Spotify Dataset

Data Mining SoSe 2022

July 20, 2022

Crucial Talking Points

DISCUSSION FLOW

1. Introduction
2. Sentiment Analysis
3. Genre
4. Popularity
5. Recommender System
6. Learnings and Outlook

1. Introduction

The Spotify Dataset

WHY SPOTIFY?

- Music brings people together.
- Helps to group music & playlists → Improve listeners experience
- Popularity prediction of a songs interesting for music industry

THE DATASET – TOP 200 CHARTS

ARTIST

- artist name
- number of followers
- …

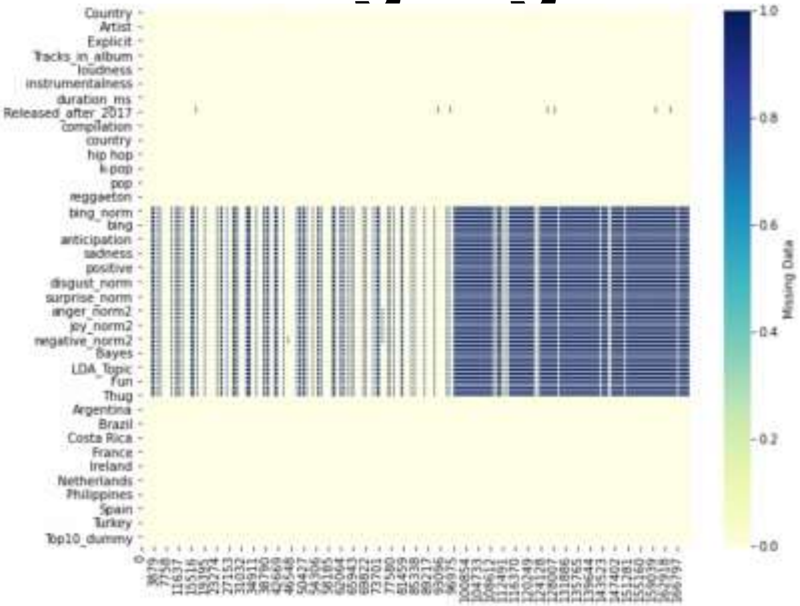
SONG

- song characteristics (duration, mode, …)
- genre
- topics / emotions
- …

PLAYLIST

- country
- popularity in the respective country playlist
- ranking position
- …

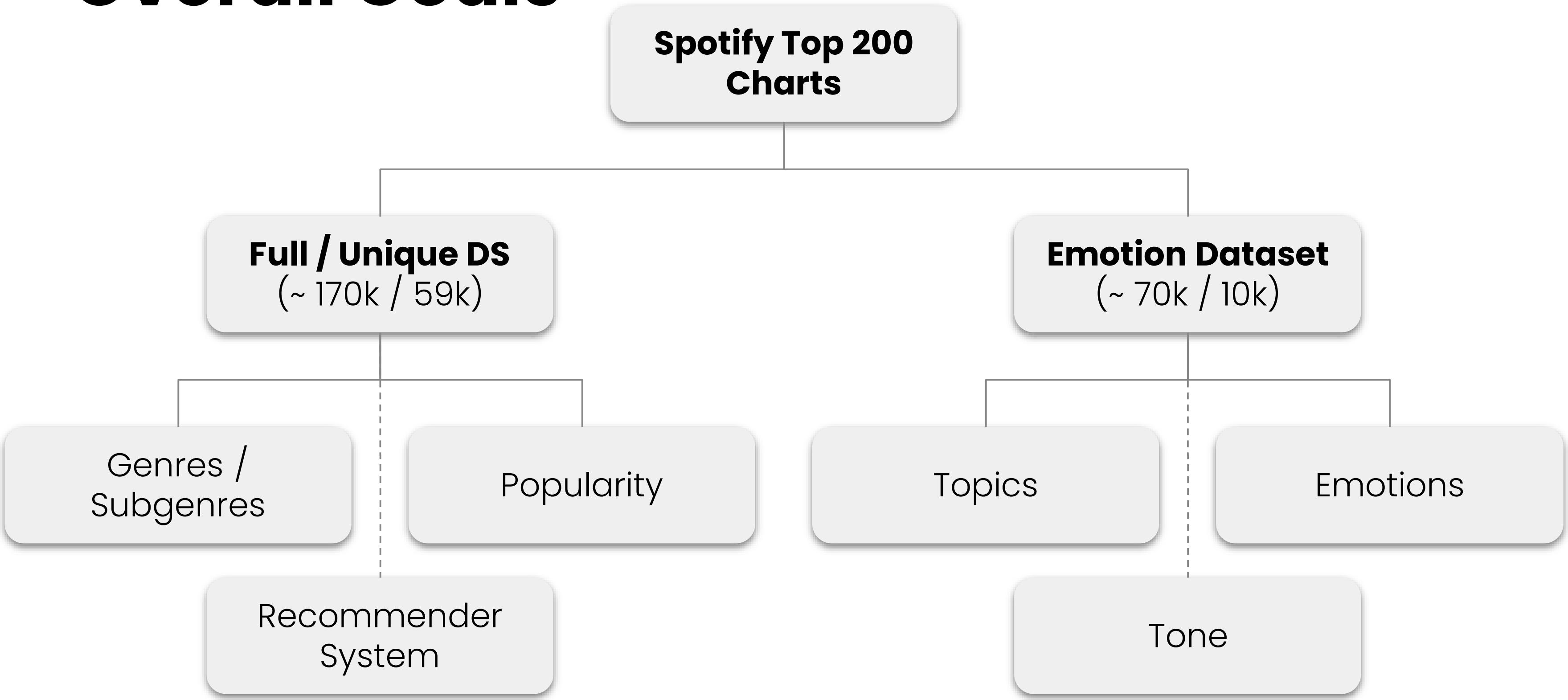
Data Wrangling



	% missing
genre	5.140272
sub_genre	2.316082
days_since_release	1.888263
released_after_2017	1.888263
duration_ms	0.011135
album	0.002930
uri	0.002930
tracks_in_album	0.002930
release_date	0.002930
track_number	0.002930
explicit	0.002930
artist_followers	0.002930
release_type	0.002930
artist	0.002930
title	0.002930

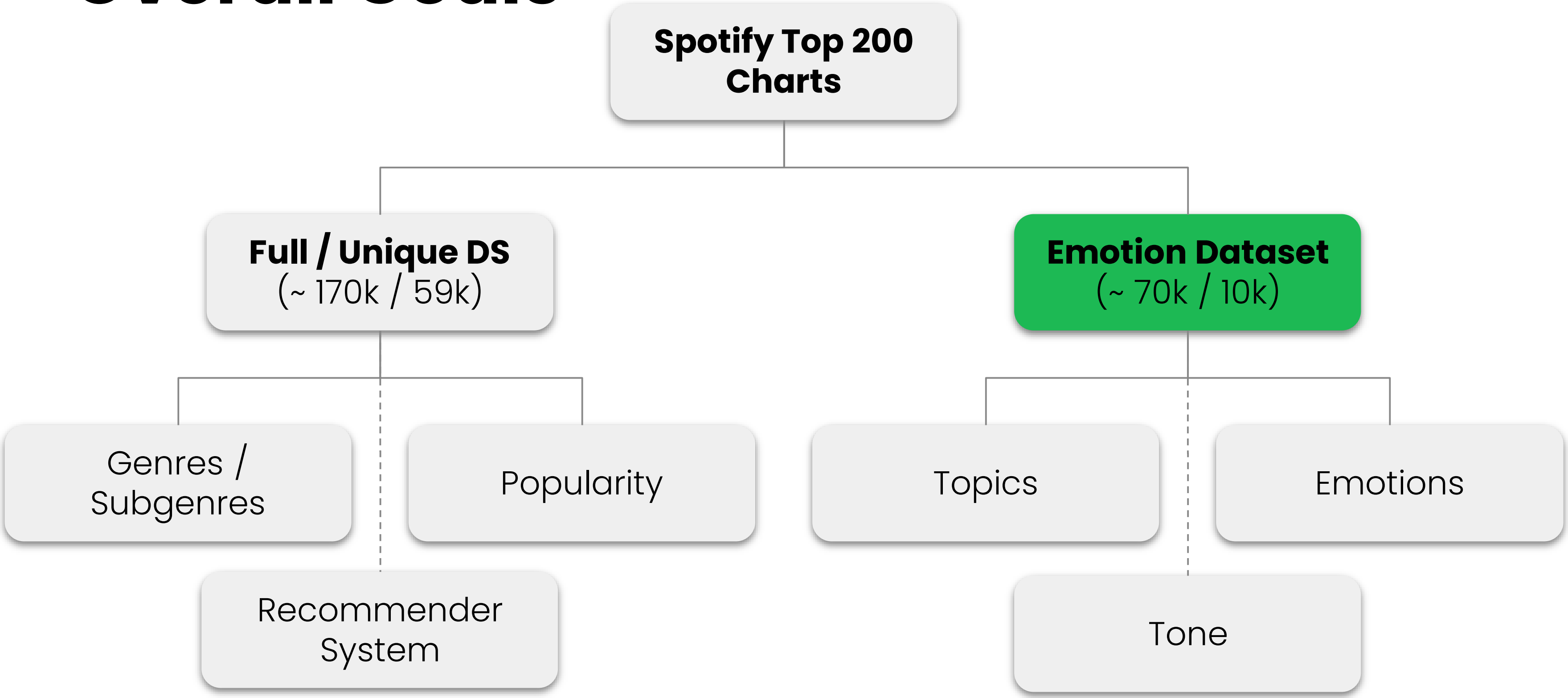
- Dropped all samples with 0.0029 %
- Imputed values with mean/mode
- Splitted datasets & created a unique song DS

