Spotify®

NICKY DESAI, ALON JACOBY, TREY ELDER | 12/8/2025

Spotify Song Genre Clustering

Big Data Analytics: CIS 5450, Professor Ives, Final Semester Project

https://newsroom.spotify.com/media-kit/logo-and-brand-assets/

Content overview

01

Data, Objective, Value Proposition

02

EDA Major Learnings

03

Modeling results

04

Implications and insights

05

Challenges/limitations

1) Dataset

- Large scale music dataset sourced from Kaggle (amitanshjoshi/spotify-1million-tracks)
- Metadata and audio descriptors for **over 1.15 million songs**
- Each row represents a single track in Spotify's database (**primary key)**

- **19 total features**
  - Basic metadata: **artist name, track title, release year, track ID**
  - Audio feature scores computed by Spotify's machine learning models (examples below):
    - **danceability**
    - **energy**
    - **acousticness**
    - **liveness**
    - **tempo**
    - **speechiness**
    - **instrumentalness**
  - **82 genres** → crawled from 3rd party source, wikepedia most likely

Quick Look

```
# Quick peek of the head of the dataset
spotify_df.head(10)
```

	artist_name	track_name	track_id	popularity	year	genre	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence
0	Jason Mraz	I Won't Give Up	53QF56cjZA9RTuuMZDrSA6	68	2012	acoustic	0.483	0.303	4	-10.058	1	0.0429	0.6940	0.000000	0.1150	0.139
1	Jason Mraz	93 Million Miles	1s8tP3jP4GZcyHDsjvw218	50	2012	acoustic	0.572	0.454	3	-10.286	1	0.0258	0.4770	0.000014	0.0974	0.515
2	Joshua Hyslop	Do Not Let Me Go	7BRCa8MPiyuvr2VU3O9W0F	57	2012	acoustic	0.409	0.234	3	-13.711	1	0.0323	0.3380	0.000050	0.0895	0.145
3	Boyce Avenue	Fast Car	63wsZUhUZLlh1OsyrZq7sz	58	2012	acoustic	0.392	0.251	10	-9.845	1	0.0363	0.8070	0.000000	0.0797	0.508
4	Andrew Belle	Sky's Still Blue	6nXIYClvJAfi6ujLiKqEq8	54	2012	acoustic	0.430	0.791	6	-5.419	0	0.0302	0.0726	0.019300	0.1100	0.217
5	Chris Smither	What They Say	24NvptbNKGs6sPy1Vh1O0v	48	2012	acoustic	0.566	0.570	2	-6.420	1	0.0329	0.6880	0.000002	0.0943	0.960
6	Matt Wertz	Walking in a Winter Wonderland	0BP7hSvLAG3URGrEvNNbGM	48	2012	acoustic	0.575	0.606	9	-8.197	1	0.0300	0.0119	0.000000	0.0675	0.364
7	Green River Ordinance	Dancing Shoes	3Y6BuzQCg9p4yH347Nn8OW	45	2012	acoustic	0.586	0.423	7	-7.459	1	0.0261	0.2520	0.000006	0.0976	0.318
8	Jason Mraz	Living in the Moment	3ce7k1L4EkZppZPz1EJWTS	44	2012	acoustic	0.650	0.628	7	-7.160	1	0.0232	0.0483	0.000000	0.1190	0.700
9	Boyce Avenue	Heaven	2EKxmYmUdAVXlaHCnnW13o	58	2012	acoustic	0.619	0.280	8	-10.238	0	0.0317	0.7300	0.000000	0.1030	0.292

Objective

- Use crawled genre labels as a **benchmark,** not absolute truth
- Apply **multiple unsupervised learning methods** to discover natural song groupings
  - k-Means
  - DBSCAN
  - Agglomerative Hierarchical Clustering
    - including dendrogram analysis + optimal reference cutoff selection
  - **Evaluate how well the learned clusters align with referenced labels**
    - Cluster to label agreement
    - Error rates and misclassification patterns
    - Statistical tests to assess significance
  - Analyze structure of audio features to understand how musical characteristics shape genre boundaries
  - Build interpretable groupings that reflect song level genre patterns

