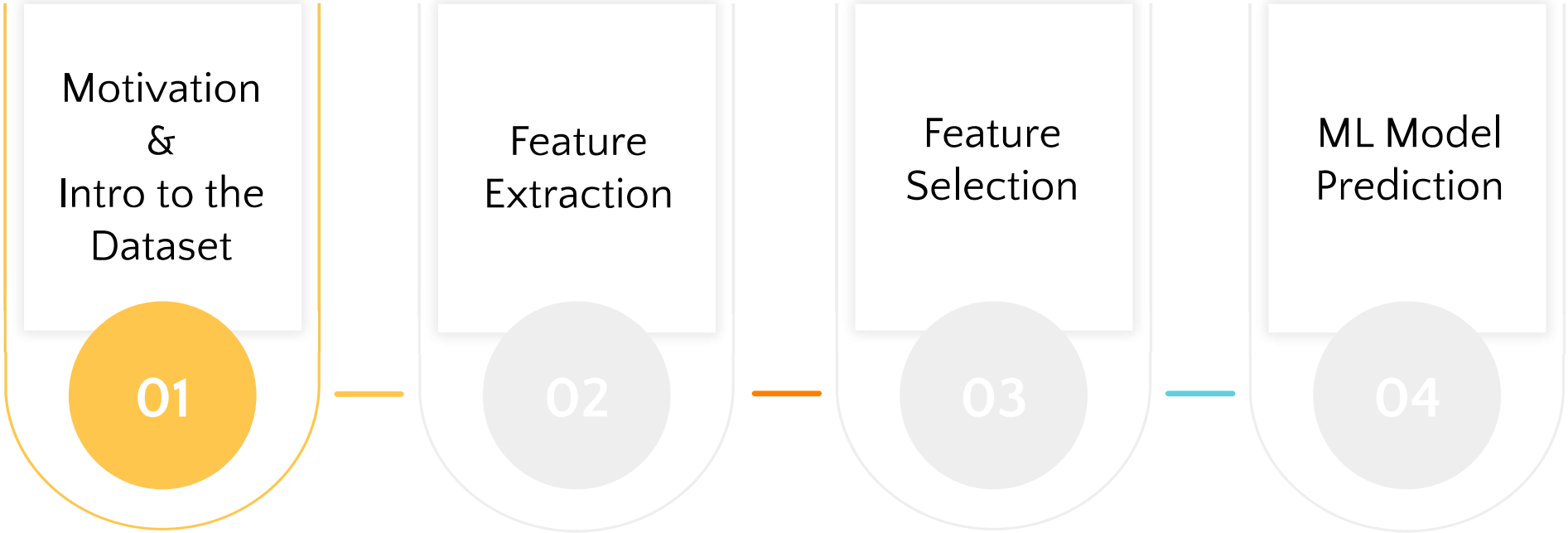
Group 11

Music Popularity Prediction

Karl Hernandez, Peng Wei Lee, Karrmanya Pande, Henri Schulz, Yue Yu

Outline


Motivation
&
Intro to the
Dataset
01
Feature
Extraction
02
Feature
Selection
03
ML Model
Prediction
04

Motivation

Is the popularity of songs predictable from **just lyrical and audio features**?