Overall Goals

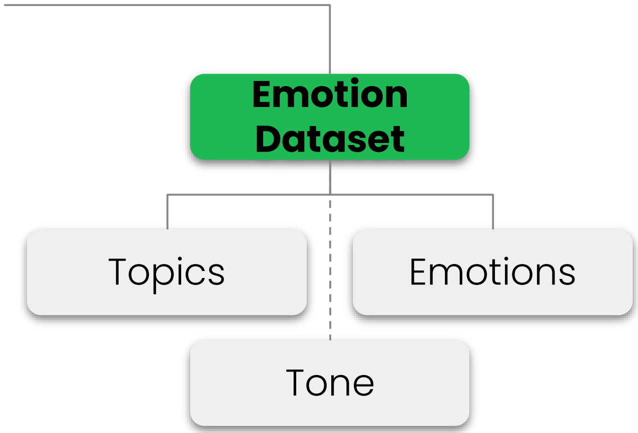


2. Sentiment Analysis

Overall Goals



Emotion Dataset



GOALS & QUESTIONS

- **Question 1:** What are popular topics?
- **Question 2:** Which emotions are characteristic for which topic?

-> Can we derive the topic of songs based on emotions?

Emotion Dataset

Q 1: WHAT ARE POPULAR TOPICS?

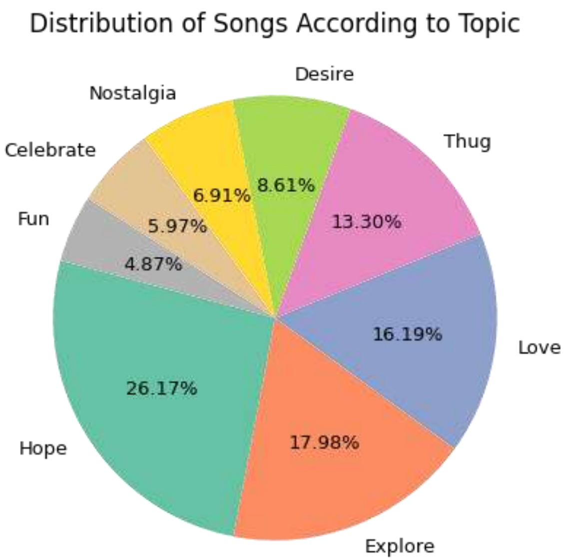
APPROACH

- Investigate the distribution of topics

PROBLEMS

- We only have emotional information for 41% of the Top 200 dataset
- Without duplicates: barely 9.5k songs

-> Statements about popularity are risky



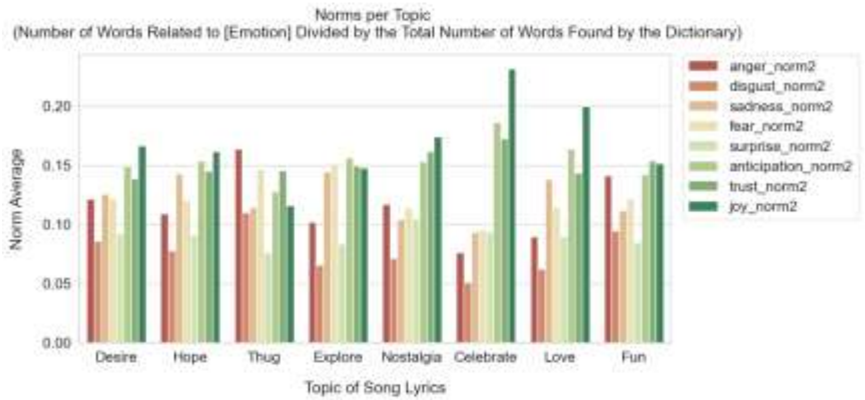
Emotion Dataset

Q 2: TOPICS VS EMOTIONS?

APPROACH

- Remove duplicate songs
- Investigate the distribution of emotions among topics

-> Tendencies but not clearly distinguishable



PROBLEM

- Uneven topic distribution

Further Remarks

Sentiment Analysis of Texts (e.g. Lyrics)

- Natural language processing problem
- Counting positive and negative words = very naive
- Does not consider word combinations

- General sentiment analysis challenges:
  - Context
  - Irony & Sarcasm
  - …

Language is very complex!

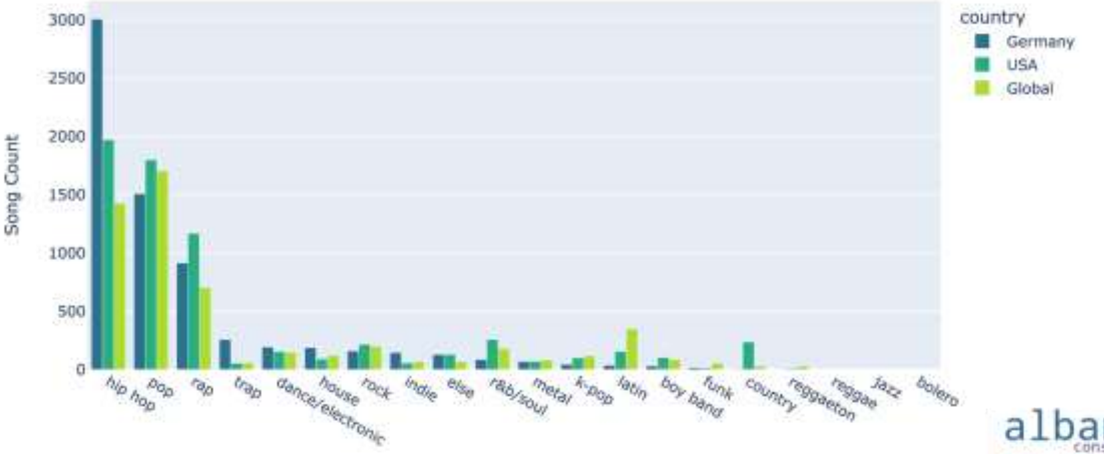
3. Genre

Overall Goals

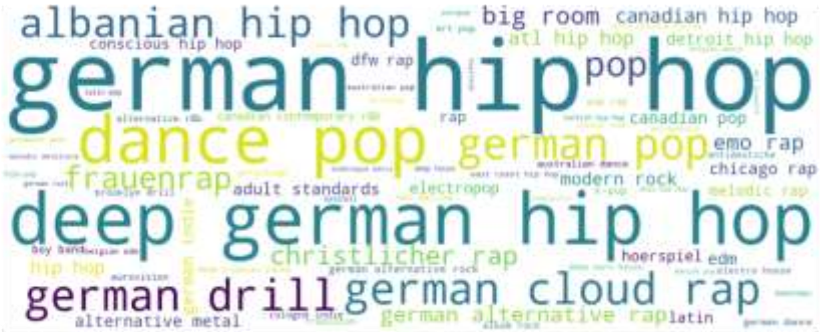
- **Spotify Top 200 Charts**
  - **Full / Unique DS** (~ 170k / 59k)
    - Genres / Subgenres
    - Popularity
    - Recommender System
  - **Emotion Dataset** (~ 70k / 10k)
    - Topics
    - Emotions
    - Tone

