Genre-Focused Clustering of Spotify Song Data

Attempting to Cluster Uniform Data

By: Niki Vasan, David Gaviria, Gianluis Hernandez

Q: How do quantitative measures of musicality (e.g. tempo, danceability, instrumentalness) correlate to the genre profile of a song or group of songs?

Kaggle Spotify Dataset

Dataset of songs in Spotify

The full list of genres

Data Card | Code (40) | Discussion (6)

About Dataset

The full list of genres included in the CSV are Trap, Techno, Techhouse, Trance, Psytrance, Dark Trap, DnB (drums and bass), Hardstyle, Underground Rap, Trap Metal, Emo, Rap, RnB, Pop and Hiphop.

Usability ⓘ
10.00

License
CC0: Public Domain

Expected update frequency
Annually

Data Pre-Processing

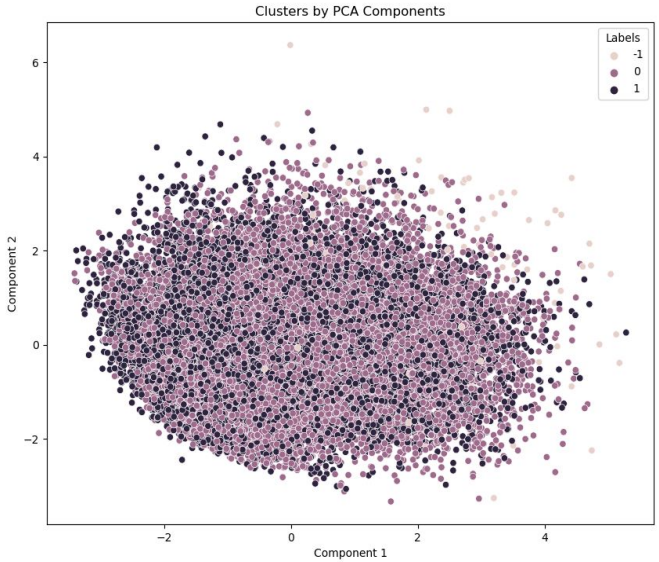
1. **Data Cleaning**
   - Remove irrelevant attributes such as duration, ID, API info etc…
2. **Feature Selection**
   - Use Pearson Correlation Matrix to mitigate collinearity among features
   - **Selected Features:** Danceability, Energy, Key, Loudness, Speechiness, Acousticness, Instrumentalness, Liveness, Valence, Tempo
3. **Feature Scaling**
   - Z-Score normalization of quantitative features

