Hierarchical Clustering

Maya Archer, Rei Kanazawa, Saumya Bothra

Can we identify groups of years with similar audio characteristics in the Spotify data set using hierarchical clustering?

How has musical style evolved overtime?

Dataset

📊 Observations: 100
📅 Years: 1921 - 2020
🎯 One data point per year (aggregated)
🎵 Audio Features generated by Spotify API
📏 Different scales

Our Variables

Acousticness (0-1)
Confidence that track is acoustic

Liveness (0-1)
Presence of audience

Instrumentalness (0-1)
Predicts no vocals

Tempo (BPM)
Speed in beats per minute

Danceability (0-1)
How suitable for dancing

Speechiness (0-1)
Presence of spoken words

Duration (milliseconds)
Length in milliseconds

Energy (0-1)
Intensity and activity

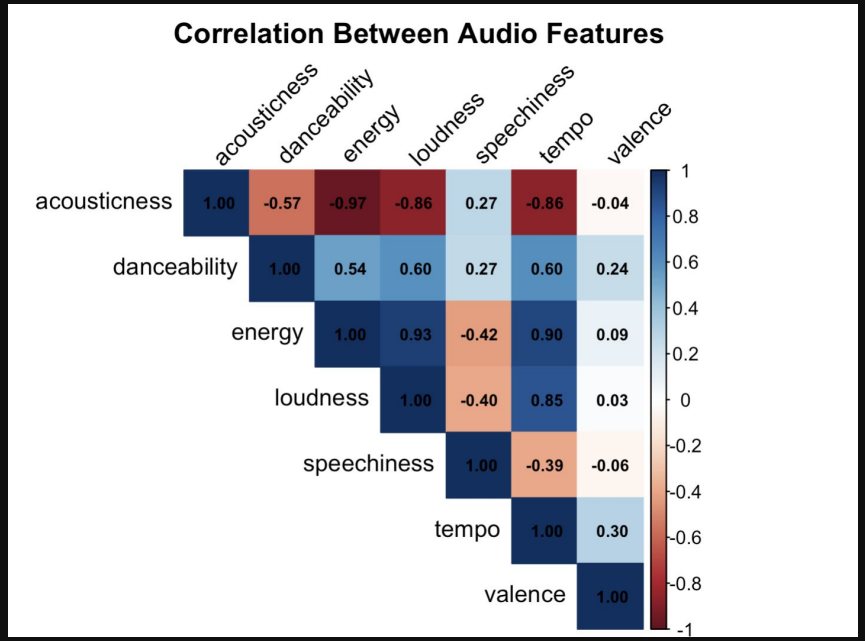
Valence (0-1)
Musical positivity/happiness

Loudness (Decibels)
Overall volume

Dataset Overview

year	Acoustic ness	Dance ability	Duration (ms)	Energy	Instrument alness	Liveness	Loudn ess	Speechine ss	Tempo	Valence	Popularit y
1921	0.8869	0.4186	260537.2	0.2318	0.3449	0.2057	-17.0 5	0.0737	101.53	0.3793	0.65
1922	0.9386	0.4820	165469.7	0.2378	0.4342	0.2407	-19.2 8	0.1167	100.88	0.5355	0.14
1923	0.9572	0.5773	177942.4	0.2624	0.3717	0.2275	-14.1 3	0.0939	114.01	0.6255	5.39
1924	0.9402	0.5499	191046.7	0.3443	0.5817	0.2352	-14.2 3	0.0921	120.69	0.6637	0.66
1925	0.9626	0.5739	184986.9	0.2786	0.4183	0.2377	-14.1 5	0.1119	115.52	0.6219	2.60

Relationships Between Features



Correlation indicates why certain years wil cluster together:

Energy ↔ Loudness (strong positive)

Acousticness ↔ Energy (strong negative)

Data preparation

Handling missing values
- ➔ None

Checking for outliers (IQR)
- ➔ Log-transformed speechiness
- ➔ Kept others due to low influence

Selecting relevant features
- ➔ Left out “year” because we use that for interpretability

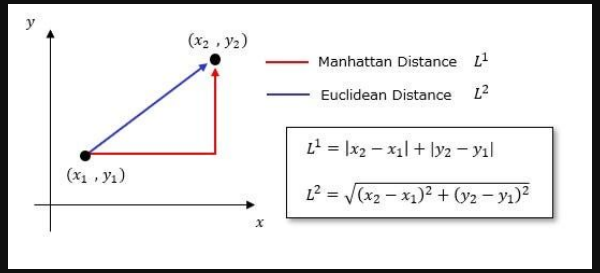
Normalizing or standardizing data
- ➔ Scaled numeric features to z-scores (mean = 0, sd = 1)

Hierarchical	Partitional
No need to specify **k** ahead	Must choose fixed number **k**
Builds step-by-step, so running again gives the same dendrogram	Results can change with different random initial seeds
Reveals substructures	Produces a single flat partition
Suitable for small datasets - compares all points with each other	Scales well to large datasets - divides data by minimizing distances

Why hierarchical clustering?