Genre Exploration

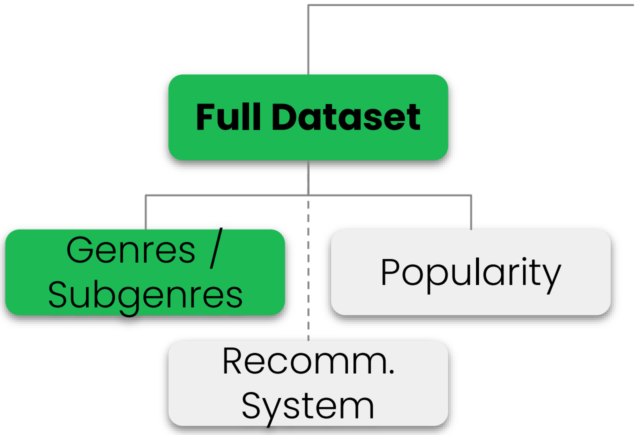
Genre distributions per country



Germany Subgenres



Genres / Subgenres

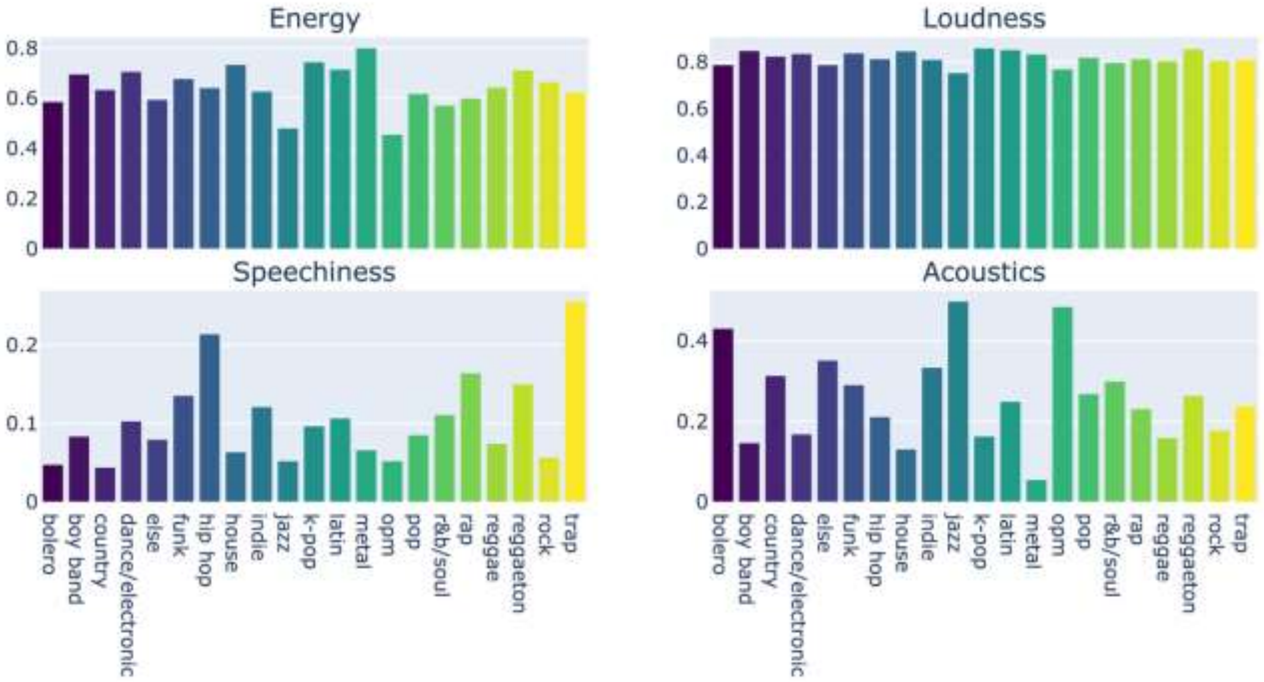


GOALS & QUESTIONS

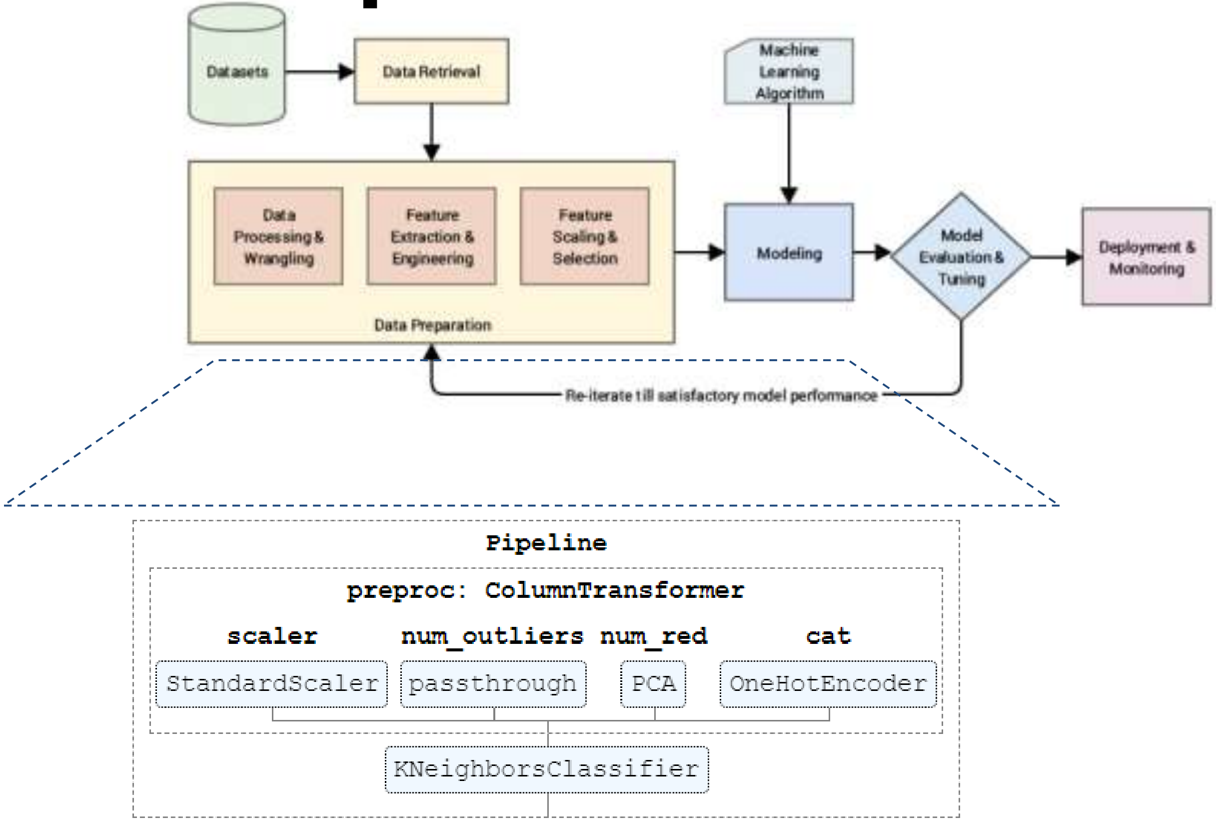
- **Question 1:** What characterizes a genre?
- **Question 2:** What and how many features can predict genres?
- **Question 3:** How heavily does the genre imbalance affect the predictions?
- **Question 4:** Can we see separable cluster in the feature space?