Intro to SpotGenTrack Dataset

Size

101,939 songs

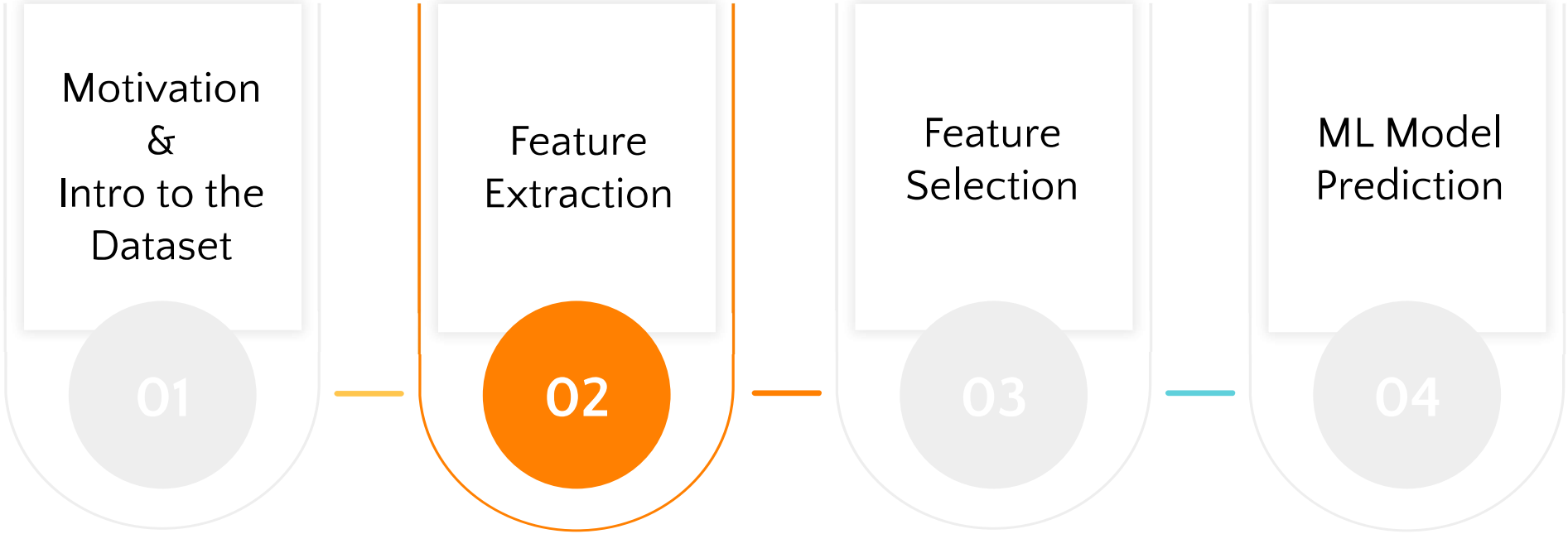
Principal features

Popularity
Genre
Lyrics
Low-level audio features

Source

Martín-Gutiérrez, David; Hernández-Peñaloza, Gustavo; Belmonte-Hernández, Alberto; Álvarez, Federico (2019), "SpotGenTrack Popularity Dataset", Mendeley Data, V1, doi: 10.17632/4m2x4zngny.1

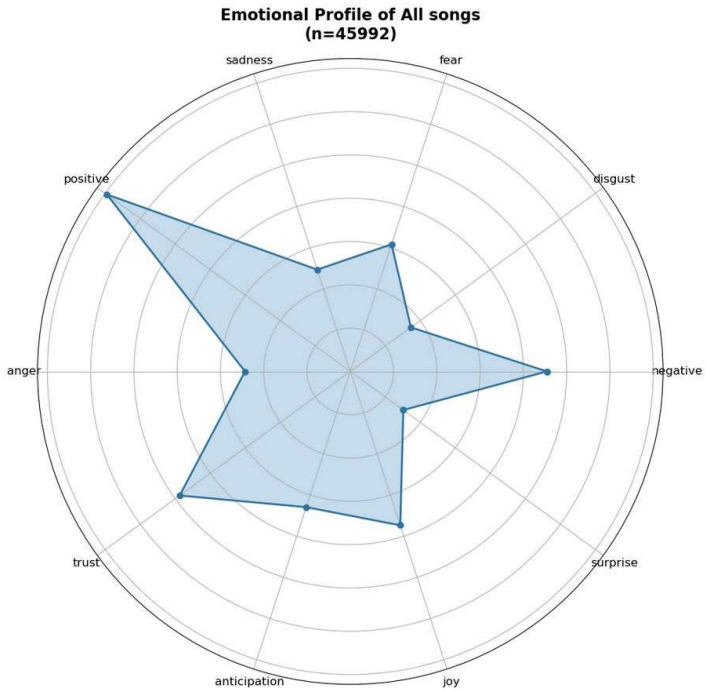
Outline

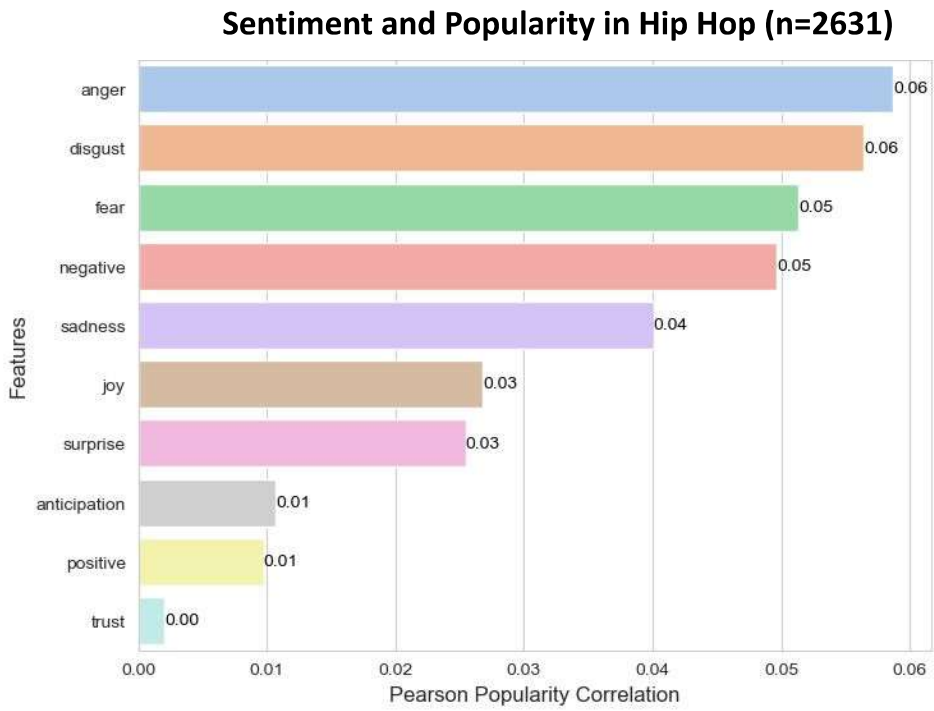

Motivation
&
Intro to the
Dataset
01
Feature
Extraction
02
Feature
Selection
03
ML Model
Prediction
04