- We don’t know in advance how many natural eras there are
- Dendrogram lets us see broader eras and smaller transitional periods all in one picture
- Small dataset (100 obs)

Distance measures

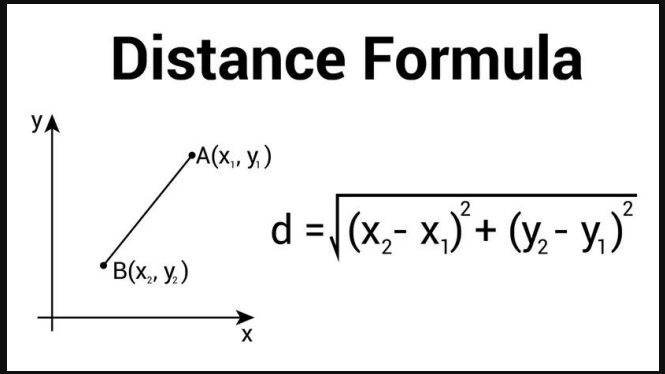


Distance	Idea	When to use
Euclidean	Straight-line distance between two points, square root of square differences	When your data is continuous and you want to measure true geometric similarity
Manhattan	Sum of absolute differences of two points	When your data is categorical, sparse, or contains outliers, and you want all differences to contribute equally

Distance measures

Euclidean distance

- Measures the overall similarity between songs across all continuous features at once,
- Continuous and scaled data
- No big outliers



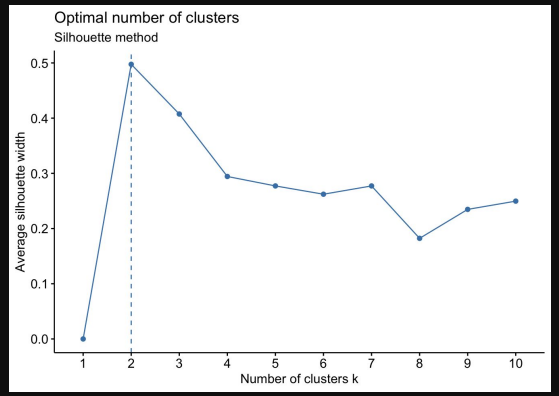
Linkage methods

Method	How it merges clusters	Pros & cons
Single	Based on the closest pair of songs	Can form long “chains” where eras could connect just because of one similar song
Complete	Using the furthest pair of songs.	Creates compact clusters but can split naturally continuous genres or eras (classic rock → indie rock)
Average	Based on the average distance between all pairs.	More balanced, but doesn’t explicitly minimize within-cluster variance
Ward’s	Merges clusters that minimize the total within-cluster variance.	Creates compact, homogeneous clusters of songs with similar feature profiles