Genre Exploration

Q 1: What characterizes a genre?



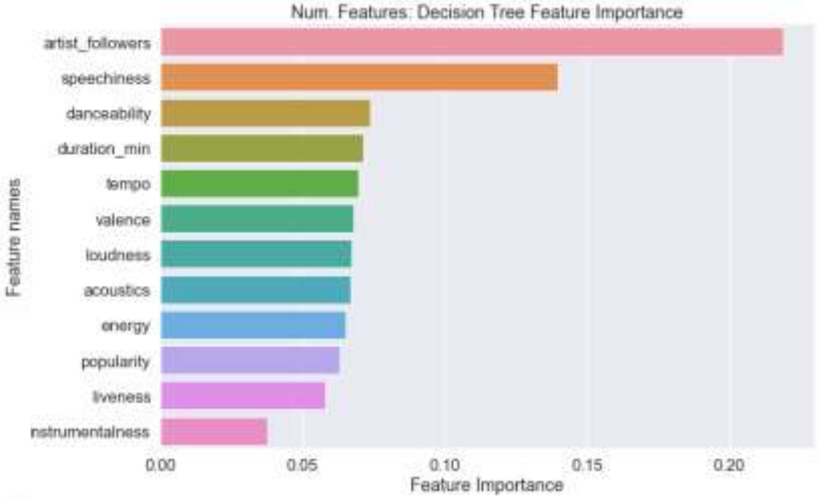
Predictive Pipeline

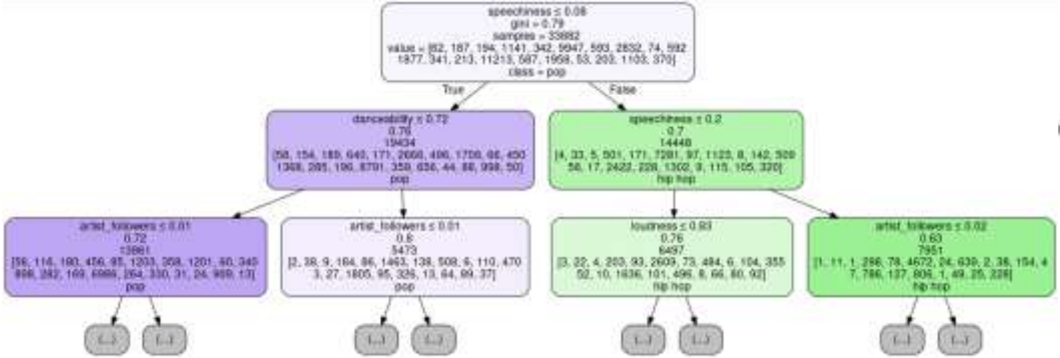


Simple Classifier

Q 2: What and how many features can predict genres?

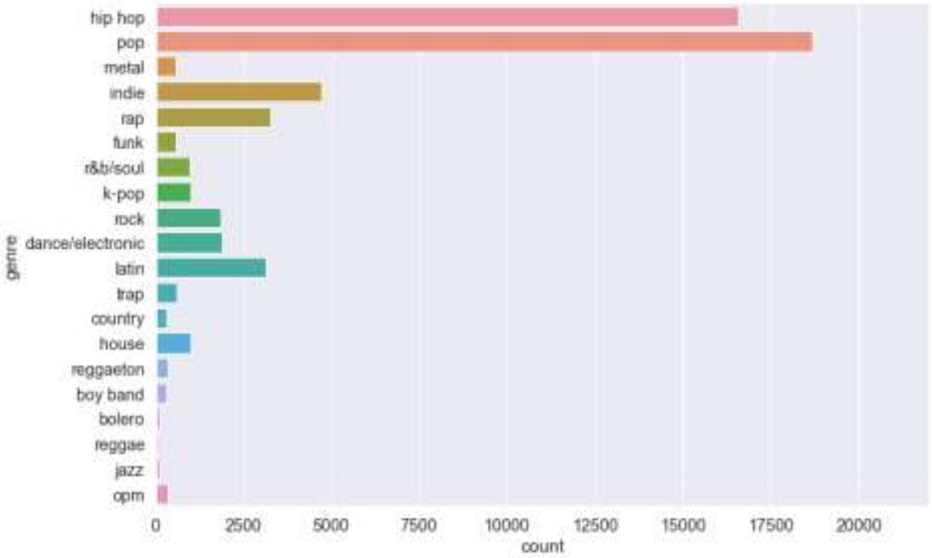
- First use **Decision Tree** for explainability
- F1-macro: ~0.35





Imbalancement

Q 3: How heavily does the imbalance affect predictions?



20 genres with a lot of variety of class sizes

Options we tried:

- Leave genres untouched
- Regroup and drop redundant genres
- Two vs Rest

Untouched Genres

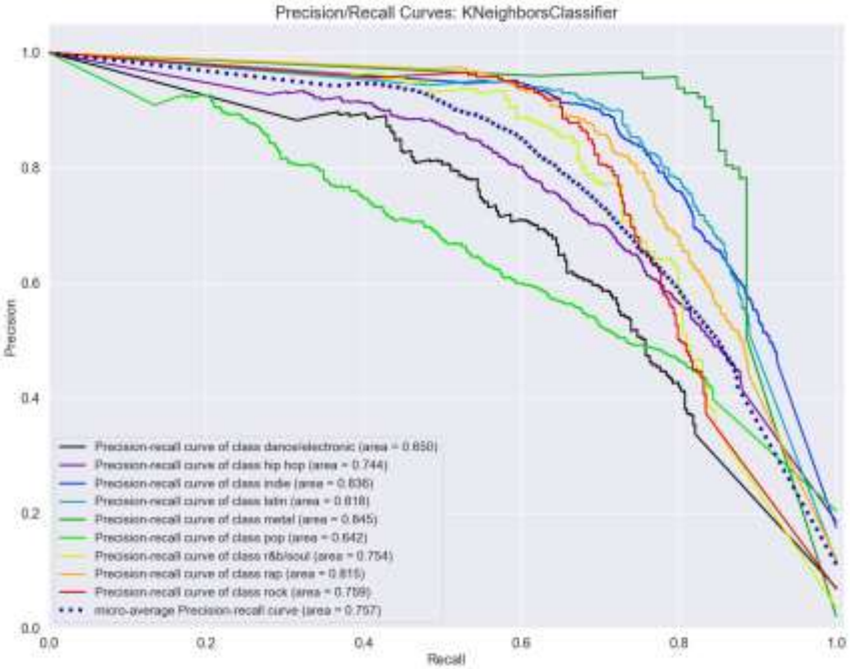
- Predictions for **20 genres** with 3 different classifiers
- Logistic Regression, RF, KNN
- Mixed f1 score for minority and majority genres

