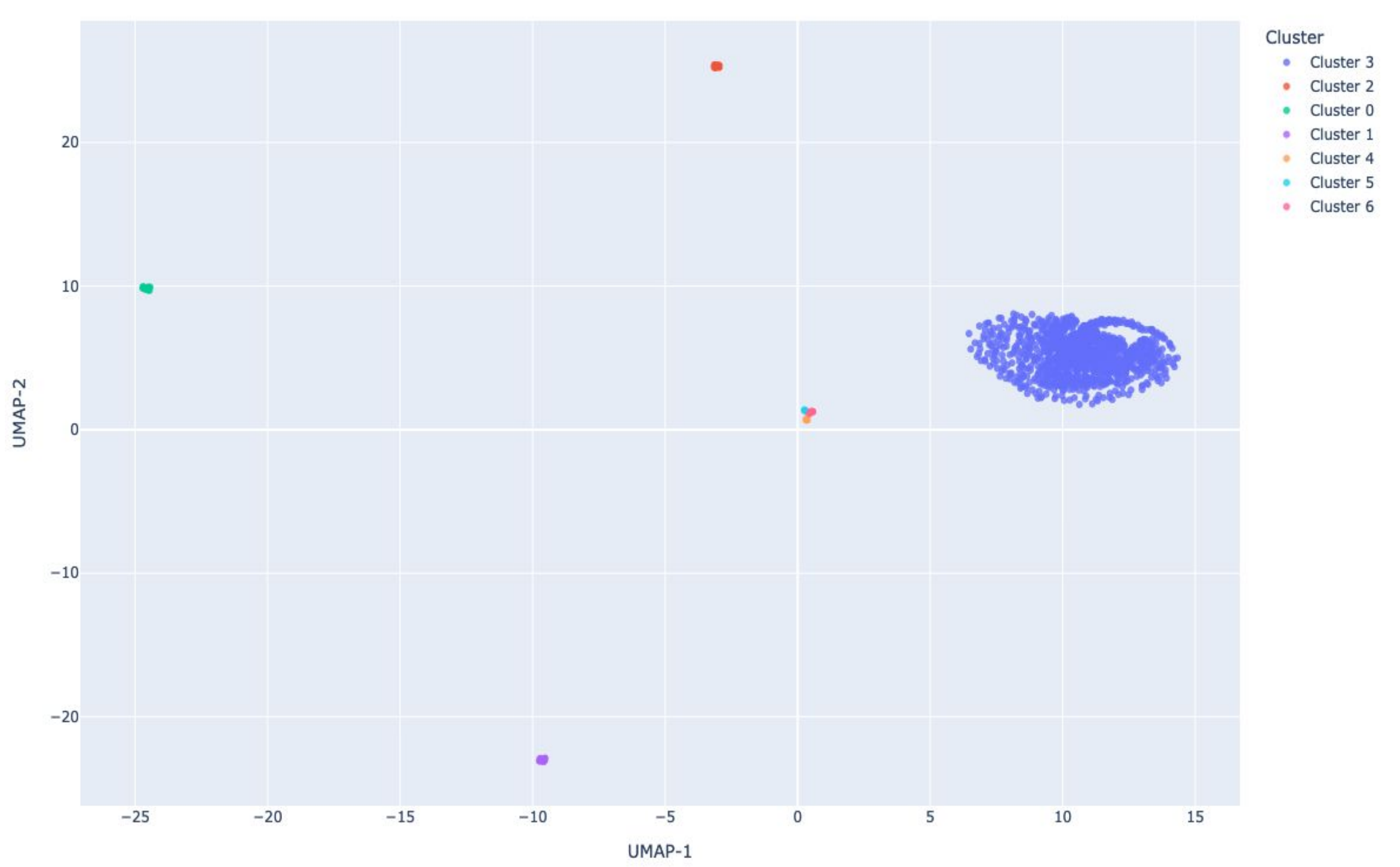
| Model | MSE | R² |
|---|---|---|
| Linear Regression (Including Artist_Popularity) | 368.78 | 0.512 |
| Ridge Regression (Including Artist_Popularity) | 368.77 | 0.512 |
| Lasso Regression (Including Artist_Popularity) | 368.76 | 0.512 |
| Random Forest Regressor (Including Artist_Popularity) | 473.25 | 0.373 |
| Gradient Boosting Regressor (Including Artist_Popularity) | 380.86 | 0.496 |
| Keras Sequential Model (Including Artist_Popularity) | 444.00 | 0.412 |
| Tuned XGBoost Regressor (Including Artist_Popularity) | 377.18 | 0.501 |
| Tuned LightGBM Regressor (Including Artist_Popularity) | 389.92 | 0.484 |
| **Baseline Linear Regression (Joe Beach Capital, 2023)** | **883.80** | **-0.0053** |
| **Baseline Gradient Boosting (Joe Beach Capital, 2023)** | **881.82** | **-0.0021** |
| **Baseline Random Forest (Joe Beach Capital, 2023)** | **944.95** | **-0.8748** |