Cluster performance measures

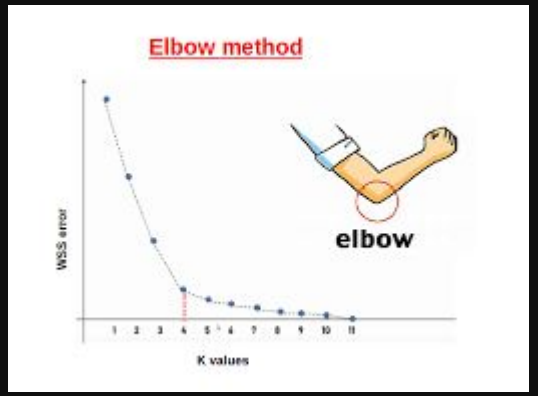
Silhouette curve

Measures how well each point fits within its cluster compared to other clusters

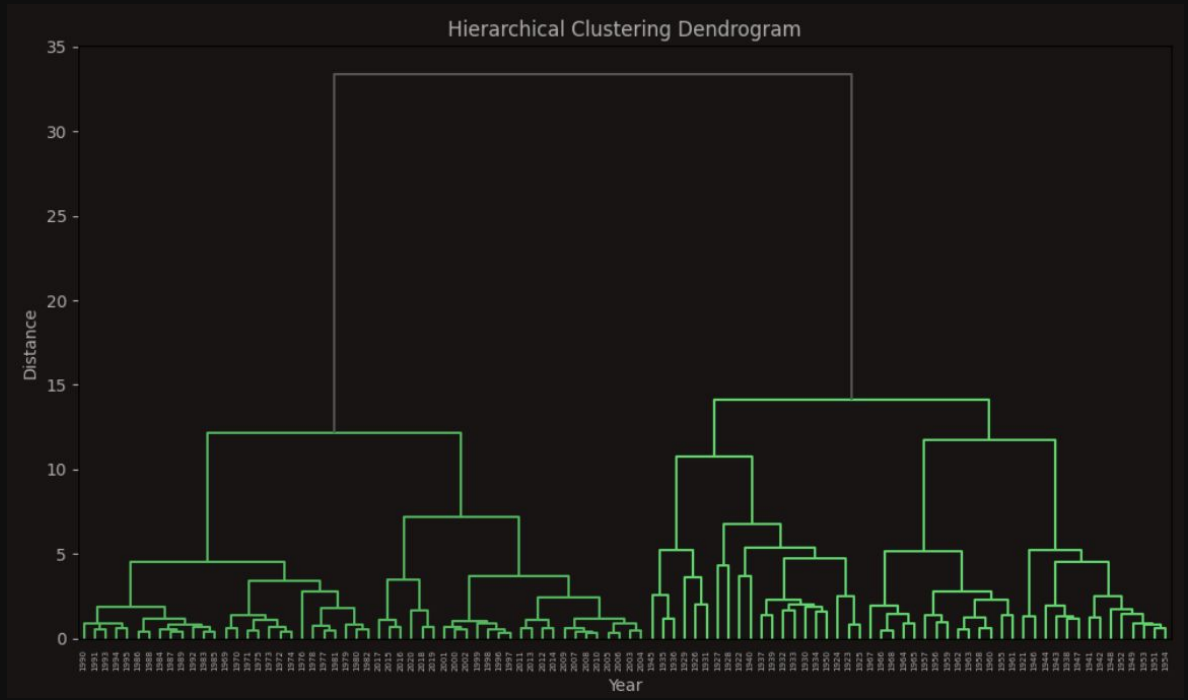


Total WSS (Elbow)

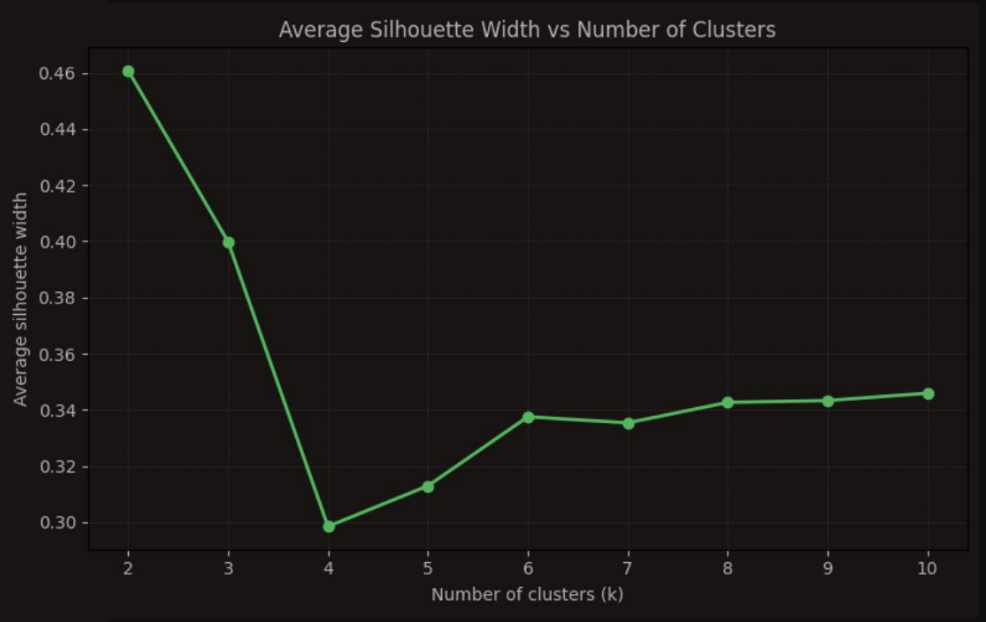
Measures measures how within-cluster variance decreases as you increase k