Sentiment Extraction Analysis

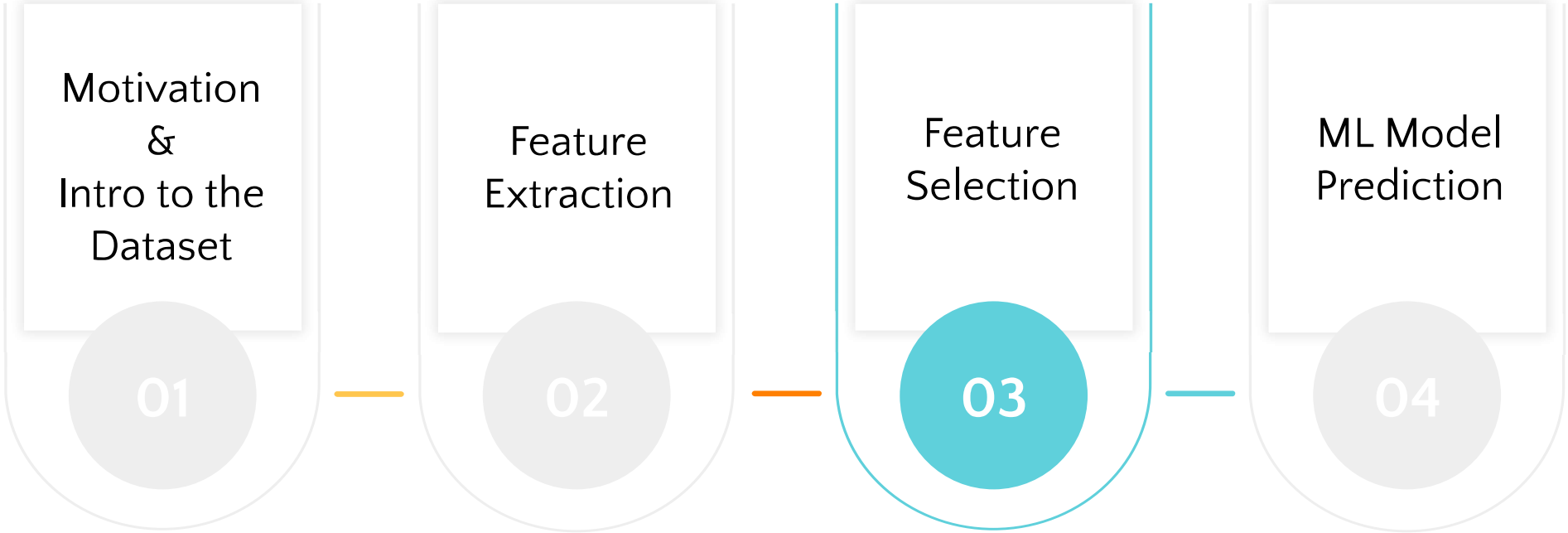
NLP for lyrics to analyze the emotional mood in correlation with the popularity of songs.

Sentiment features: negative, disgust, fear, sadness, positive, anger, trust, anticipation, joy, surprise (additional features normalized over word count)