Table 2: Clustering Model Performance Metrics

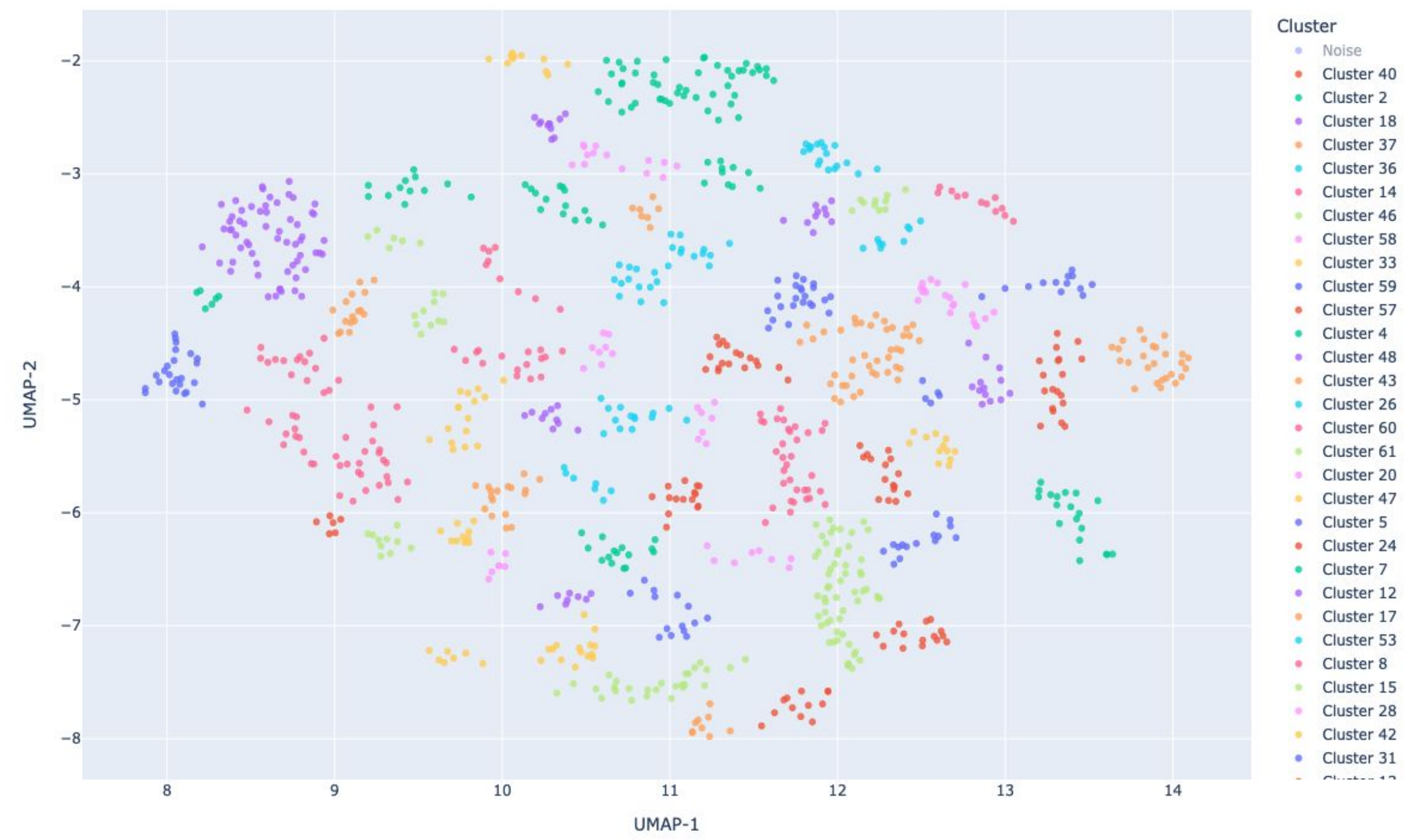
| Model | Silhouette Score |
|---|---|
| **Benchmark Model (Time Period Bin Only)** | **0.8084 (Bad Clusters)** |
| sBERT Model (Time Period Bin Only) | 0.4286 (Decent Clusters) |
| sBERT Enhanced Model (Time Period Bin + Selected Features) | 0.7464 (Excellent Clusters) |

Clustering progression visualizations of time binning baseline, baseline + sbert, and baseline + sbert + selected features.