Dendrogram


Hierarchical Clustering Dendrogram
Distance
35
30
25
20
15
10
5
0
Year

Silhouette Curve

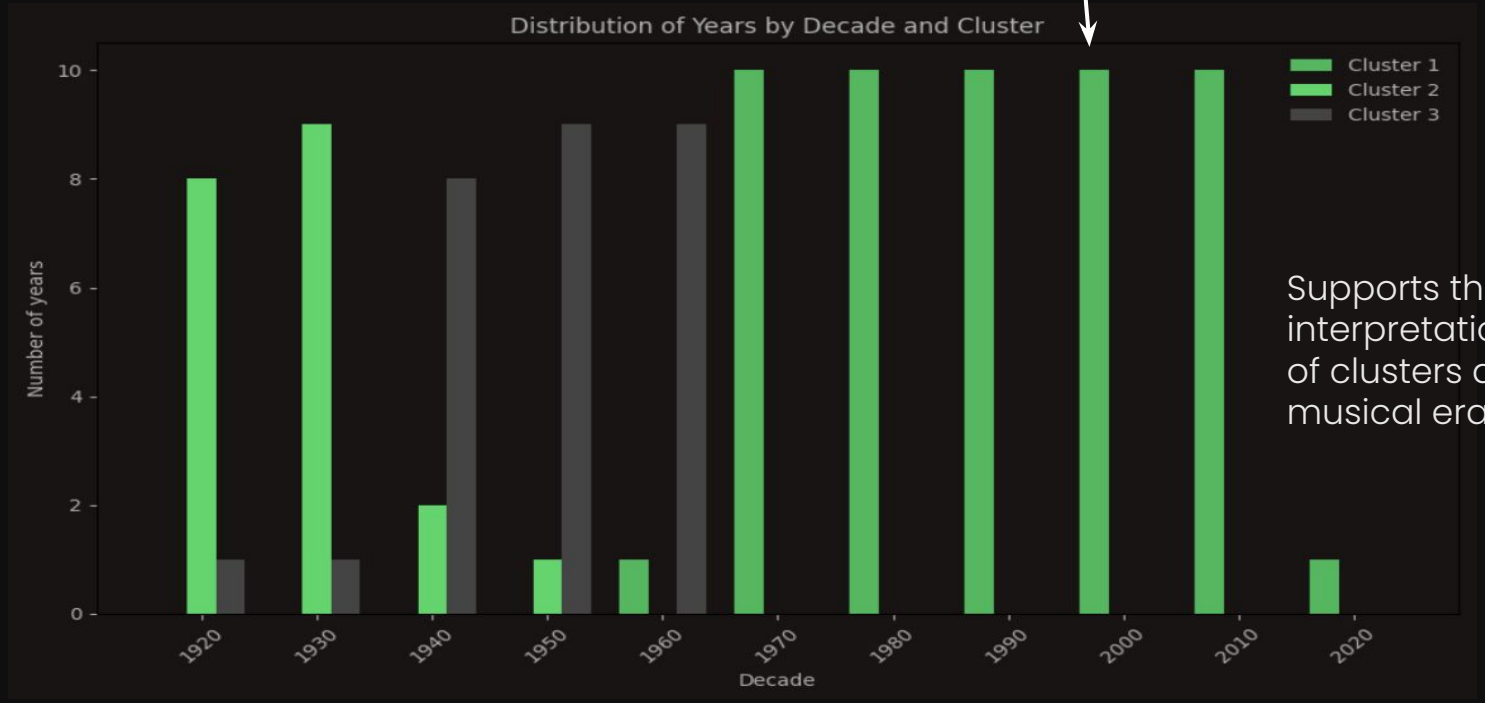


highest average silhouette width at k=2 and decreases to ≈0.40 for k=3, but k=3 was chosen for coherence