→ Selection: KNN

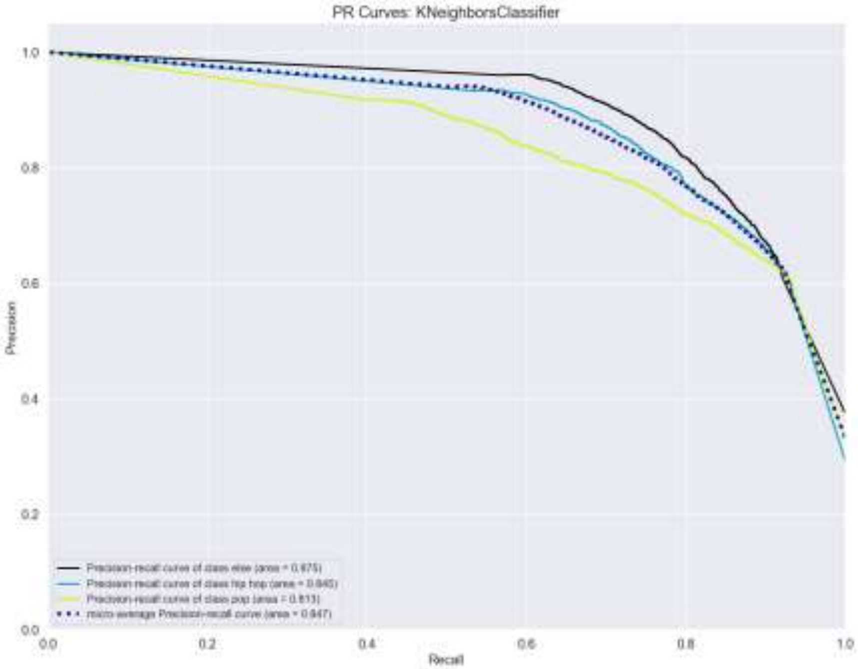
Tuned K-Neighbors Classifier				
Model Performance				
	precision	recall	f1-score	support
hip hop	0.733	0.819	0.773	3316
house	0.878	0.548	0.675	197
indie	0.782	0.668	0.721	944
metal	0.833	0.885	0.858	113
opm	0.710	0.310	0.431	71
pop	0.709	0.803	0.753	3738
...				
accuracy			0.747	11294
macro avg	0.806	0.633	**0.699**	11294

Classifiers

Reorganized and dropped (9 genres)



Two vs. Rest



Classifiers

- Compare rebalancing for pop and hip hop genres
- Tried 3 different options and trained a KNN classifier on each

	Untouched (**unbalanced**)	Reorganized & dropped (**balanced**)	Two vs. Rest (**balanced**)
Pop	0.77	0.67	0.79
Hip-Hop	0.75	0.59	0.76

f1-macro

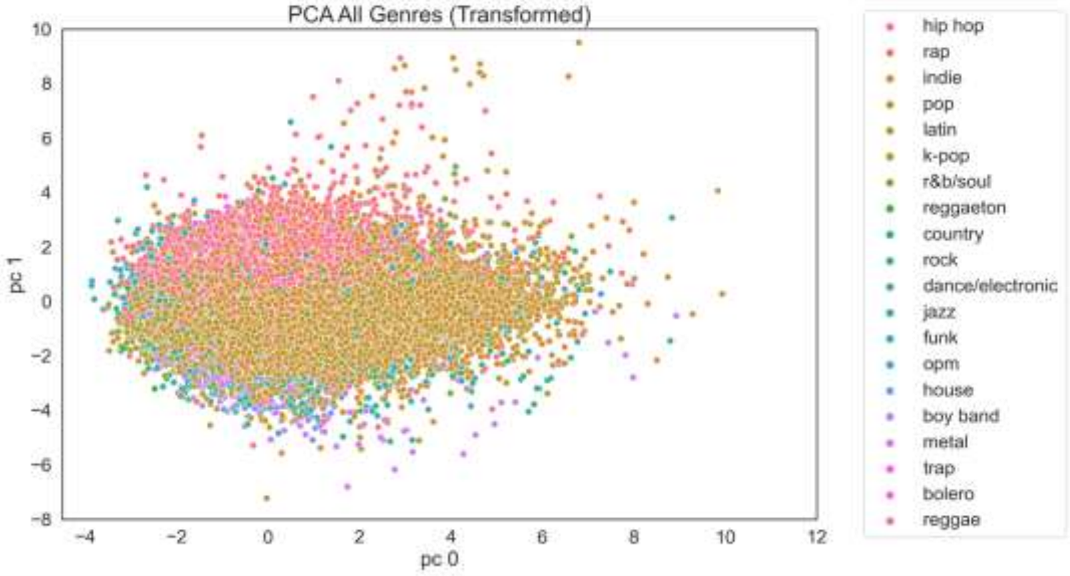
~ similar

Rebalancement doesn't affect performance of majority genres (pop/hip-hop)

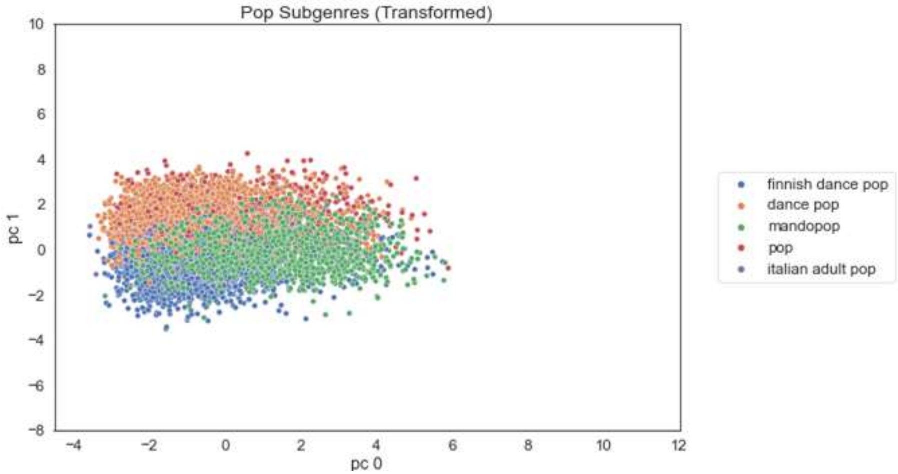
Genre Clusters

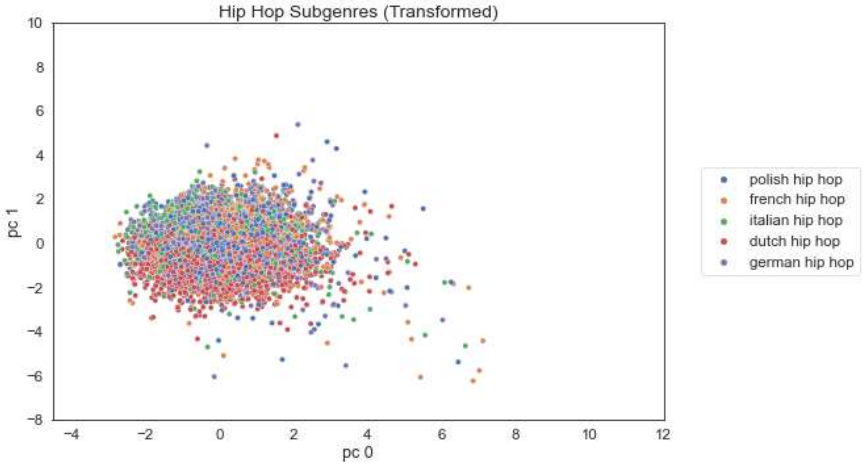
Q 4: Can we see separable clusters in feature space?

- Tried PCA in 2D &3D
  → **Hard to cluster**
- TSNE with different perplexity
  → Results didn’t improve