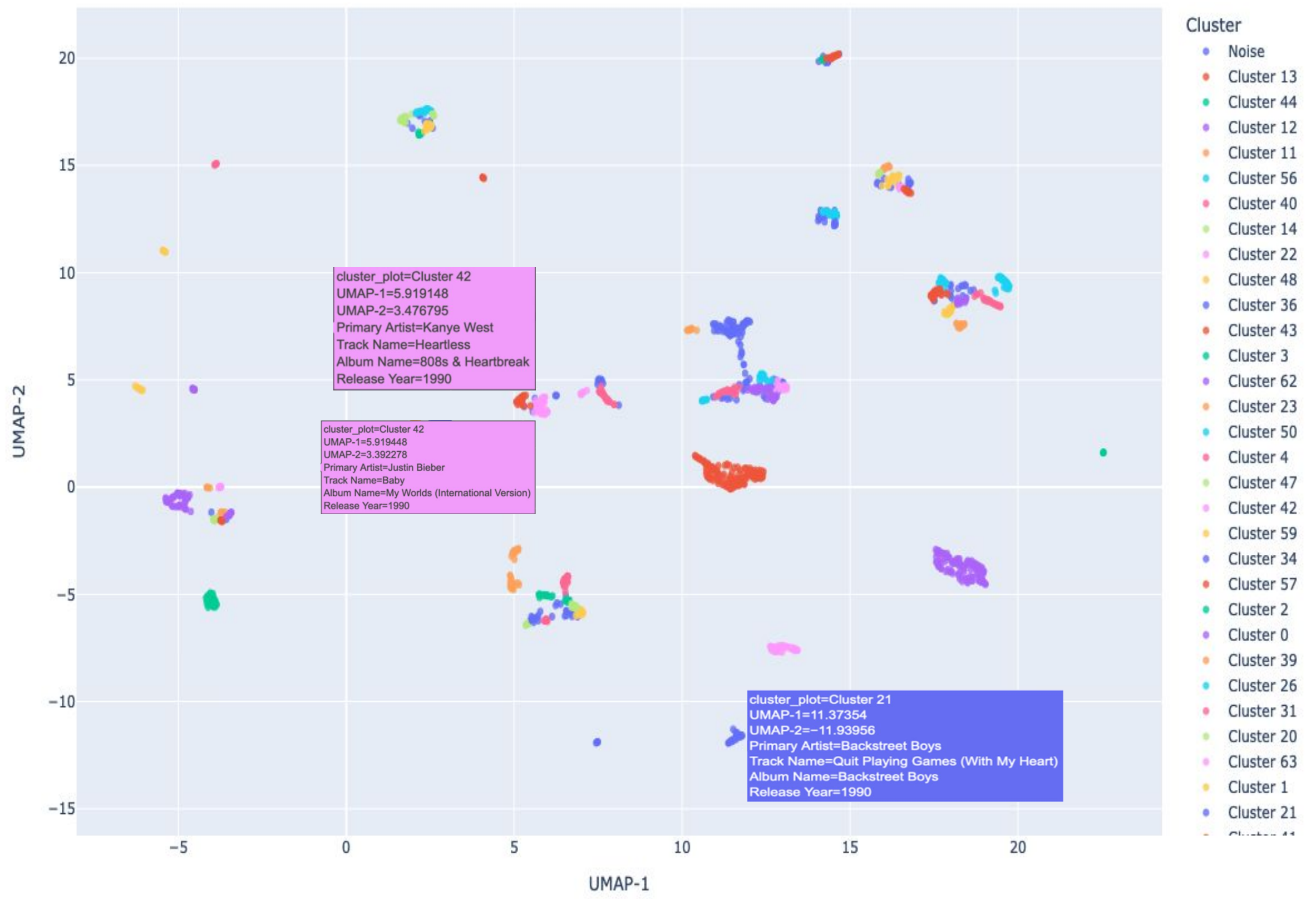
HDBSCAN Clustering on UMAP time feature Sihouette Score: 0.8005