Early **Mid** **Modern**

Decade Distribution

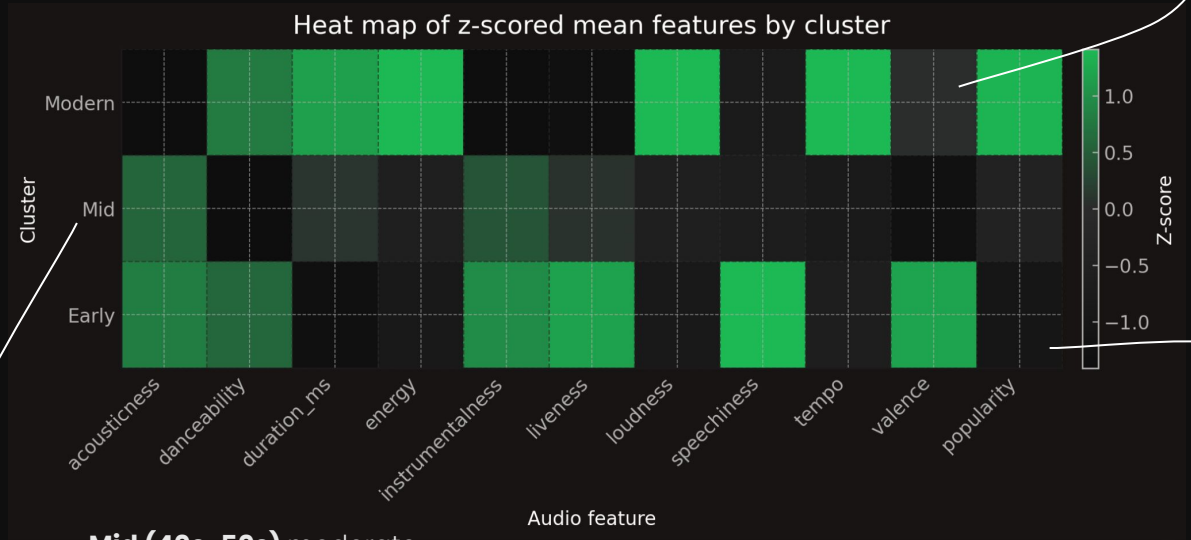
You can't tell but this green is darker than the other green



Supports the interpretation of clusters as musical eras.

Cluster Heat Map

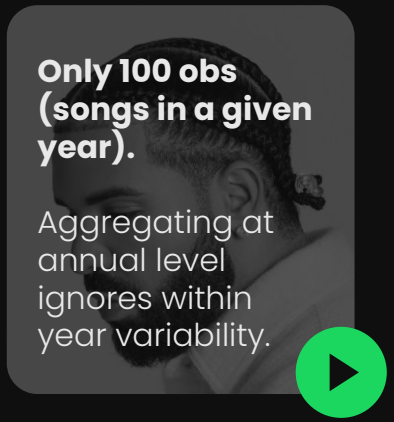
Modern (60s-present) Lowest acousticness, highest energy, danceability, and popularity. Higher loudness and shorter durations



Mid (40s-50s) moderate acousticness, lower danceability and popularity than modern but more than early

Early (20s-30s) high acousticness and valence, lowest energy tempo and popularity.

Limitations



Only 100 obs (songs in a given year).

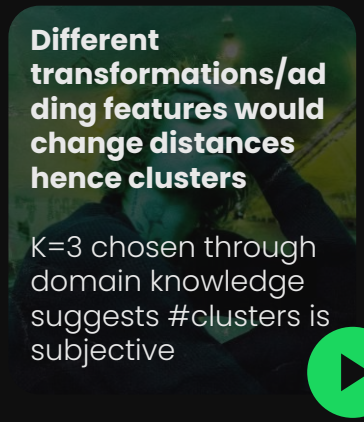
Aggregating at annual level ignores within year variability.

Dataset

Ward's linkage build tree recursively and cannot undo earlier merges.

Early decisions can lead to sub-optimal cluster structures

Greedy HC



Different transformations/adding features would change distances hence clusters

K=3 chosen through domain knowledge suggests #clusters is subjective

Sensitivity

What do you want to play ?

Conclusions

Summary
Artist

Patterns
Artist

Early Era
Artist

Mid Era
Artist

Modern Era
Artist

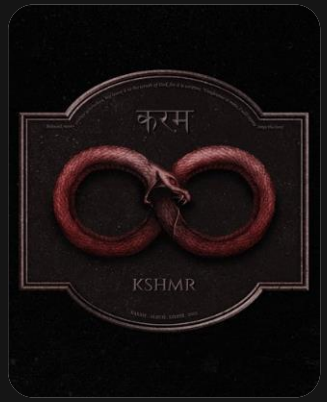
Usefulness and Future Work

Music
Method provides intuitive dendrogram and doesn't require pre-specifying clusters.

Podcasts
Collect song level data or more features like genre and lyrics for nuanced patterns

Live Events
Validate results with other methods and domain knowledge

KARAM



Khoya Sab
KSHMR

Next in Queue

KSHMR
Enemies

Thanks for Listening