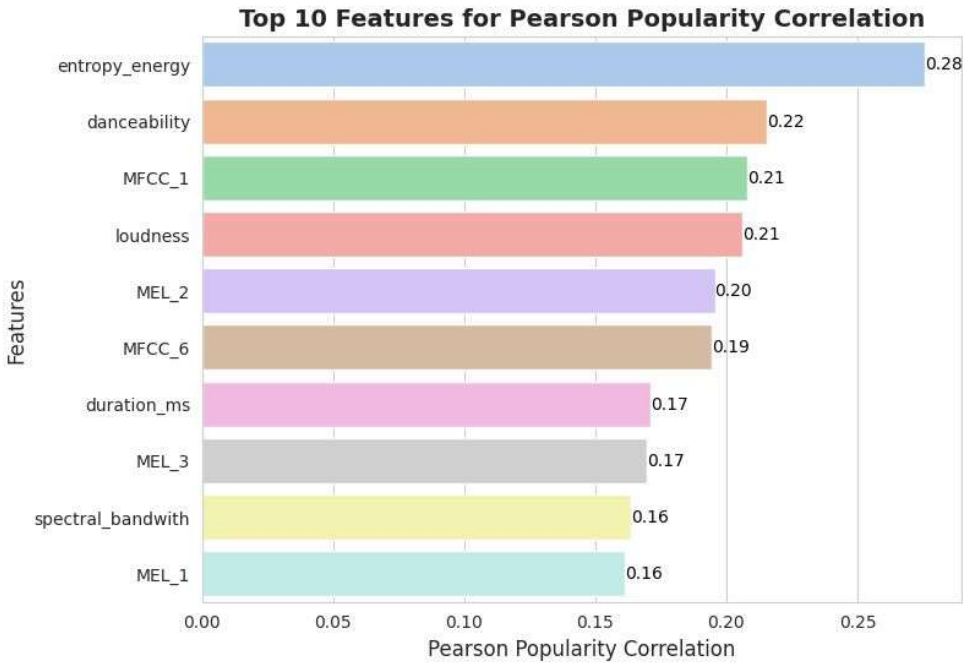
Outline

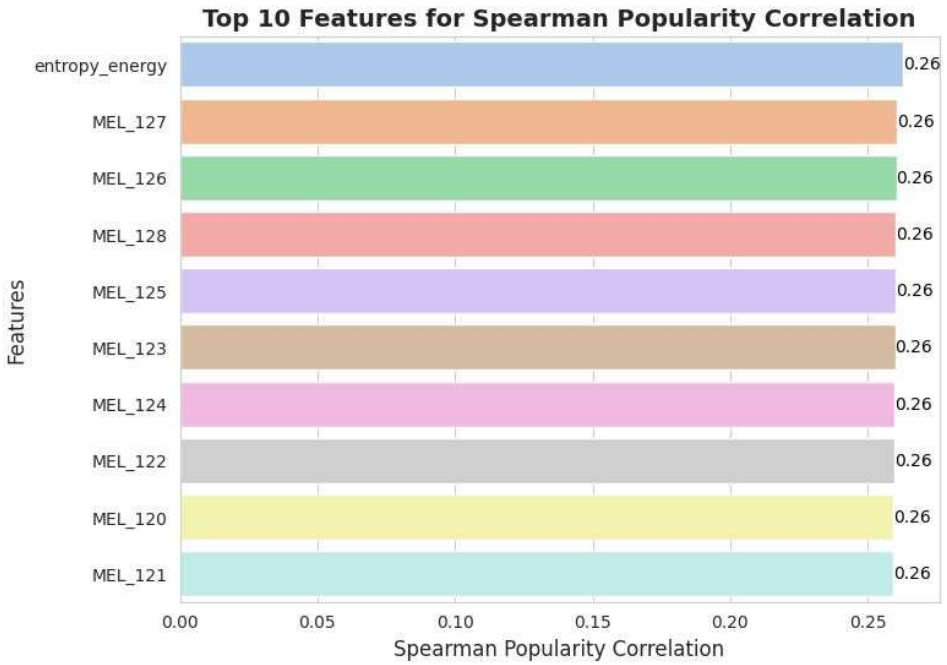

Motivation & Intro to the Dataset
01
Feature Extraction
02
Feature Selection
03
ML Model Prediction
04

Correlation Result Across All Music

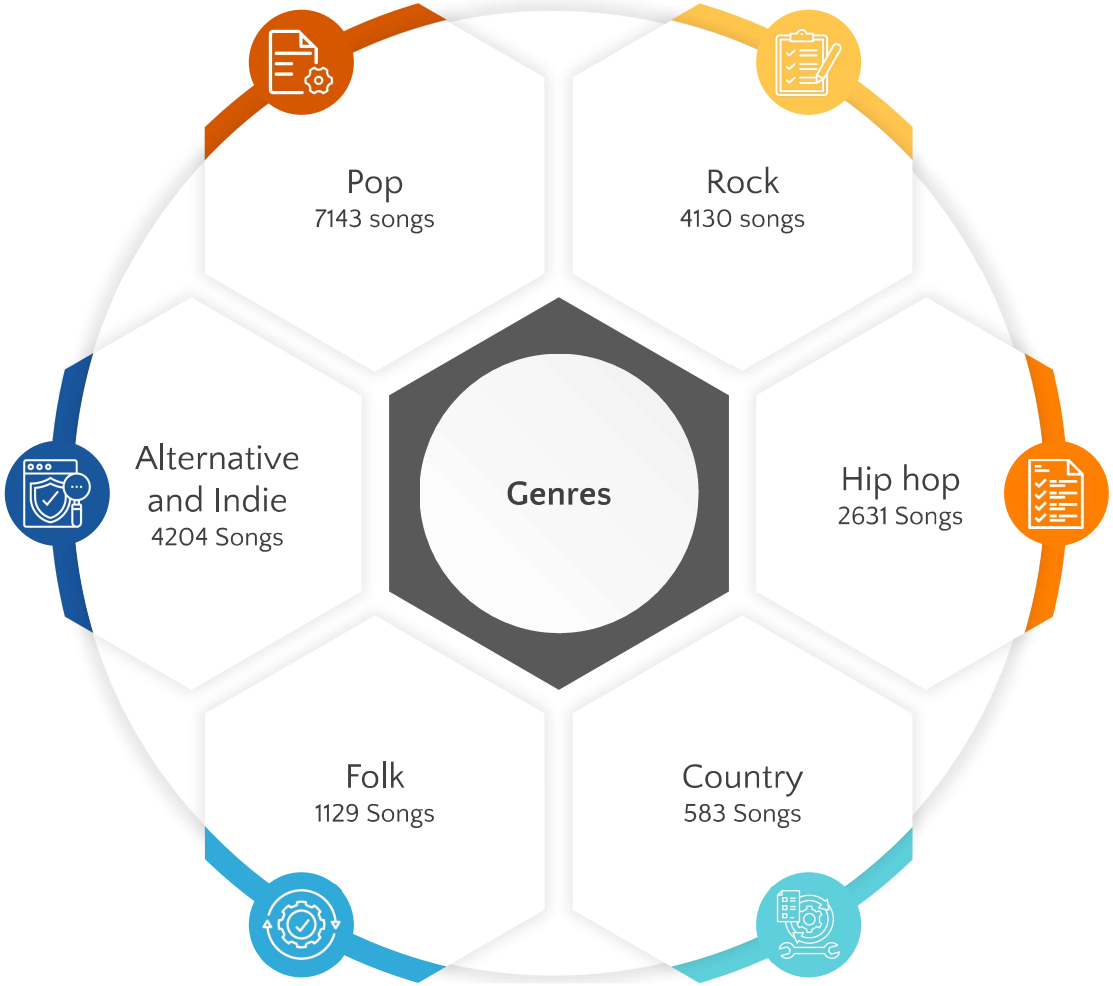
Pearson correlation: a straight line relationship of popularity to features