HDBSCAN Clustering on UMAP sBERT Embeddings Sihouette Score: 0.4466



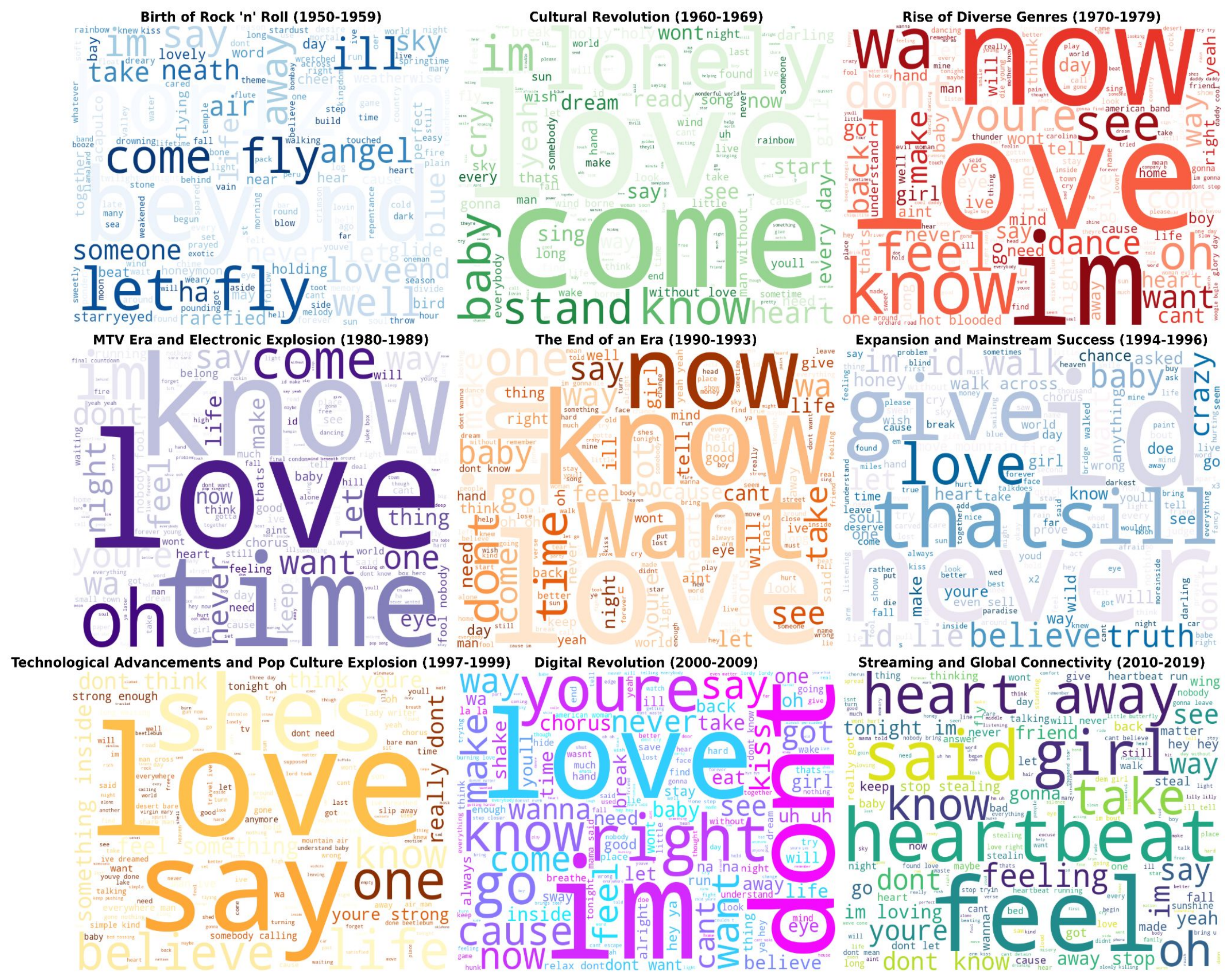
HDBSCAN Clustering on All features Sihouette Score: 0.7464