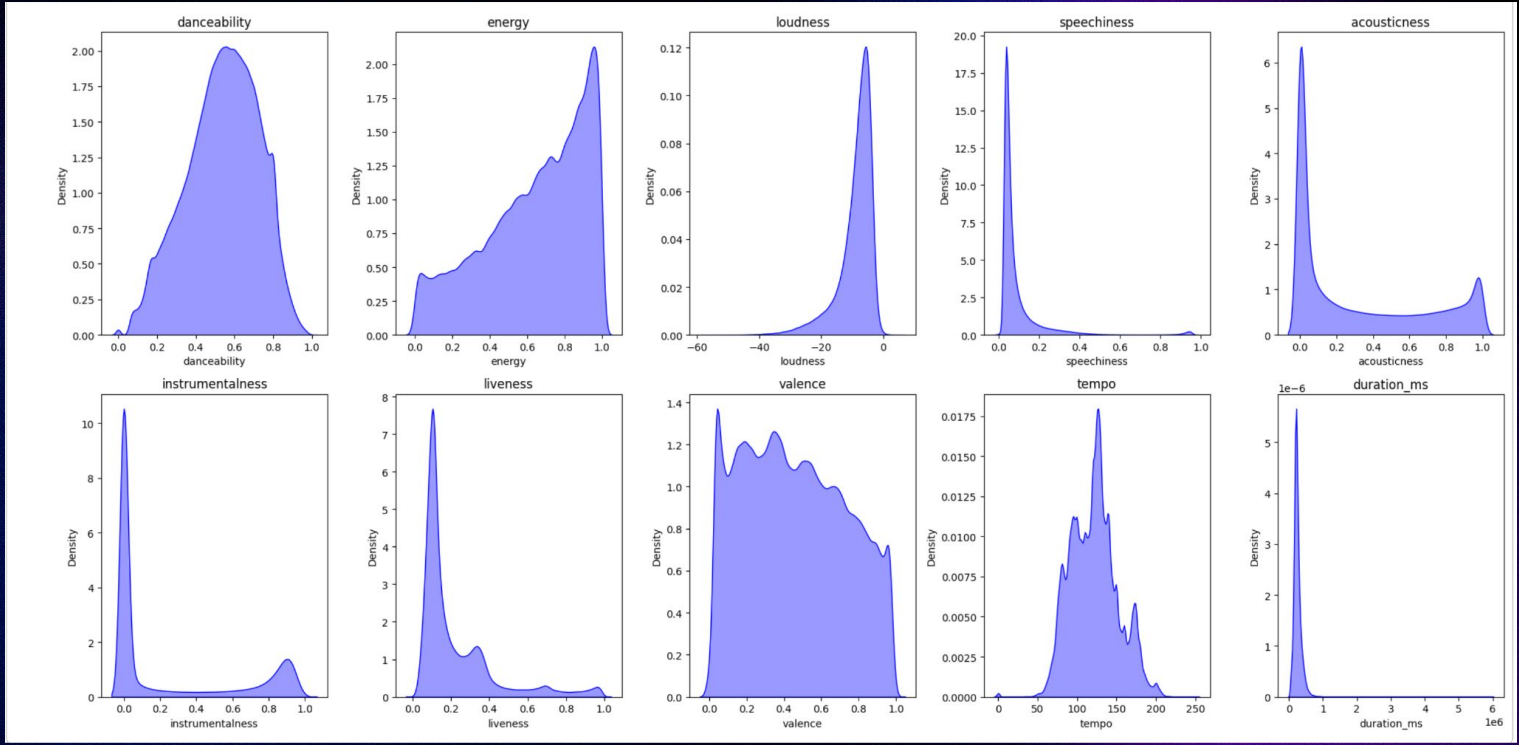
Value Proposition

- Provides a song-level genre grouping system-***something that Spotify does not offer***
- Reveals a natural structure in music based on acoustic and musical attributes, independent of noisy or inconsistent external labels
- Helps illustrate how audio features drive genre distinctions, enabling a **data-driven understanding of music similarity**
- Enables **potential applications** such as:
  - improving playlist generation
  - enhancing music discovery systems
  - identifying cross-genre or genre-blending songs
- Demonstrates how unsupervised learning can uncover meaningful insights in large-scale music datasets