Spearman correlation: a rank order relationship of popularity to features

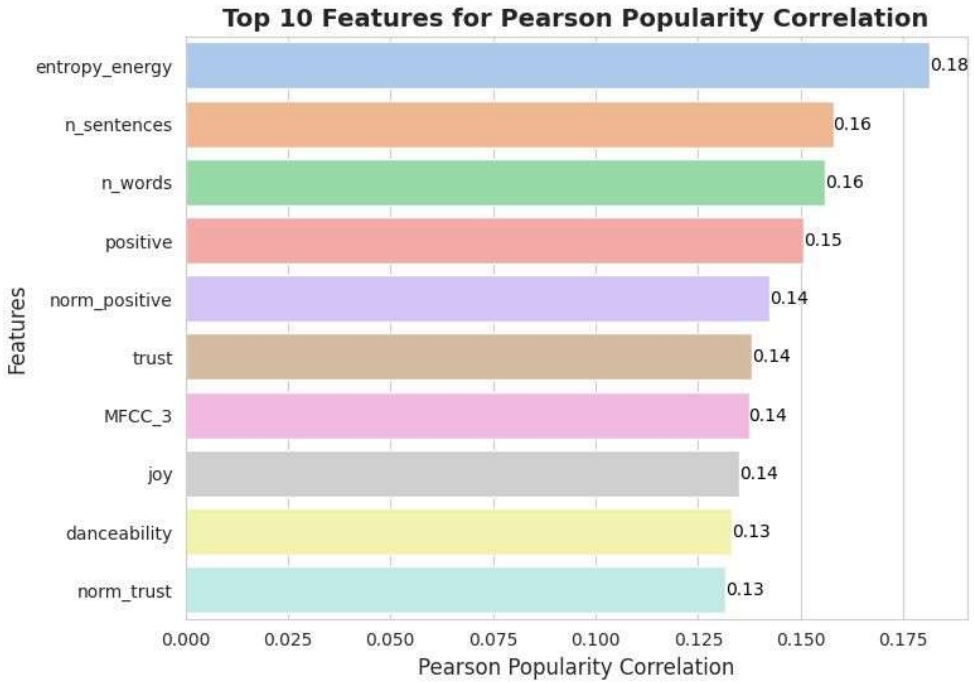


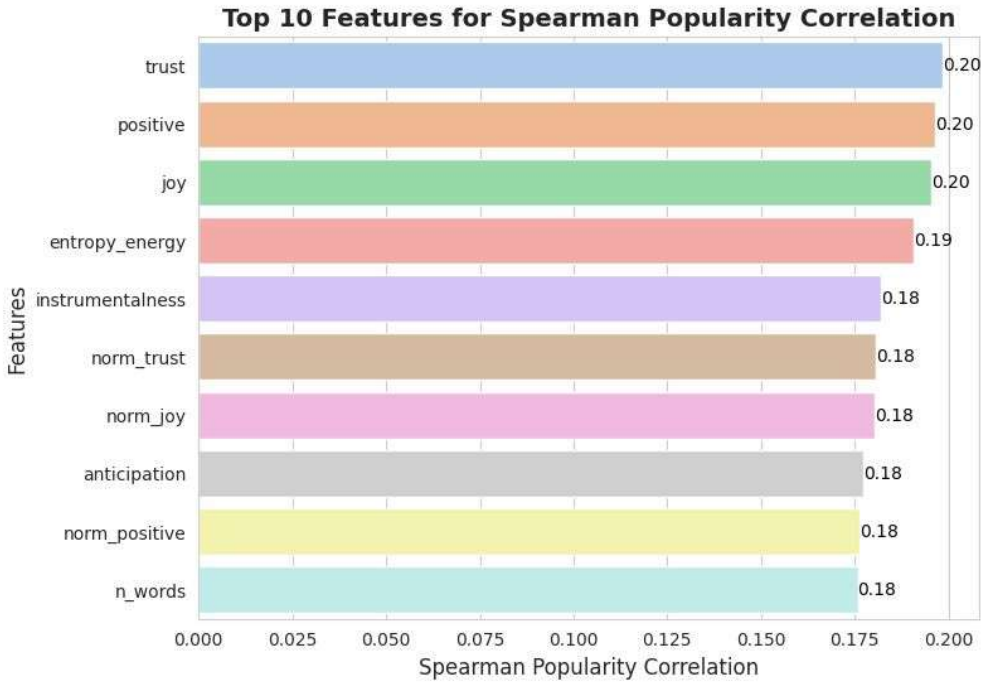


Dataset Clustering – Genres

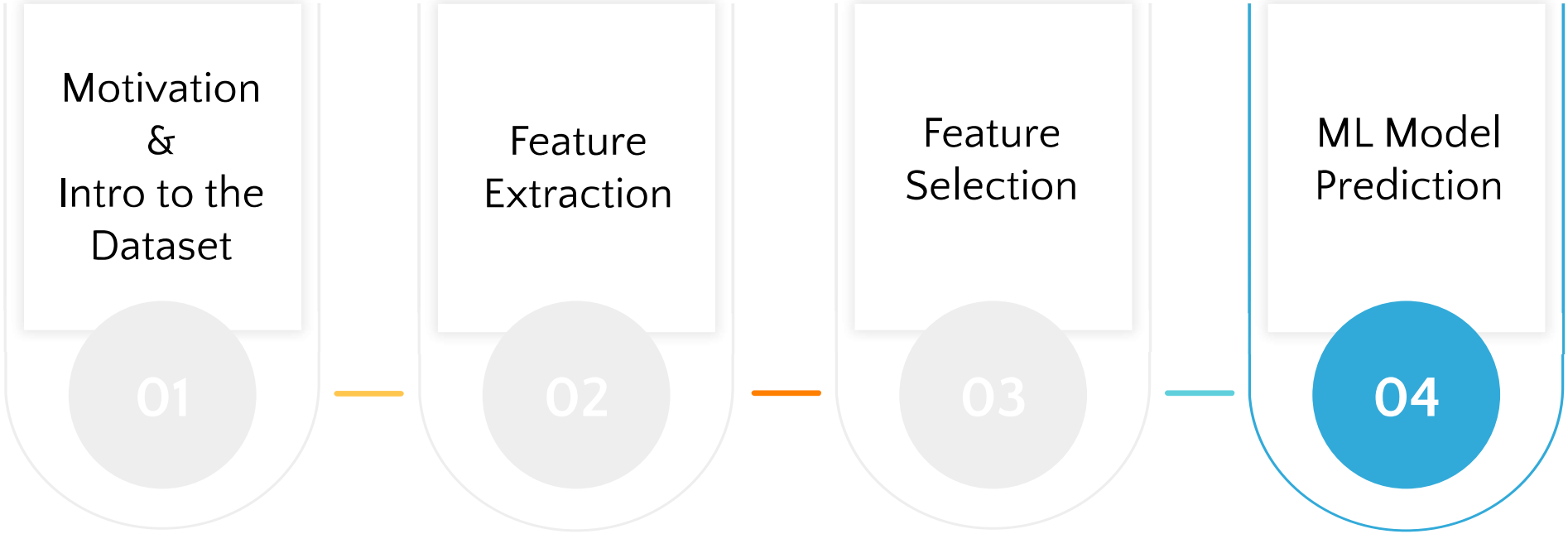