Subgenre Clusters



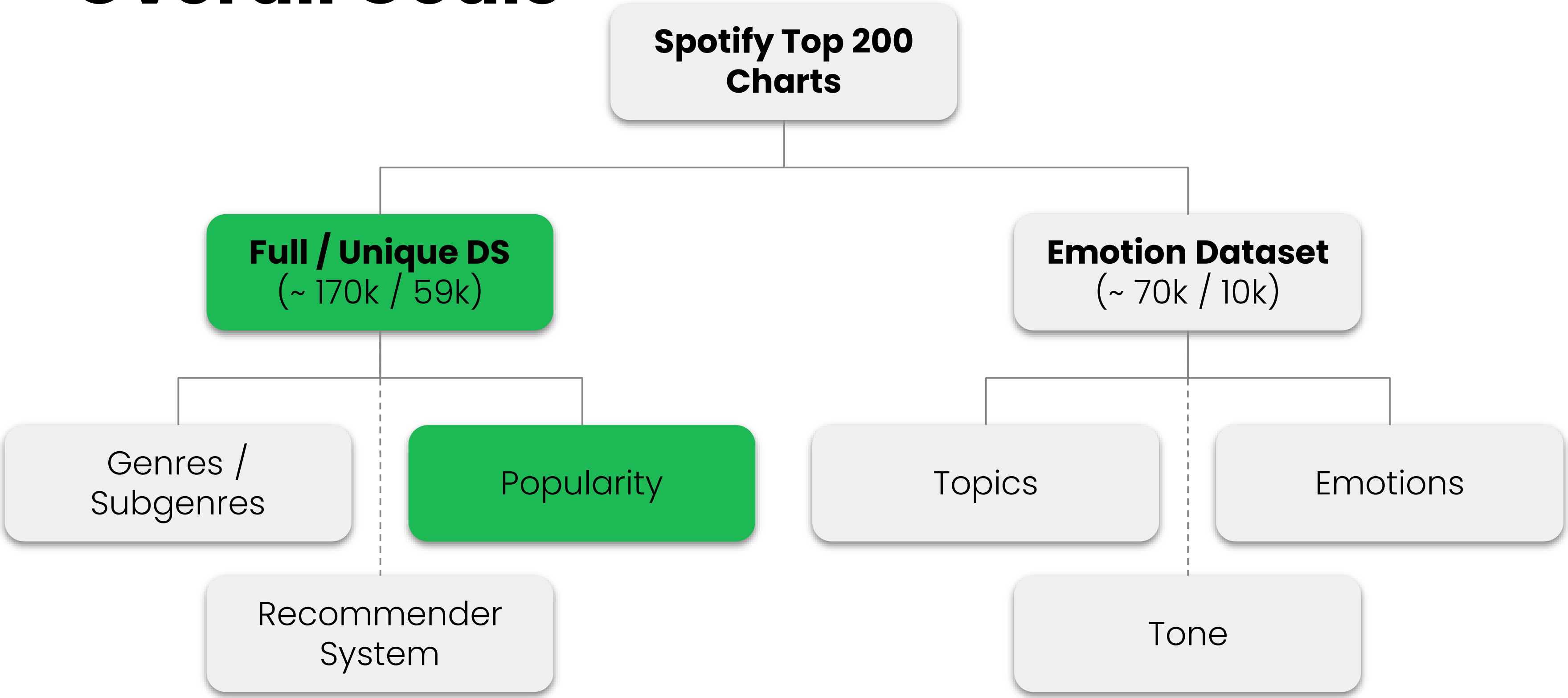


Key Learnings:

- Some majority subgenres have distinct groups
- Hip-Hop subgenres focus more on language characteristics than music features

4. Popularity

Overall Goals



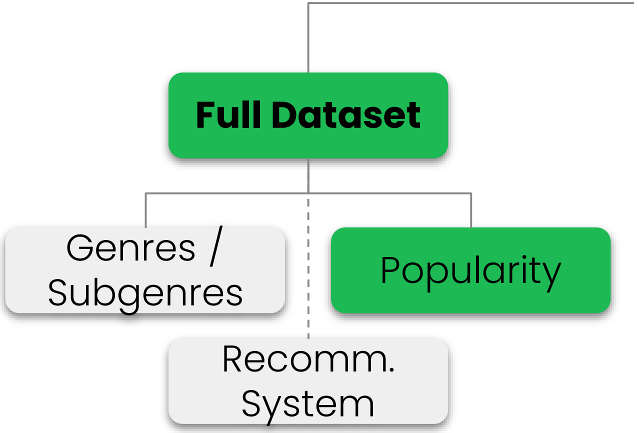
Popularity

	DATASET	SPOTIFY
Definition	Calculated **the number of days a song stayed in the Top200** and the **position** it stayed in everyday.	Calculated by algorithm and is based on the **total number of plays the track has had** and **how recent those plays are**.
Value	0 - 233,766.9 (highest in our dataset)	0 - 100

Any problems?

Not really…but we couldn't add any data anymore & the popularity score is not comparable

Popularity

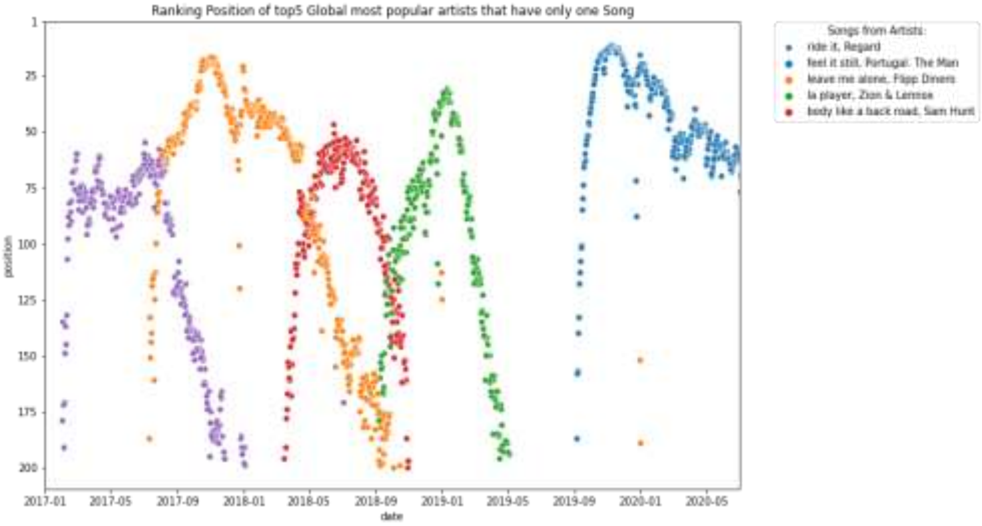


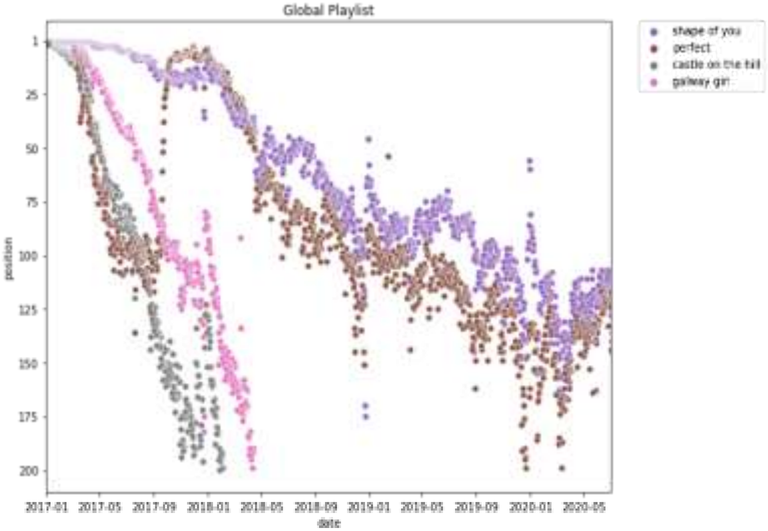
GOALS & QUESTIONS

- **Question 1:** How long do popular and not popular artist songs stay in the playlist position?
- **Question 2:** What is the average popular song duration & release day?
- **Question 3:** How reliable is our data to predict popularity?
- **Question 4:** Which features contribute the most to song popularity?

Popularity Exploration

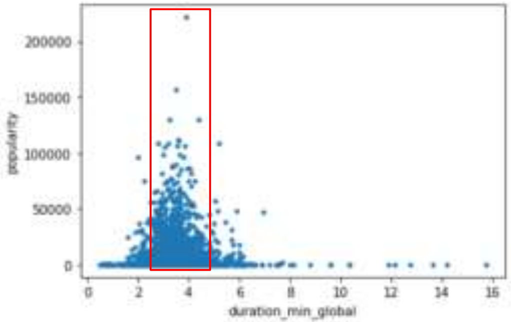
Q 1: How long do popular and not popular artist songs stay in the playlist position?