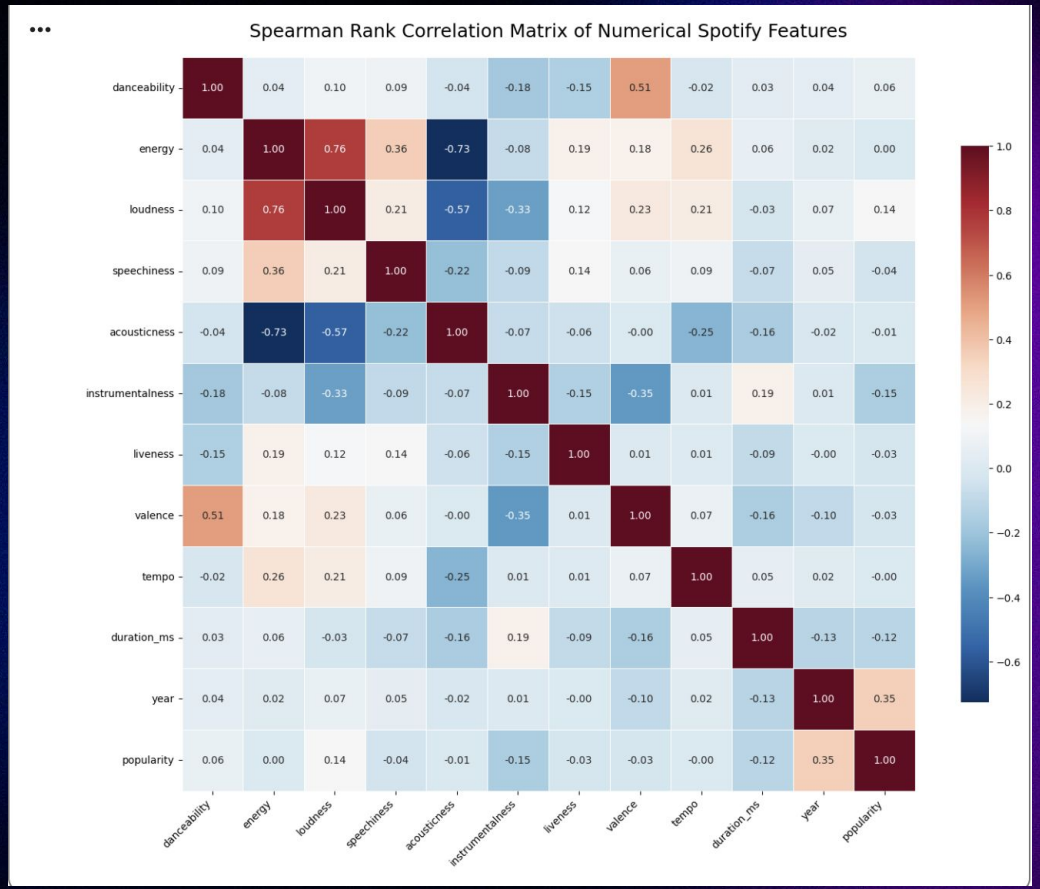
2) EDA Major Learnings

- There were 4 charts that helped deepen our understanding of the dataset and helped to inform our downstream model
  - 1) Distribution Plots of the Spotify Machine Learning Audio Features
  - 2) Spearman Rank Correlation Matrix
  - 3) Distribution of genre counts
  - 4) Cosine Similarity of Track Embeddings (Example Songs for Understanding)