Correlation Results for Indie Music

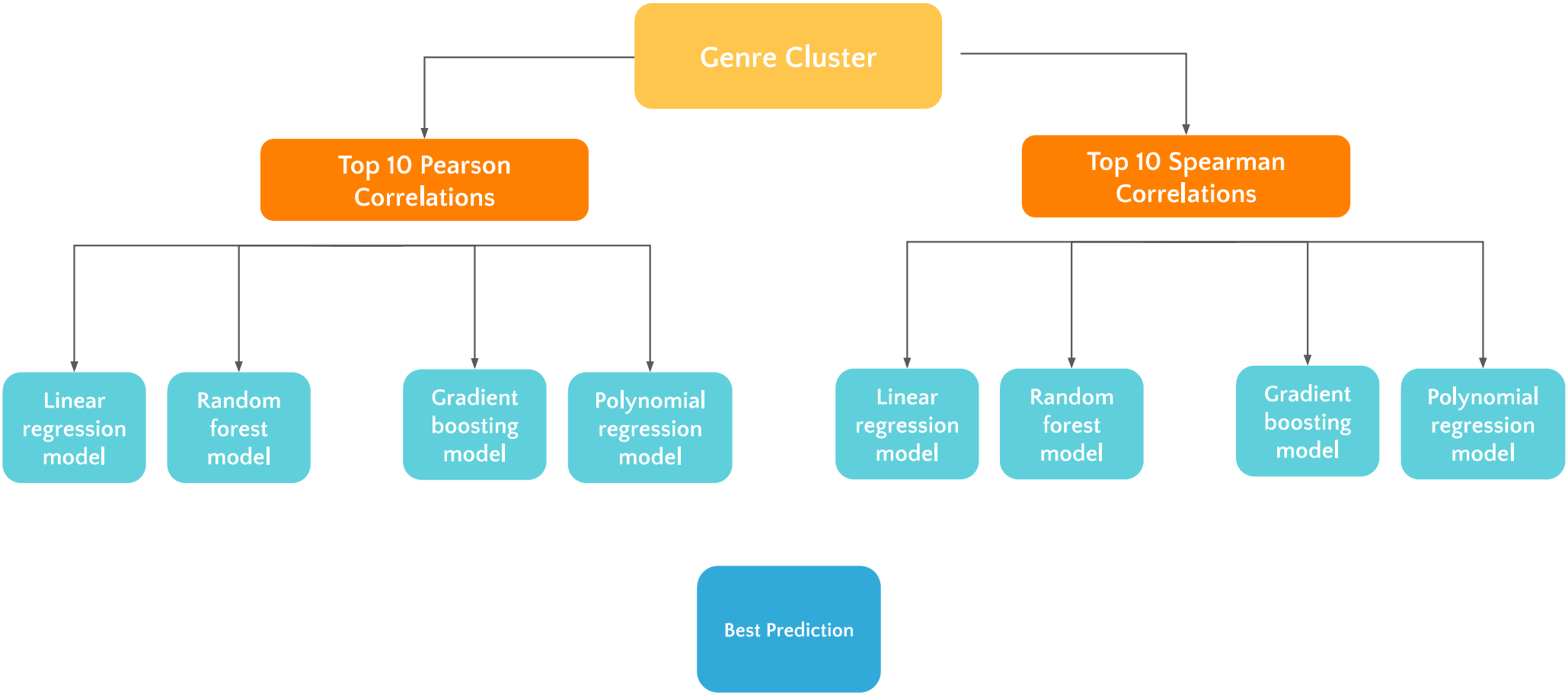




Outline


Motivation & Intro to the Dataset
01
Feature Extraction
02
Feature Selection
03
ML Model Prediction
04