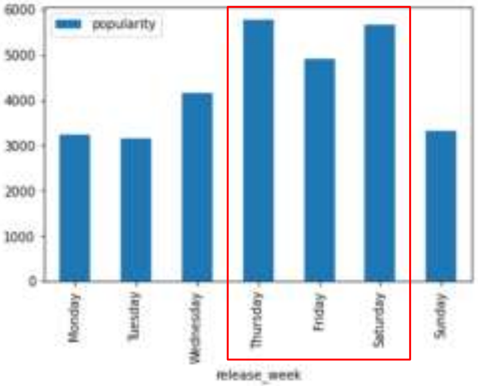




Popularity Exploration

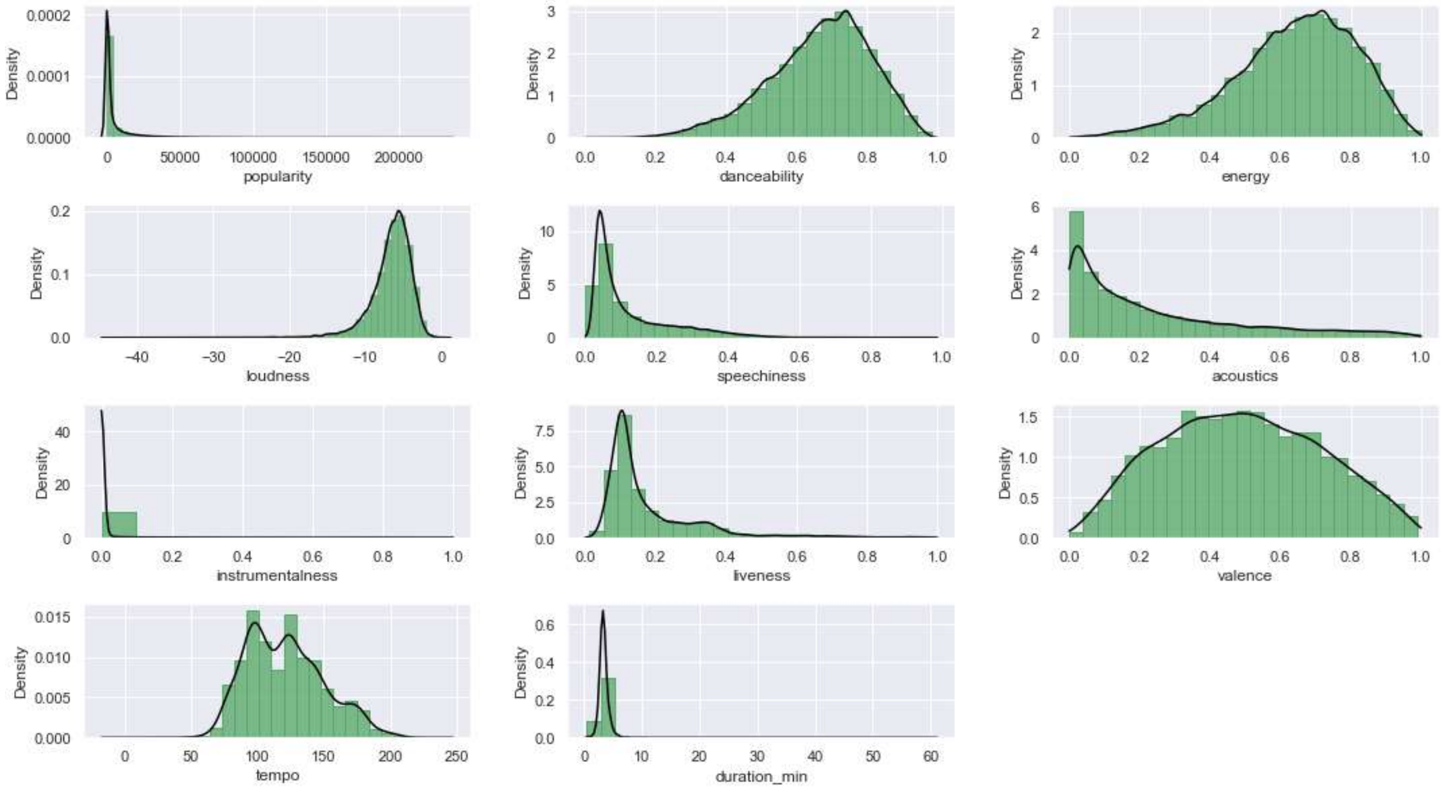
Q 2: What is the average popular song duration & release day?



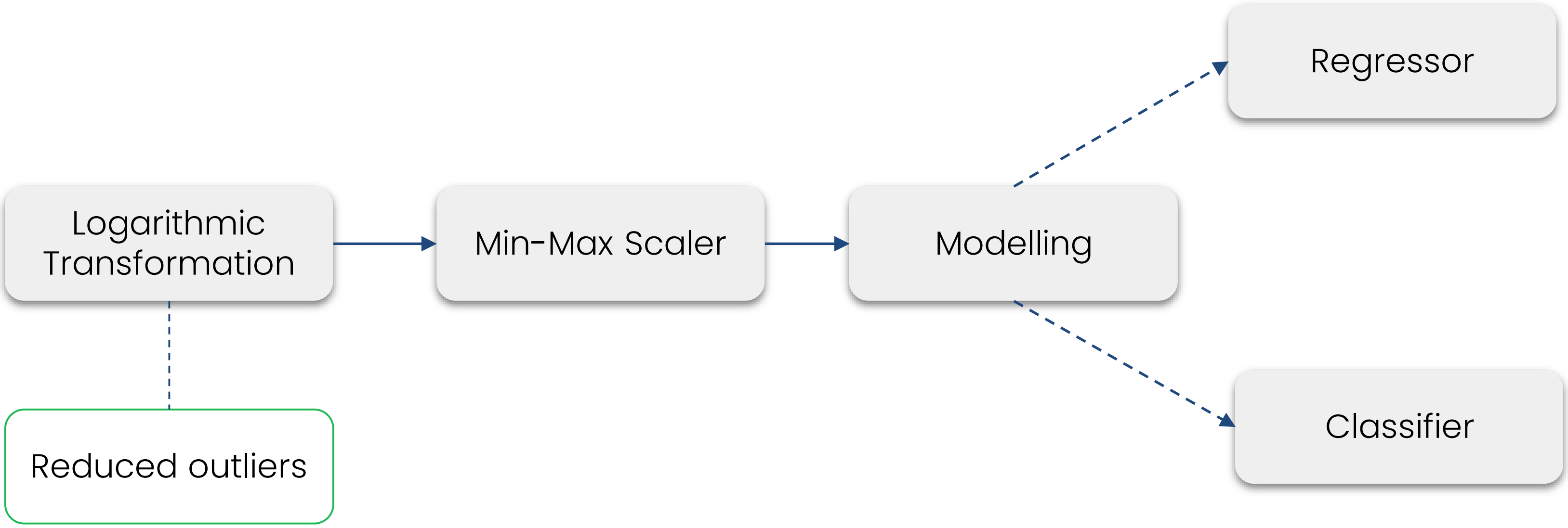


- Most popular songs released on **Thursday, Friday & Saturday**
- Popular song duration is around **2.5 - 4.5** mins

Numerical Features Overview



Predictive Pipeline



Popularity Regression

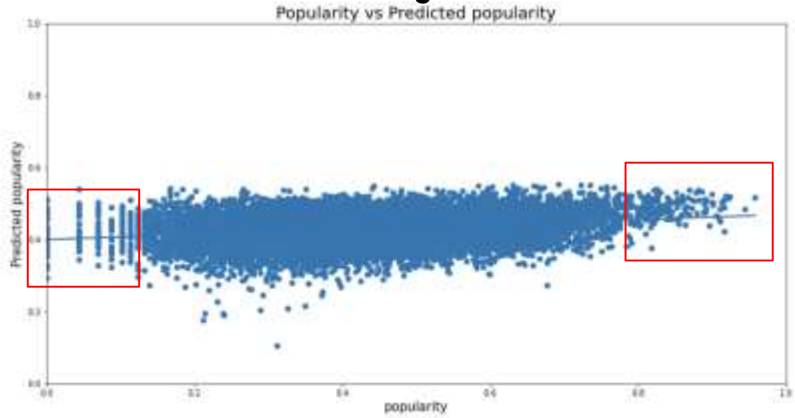
Q 3: How reliable is our data to predict popularity?