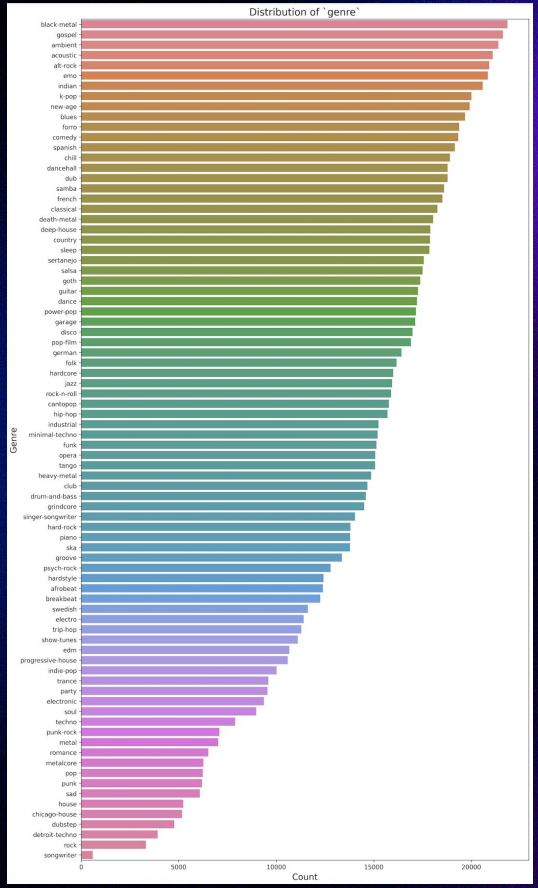
Distribution Plots



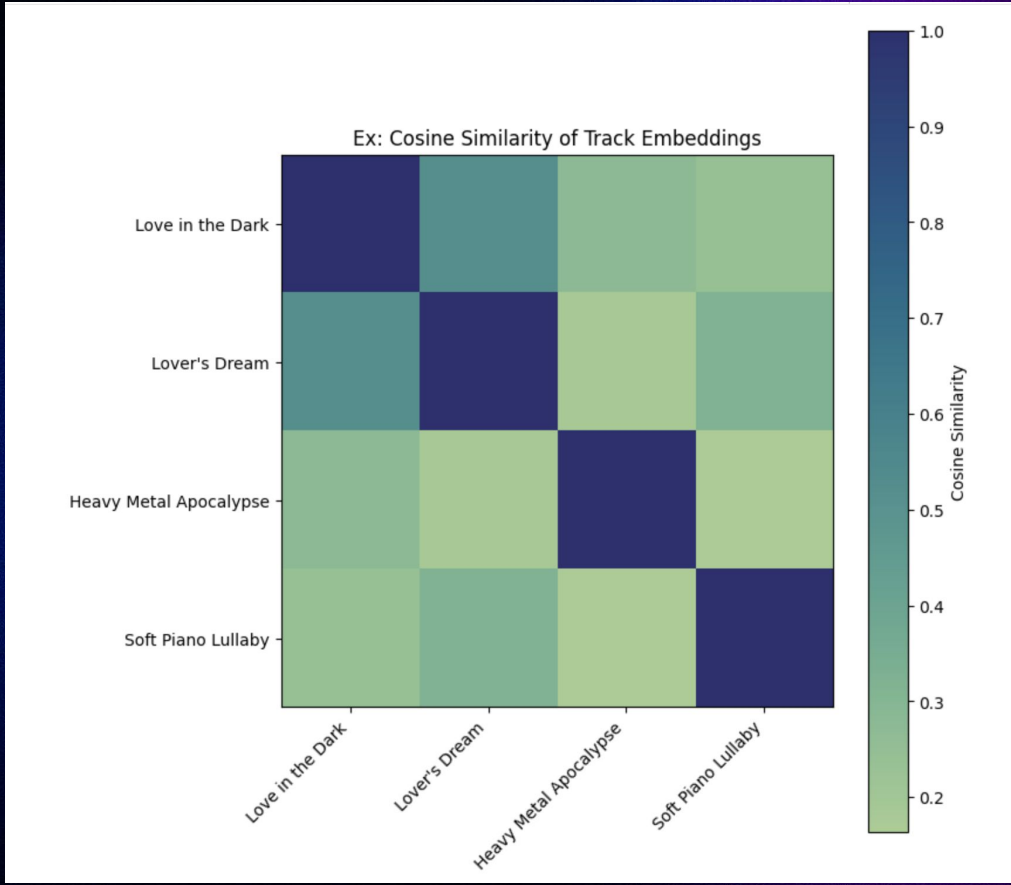
Spearman Correlation Matrix



Distribution of Genre



Cosine Similarity of Track Embeddings