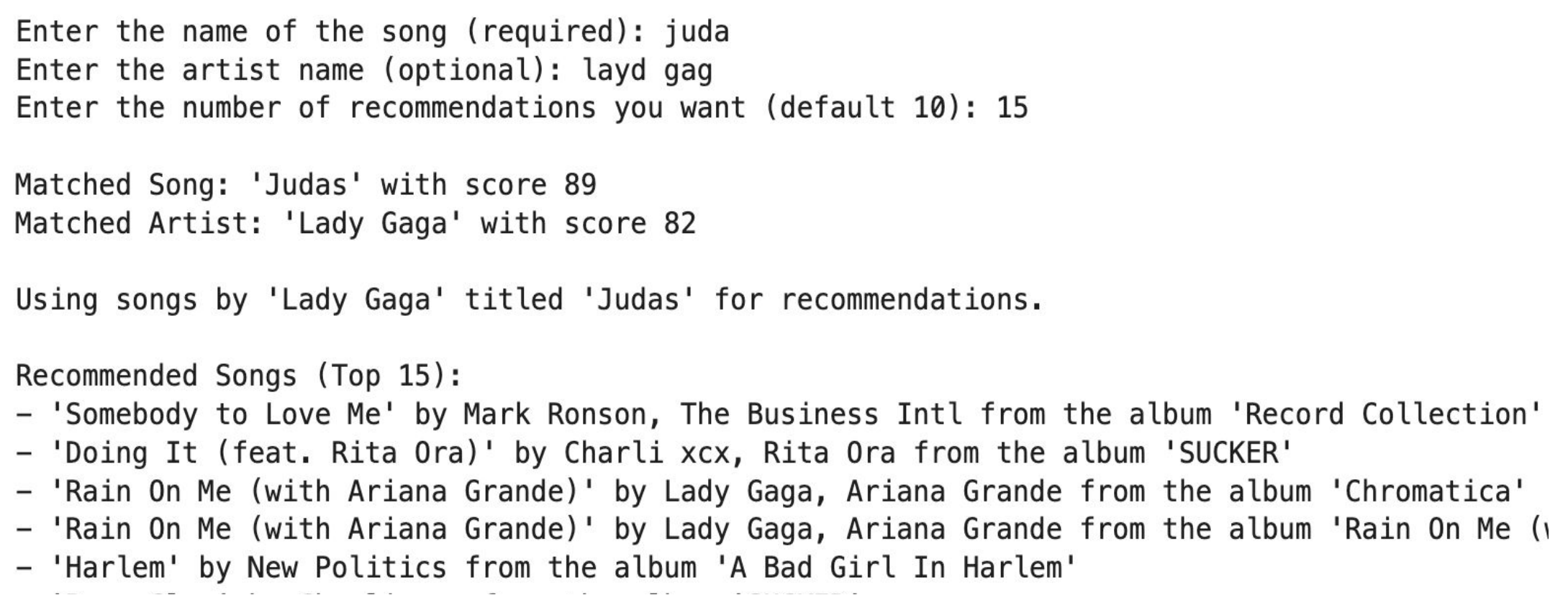
The first recommender system utilized only the selected features. The second recommender system utilizes sBERT, time binning, and the selected features. These features are shown in the final HDBScan clustering results. Very good cluster with Sihouette Score of 0.7464.

## 5. Key Learnings and Discussions

Word Cloud analysis revealed trends in lyrical themes of expressions across the decades. Shap analysis selected good features for a baseline recommender system. sBERT embedding maps the cultural relationship between music/musicians. These features improved our recommender system from one that recommended  songs of similar "mood" to one that capture both "mood" and "semantic meaning". The integration of audio features and lyrical semantics lead to a personalized recommender system.