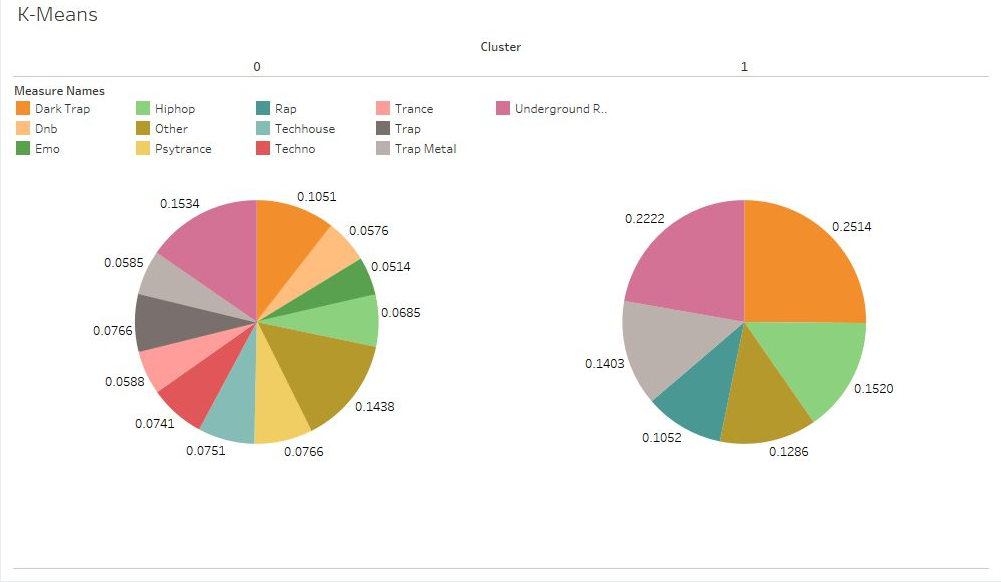
Data Mining Methods

K-Means and DB-Scan

K-Means

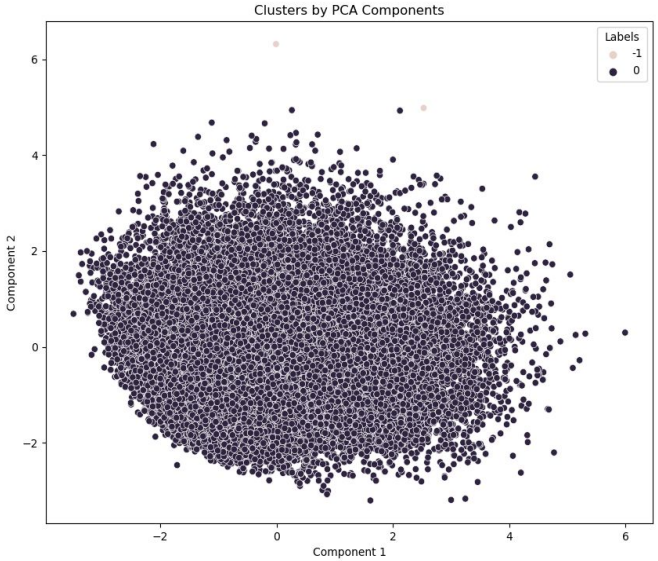
Optimal K	Silhouette Coef
2	0.1

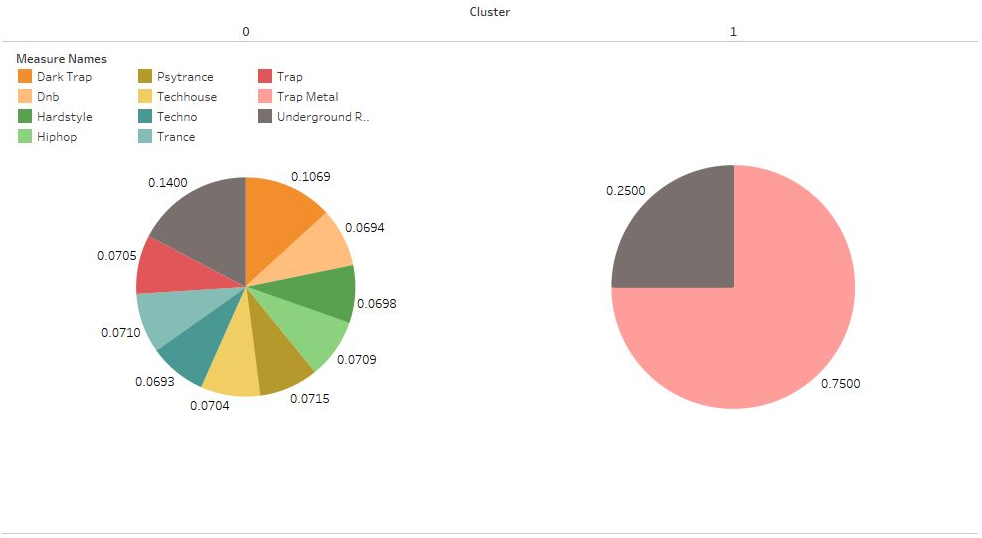


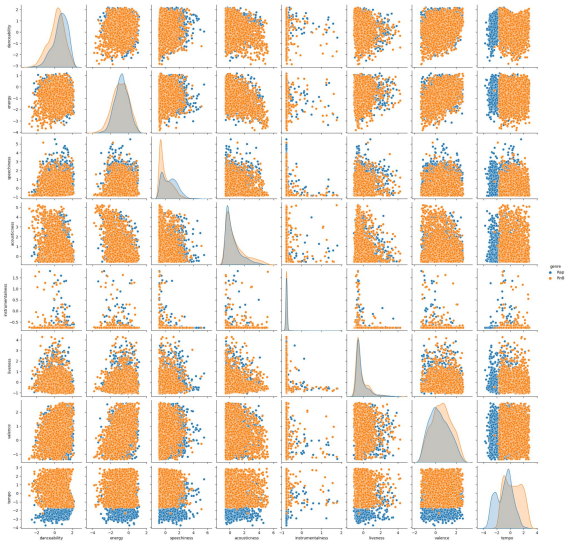


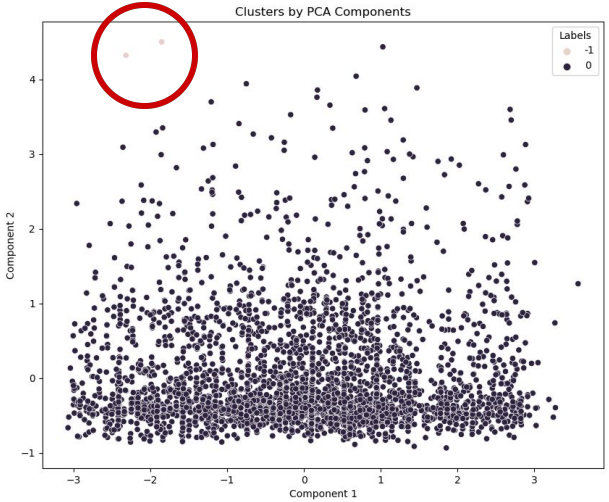
DB-Scan

Optimal Epsilon	Optimal MinPts	Silhouette Coefficient
4.0	3.0	0.5









DB Scan

Subsetting the Data

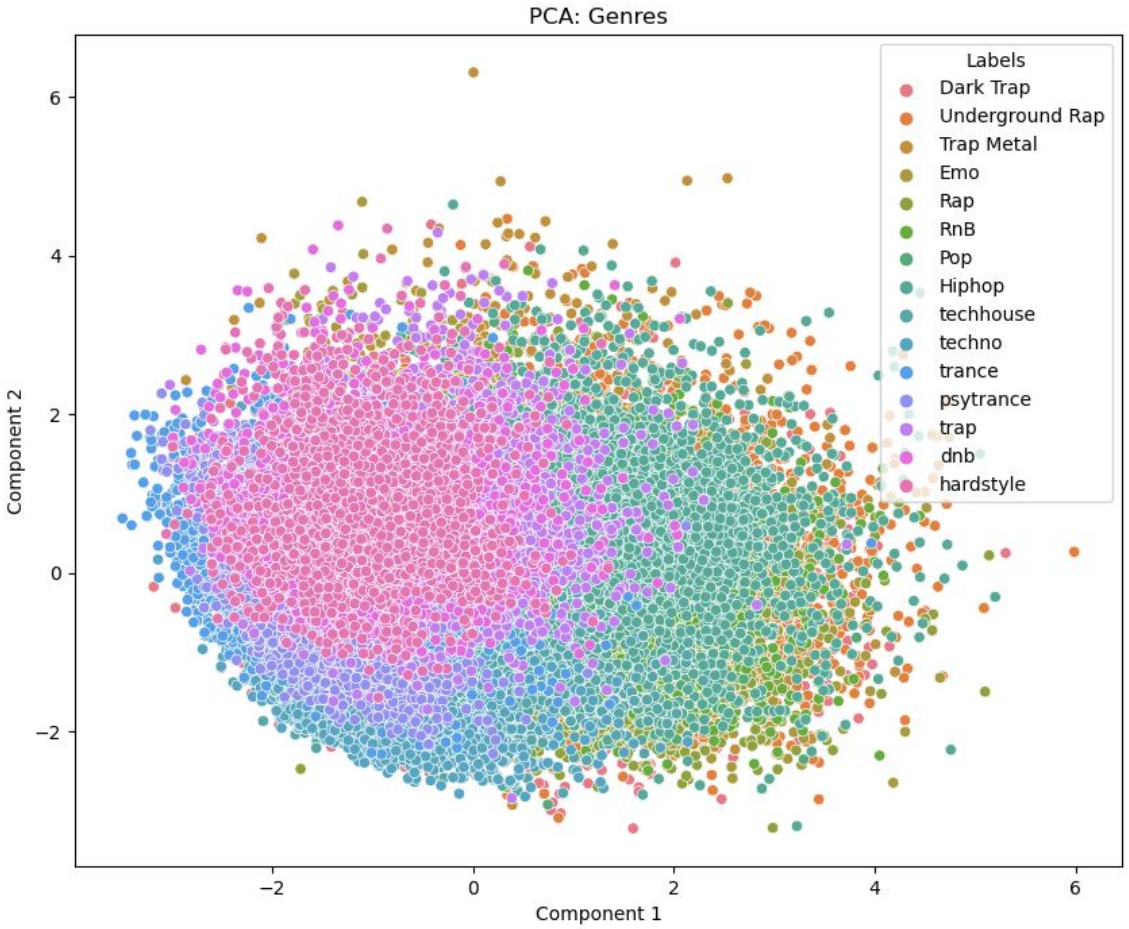
- Used pairwise scatterplot to find attributes that could spatially isolate the data better e.g. tempo, liveness
- Tried to identify genres that do not overlap e.g. Rap and R&B
- **Result:** *neither K-Means nor DB-Scan was able to create a meaningful cluster*

Conclusion

An evaluation on why our models don’t work.

1. Clustering Tendency
2. Fuzzy vs Hard Clustering for Genres
3. Hyperparameter Optimizations

Hopkin's Statistic = 0.14655



Thank you!