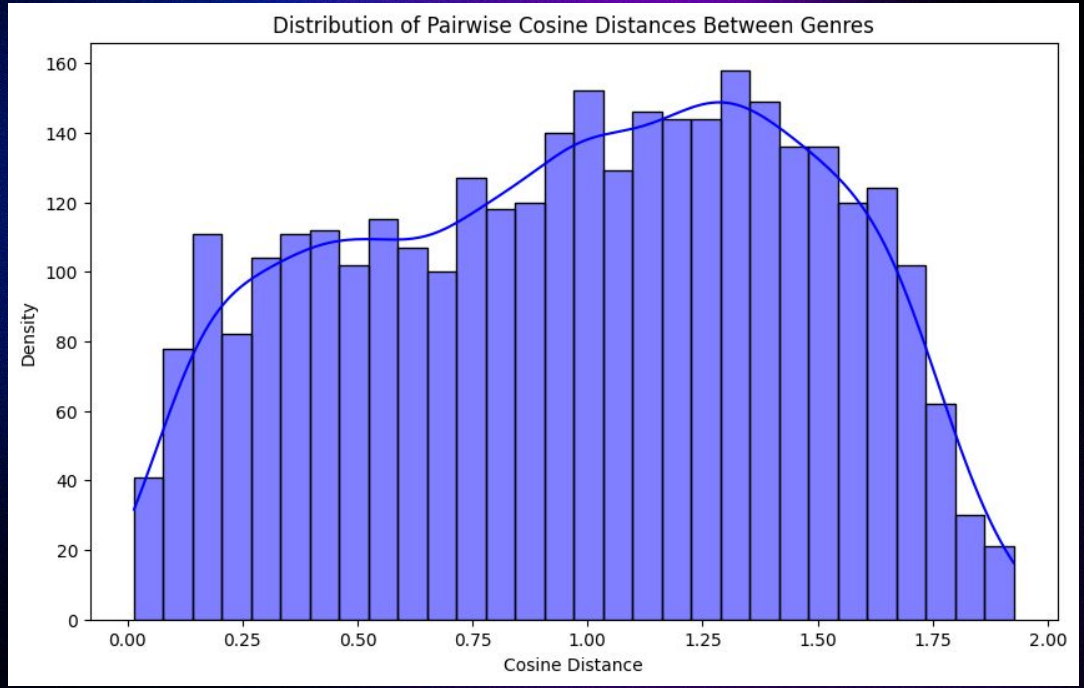


3) Modeling Results

Modelling objective: Learn the structure and distribution of genres within the Spotify data.

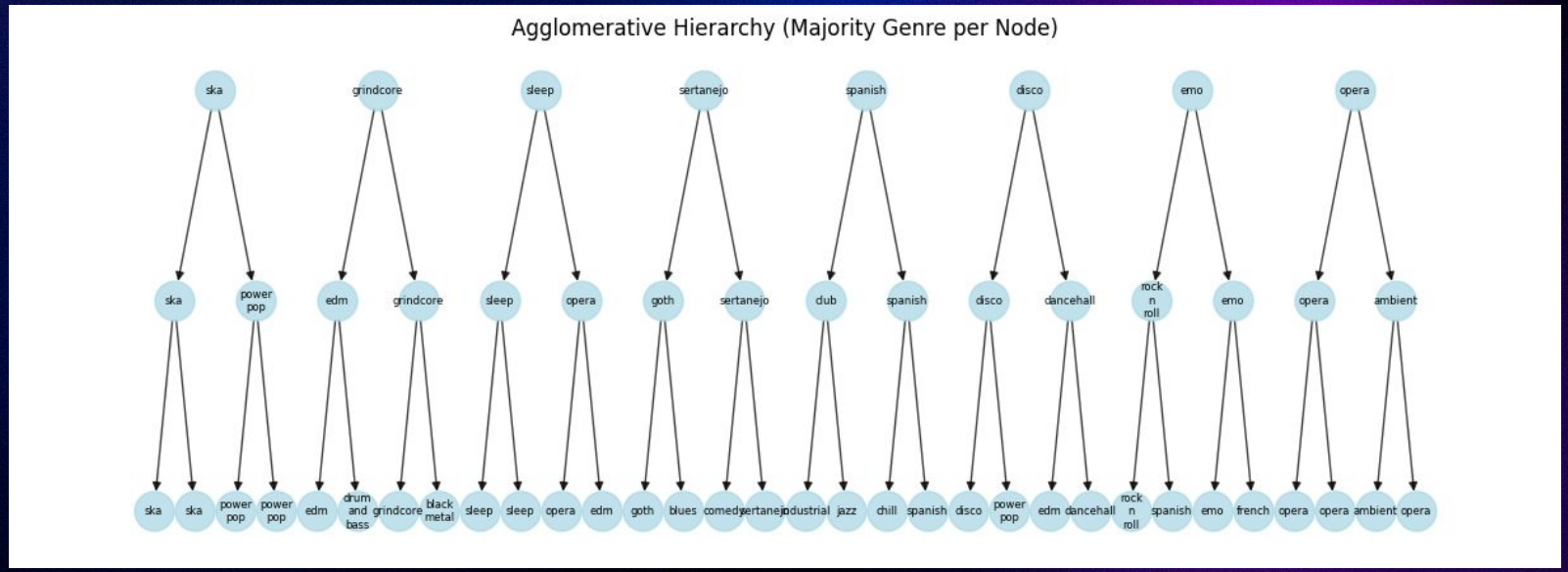
Preliminary observation:

Many genres are generally similar to one another, presenting a difficulty in discerning differentiating features.



Hierarchical Structure

Agglomerative Hierarchical Modeling confirms the preliminary results:
Strict hierarchy provides surprising merges at the intermediate levels due to difficulty in discerning genres.

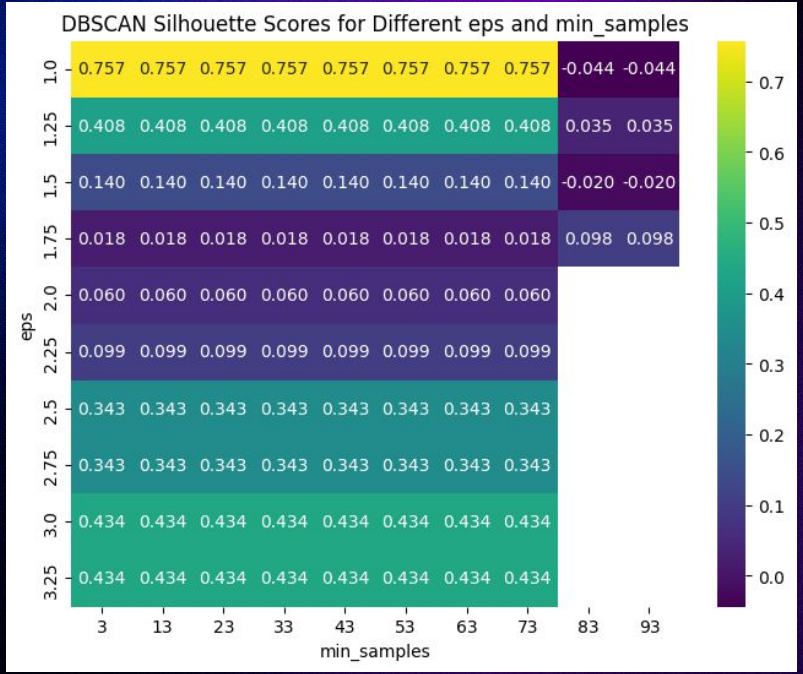


Attempt #1: DBSCAN

DBSCAN clusters points according to local densities and minimal cluster size parameters, allowing for non-convex cluster shapes.