Base Features Recommender System

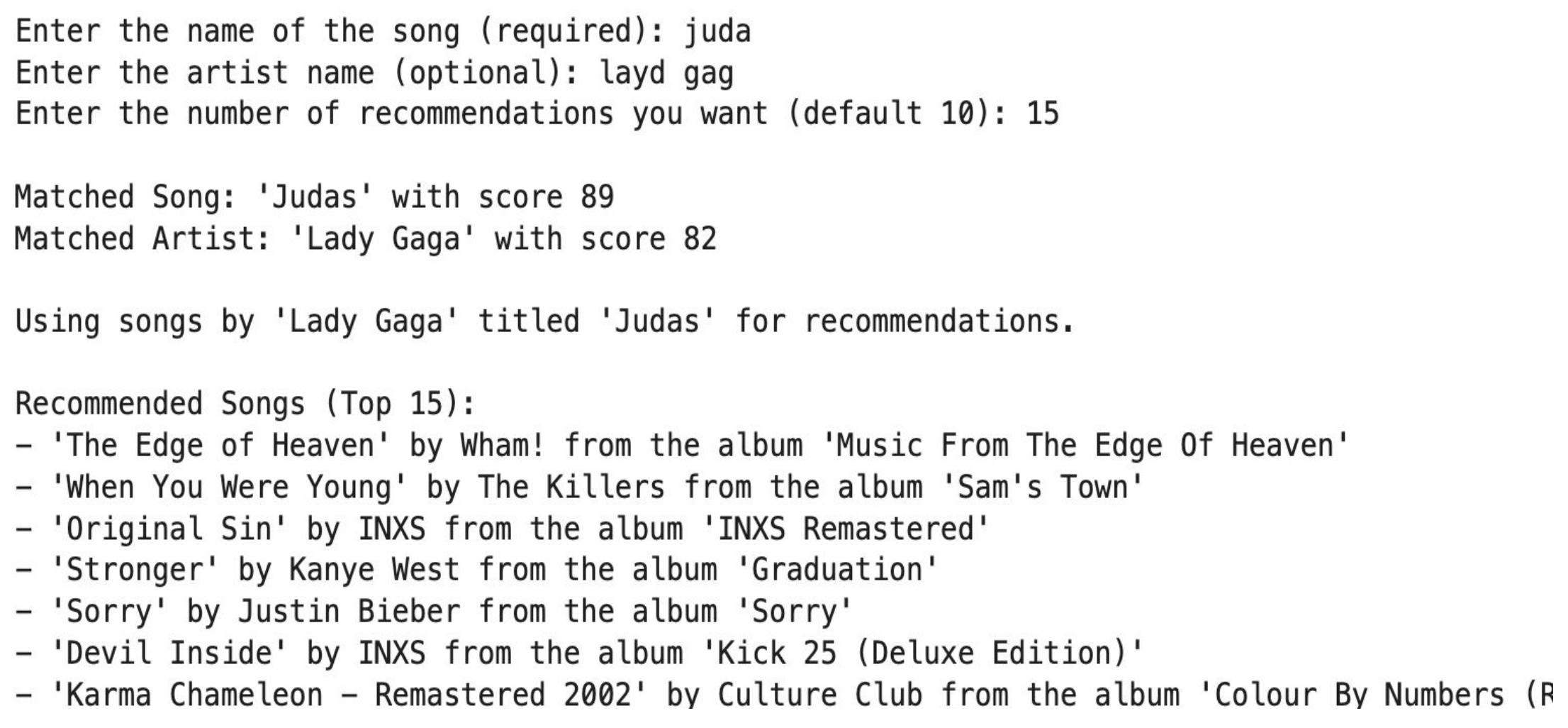

```
Enter the name of the song (required): juda
Enter the artist name (optional): layd gag
Enter the number of recommendations you want (default 10): 15

Matched Song: 'Judas' with score 89
Matched Artist: 'Lady Gaga' with score 82

Using songs by 'Lady Gaga' titled 'Judas' for recommendations.

Recommended Songs (Top 15):
– 'Somebody to Love Me' by Mark Ronson, The Business Intl from the album 'Record Collection'
– 'Doing It (feat. Rita Ora)' by Charli xcx, Rita Ora from the album 'SUCKER'
– 'Rain On Me (with Ariana Grande)' by Lady Gaga, Ariana Grande from the album 'Chromatica'
– 'Rain On Me (with Ariana Grande)' by Lady Gaga, Ariana Grande from the album 'Rain On Me (
– 'Harlem' by New Politics from the album 'A Bad Girl In Harlem'
```

Refined Features Recommender System


```
Enter the name of the song (required): juda
Enter the artist name (optional): layd gag
Enter the number of recommendations you want (default 10): 15

Matched Song: 'Judas' with score 89
Matched Artist: 'Lady Gaga' with score 82

Using songs by 'Lady Gaga' titled 'Judas' for recommendations.

Recommended Songs (Top 15):
– 'The Edge of Heaven' by Wham! from the album 'Music From The Edge Of Heaven'
– 'When You Were Young' by The Killers from the album 'Sam's Town'
– 'Original Sin' by INXS from the album 'INXS Remastered'
– 'Stronger' by Kanye West from the album 'Graduation'
– 'Sorry' by Justin Bieber from the album 'Sorry'
– 'Devil Inside' by INXS from the album 'Kick 25 (Deluxe Edition)'
– 'Karma Chameleon – Remastered 2002' by Culture Club from the album 'Colour By Numbers (R
```

## 6. Future Work

We will focus on making different types of  recommender systems that have bias towards genres, artists, or specific song feature. Our study limitations lies our merged dataset having around 2000 songs. We can revisit the merging process to incorporate more lyrics into our original 10,000-song dataset.