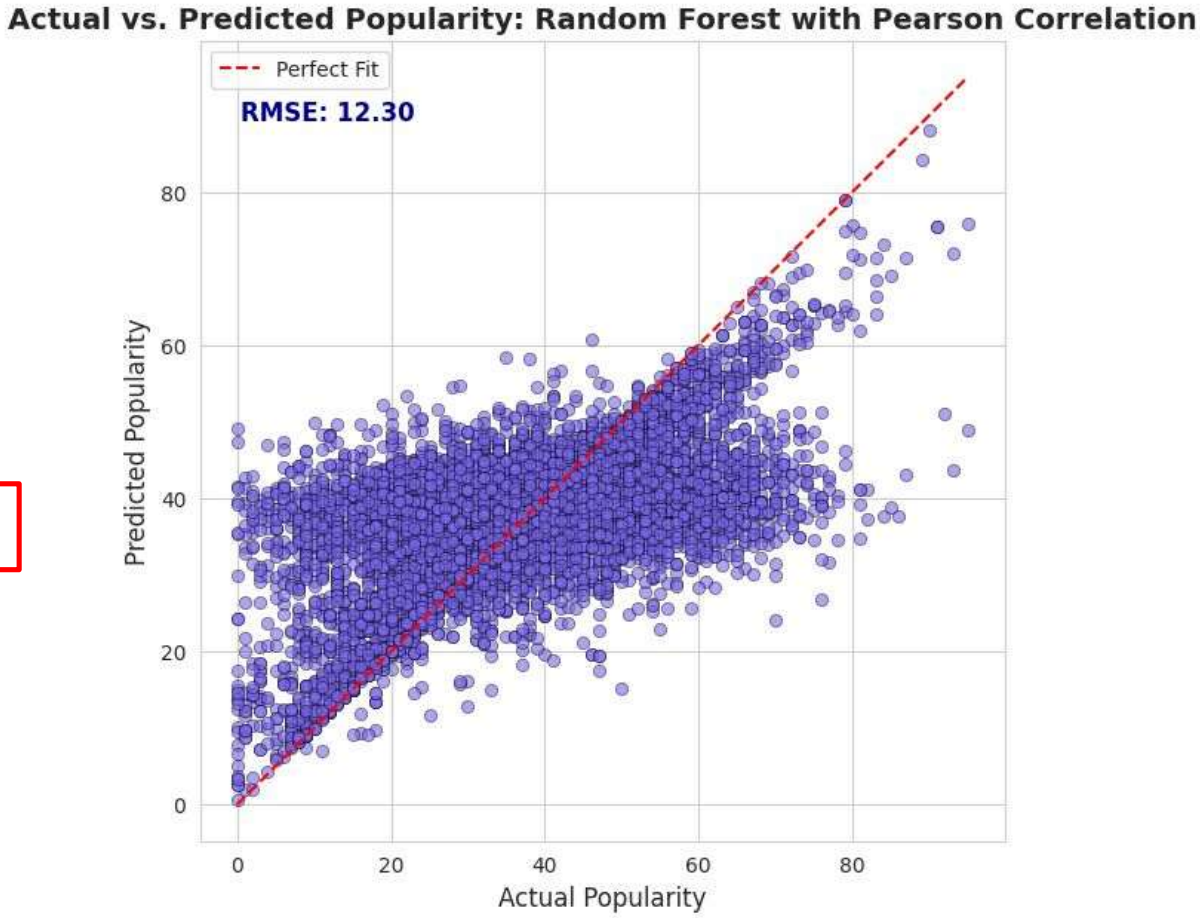
ML Model Predictions



Model Results Across All Songs

Using Top 10 Pearson Correlations

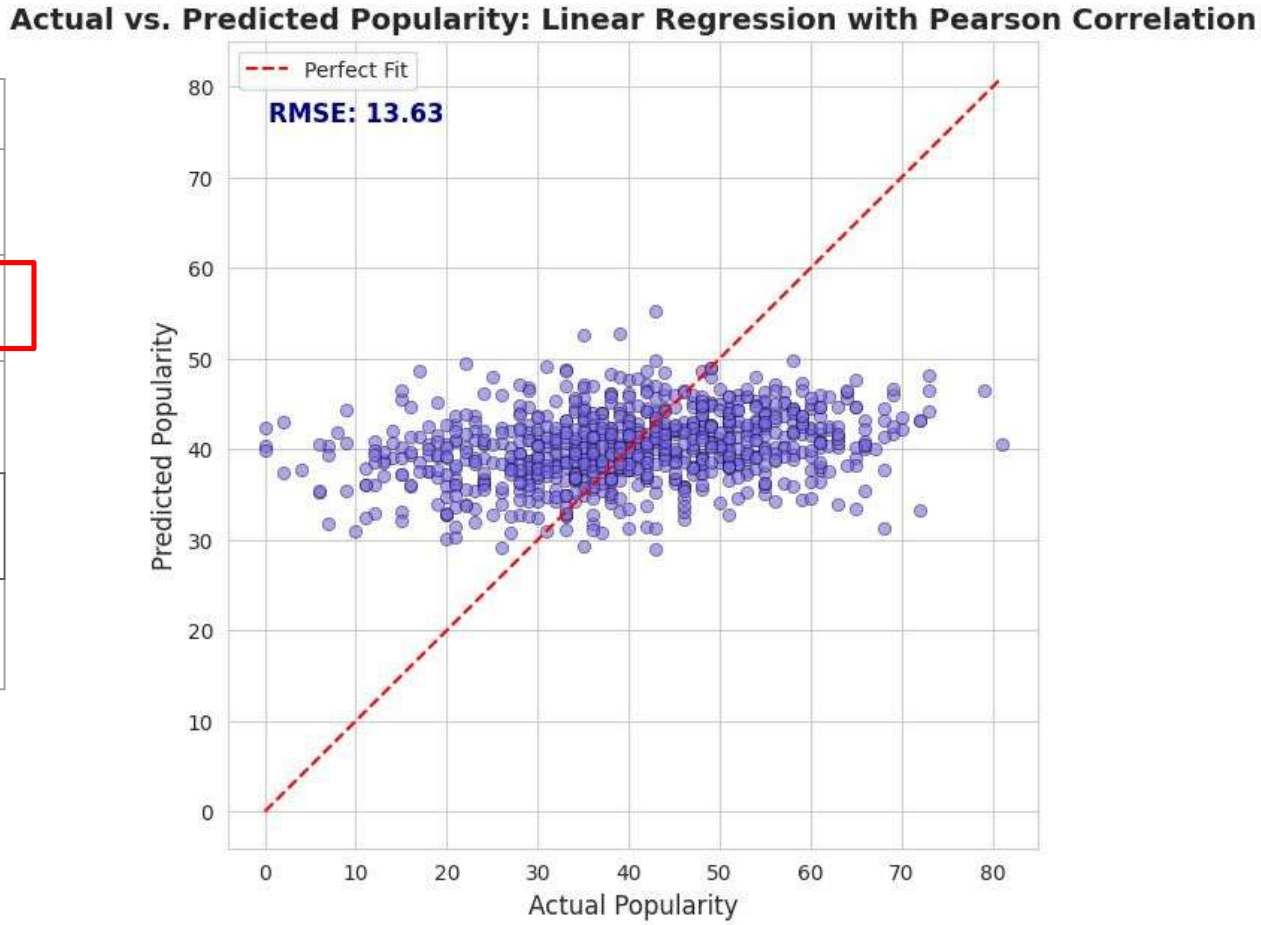
Model	RMSE		R2	
	Top 10	Random 10	Top 10	Random 10
Linear Regression	14.9	15.5	0.027	0.005
Polynomial Regression	15.0	17.6	0.015	-0.285
Random Forest	13.4	**15.9**	0.094	**-0.045**
Gradient Boosting	14.8	16.3	0.034	-0.193



Model Results Indie Songs

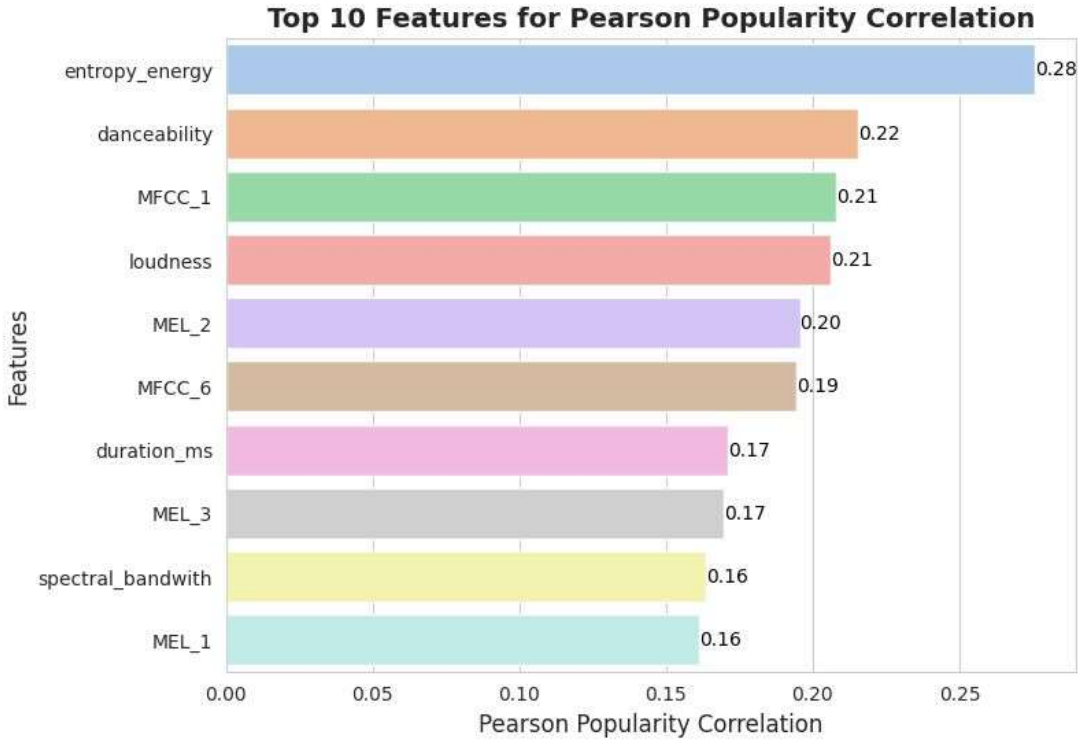
Using Top 10 Pearson Correlations

Model	RMSE		R2	
	Top 10	Random 10	Top 10	Random 10
Linear Regression	**13.63**	**14.11**	**0.078**	**0.01**
Polynomial Regression	13.69	14.42	0.069	-0.03
Random Forest	13.68	13.96	0.070	0.03
Gradient Boosting	13.97	14.11	0.022	-0.05



Summary

The popularity of songs is not confidently predictable from just lyrical and audio features.



🎵Music isn't meant to be overanalyzed — just hit play and vibe away🎵

Thank you for listening!!