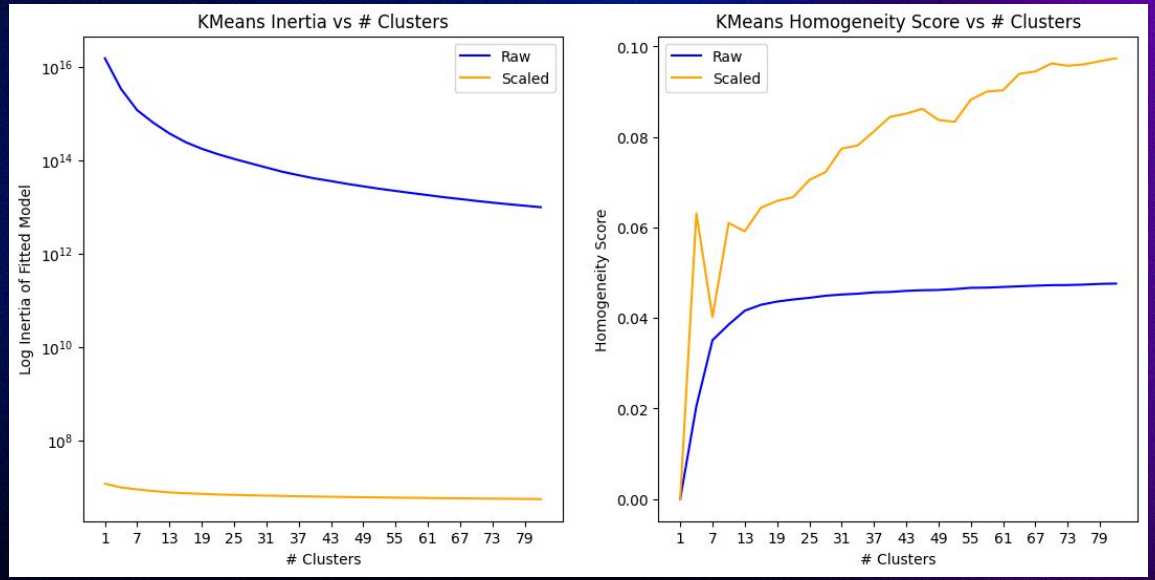
Silhouette scores supposedly show good modelling outcomes, with highly separable clusters.

Closer inspection reveals optimized solution tends towards many - but uninterpretable - clusters.



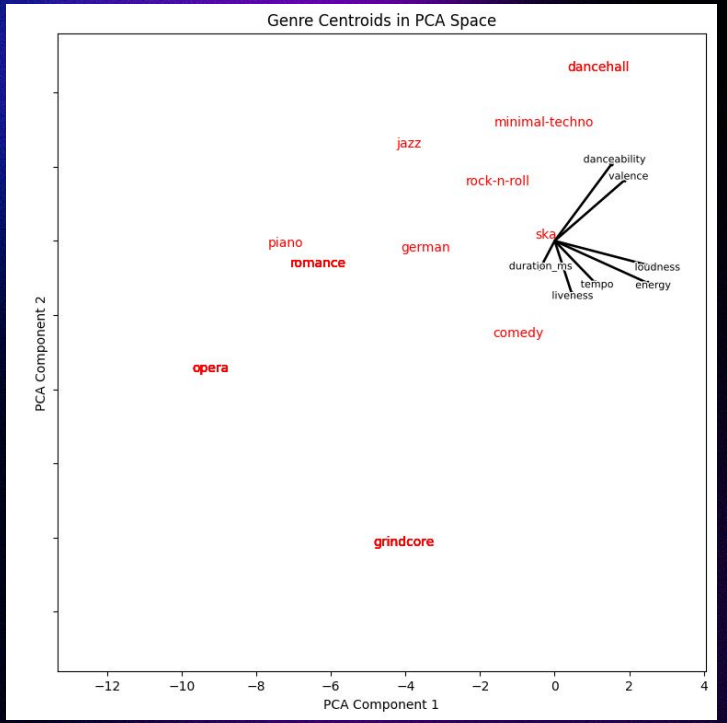
Attempt #2: K-Means

Turning to K-Means and using prior knowledge on number of expected genres, we choose an optimal K (number of clusters) based on both scaled and unscaled data, considering both model inertia and homogeneity scores.



Attempt #2: K-Means

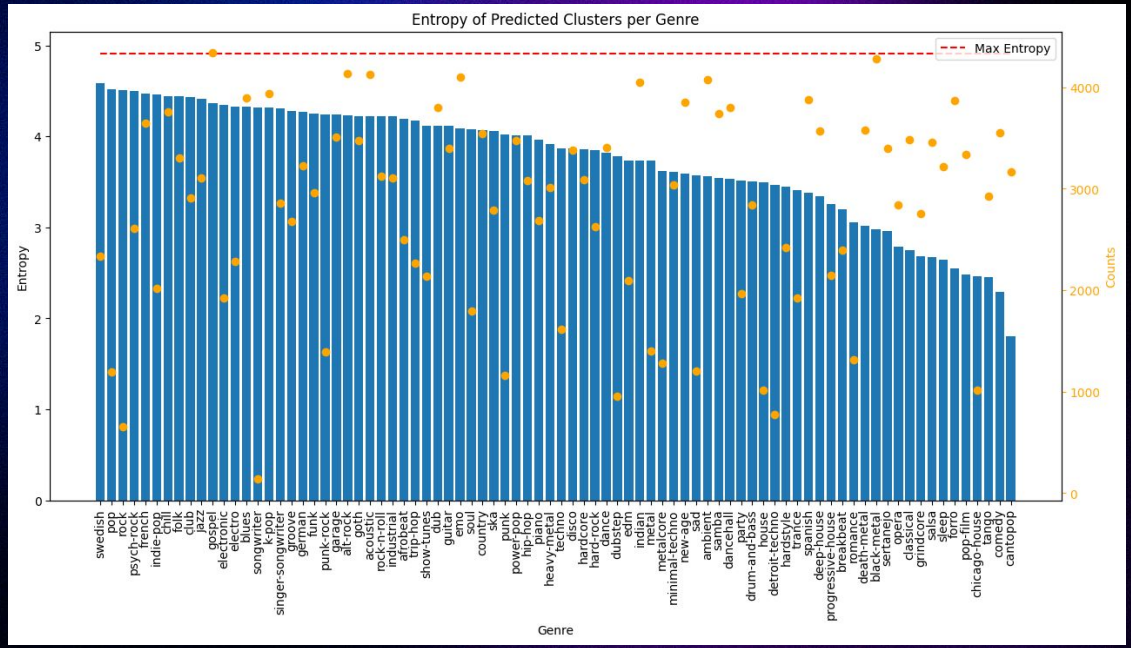
We correlate features to genres by producing the PCA bi-plot: The loadings of features plotted on top of the genre centroids scatter plot reveals that, e.g, Jazz is least with loudness and energy and minimal-techno most with danceability.



Attempt #2: K-Means

We confirm the quality of clustering *per genre* by considering the entropy of cluster assignments: The higher the entropy, the more uniform the cluster distribution, and vice versa.

Results show that “cantopop” and “comedy” are mostly associated with few clusters, while “swedish”, “pop” and “rock” appear in many clusters more uniformly, suggesting their feature values are not as unique.



Implications and Insights

- DBSCAN struggled to produce meaningful clusters
  - Despite hyperparameter tuning, clusters were of equal size and did not reflect the underlying genre structure
  - May be due to sensitivity to parameter choices and high dimensionality

- K-Means performed better
  - Scaling the data significantly improved genre separability
  - Some clusters still appeared redundant in PCA visualizations
  - Certain genres were more challenging to separate, suggesting a hierarchy of genre commonality

While K-Means provided valuable insights, ultimately both methods failed to perfectly capture the structure of genres in the data, suggesting that more sophisticated algorithms are likely needed

Challenges, Limitations and Future Work

Data challenges addressed through feature engineering and external model integration

- Binning quasi-binary audio features, one-hot encoding
- Transformers turning song names into semantic embeddings

The limitations of models due to assumptions about cluster shapes and distributions

- DBSCAN's reliance on density-based clustering made it sensitive to parameter choices and less effective in high-dimensional spaces.
- K-Means assumes spherical clusters of similar sizes
- Both struggled with the inherent complexity and overlap between genres

Future work could explore advanced clustering techniques and complex data

- Gaussian mixture models or deep learning-based approaches could better capture complex relationships in the data
- More domain-specific features or leveraging external metadata
- Investigating temporal dynamics of popularity

Thank you!

Contact

Date