- The MAE result is very low (within range 0-1)
- Predicted popularity score doesn't go above 0.6 and below 0.1

Not reliable for extremely low and high popular songs

Model	MAE	MSE	RMSE
Linear Regression	0.130738	0.025311	0.159093
Decision Tree Regressor	0.174970	0.048383	0.219961
Tuned Decision Tree Regressor	0.131463	0.025565	0.159891

Linear Regression

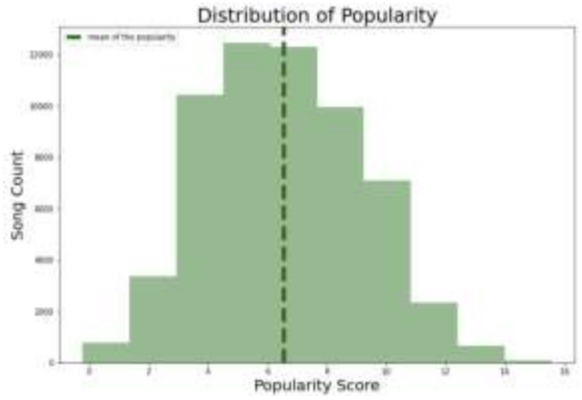


Popularity Classification

APPROACH

- Song labelling -> Popular: 48.1%, Not Popular 51.9%
- Used song numerical features*
- Added song categorical features**

Model	Accuracy*	Accuracy (Tuned)**
Logistic Regression	0.59	0.59
Decision Tree	0.56	0.59
Random Forest	0.61	0.62

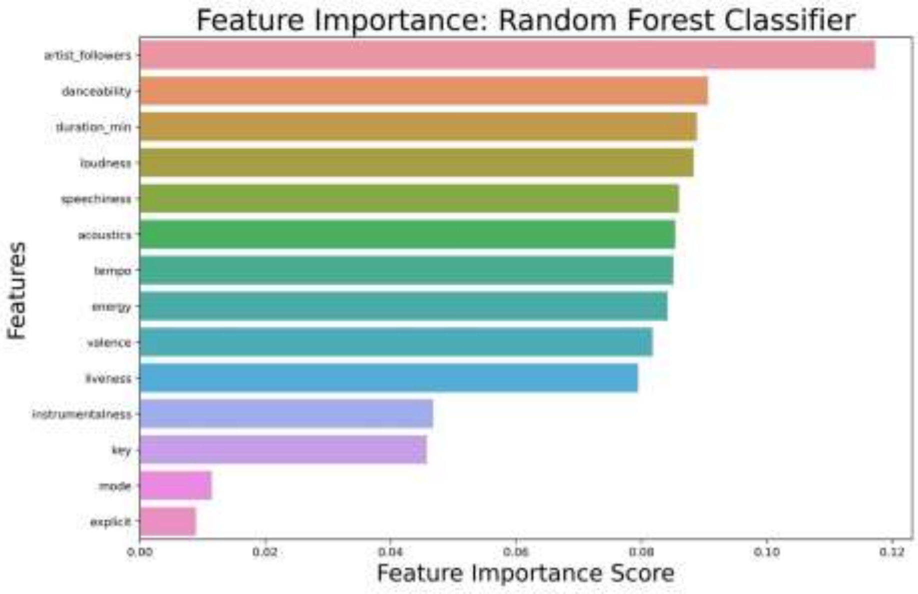


Popularity Classification

Q 4: Which features contribute the most to song popularity?

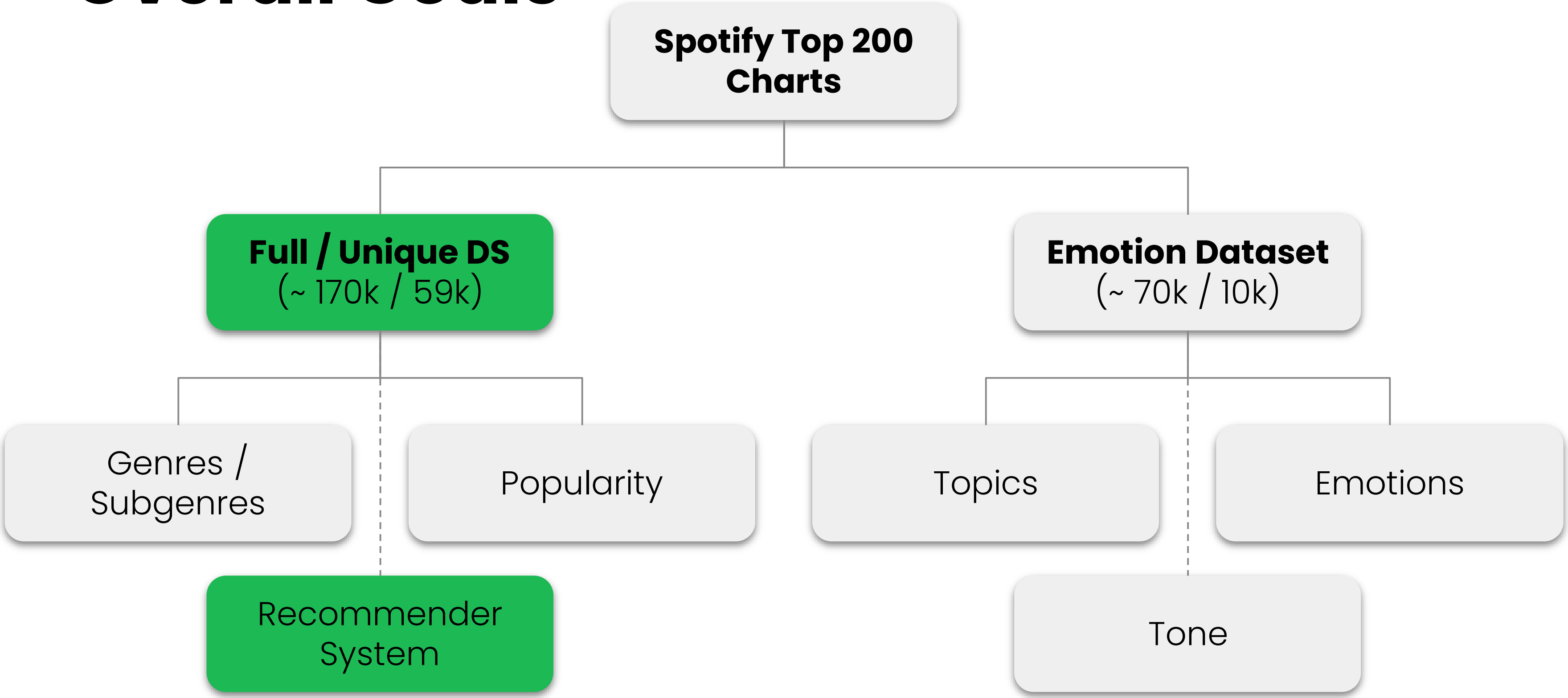
However, the gap still remains..

There are other factors that we didn't take into account, like lyrics, artist collaboration, language, etc.



5. Recommender System

Overall Goals



Overview

Building a recommender system for the songs prediction

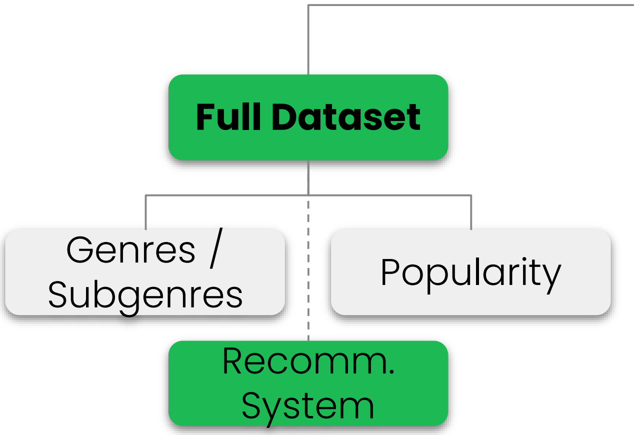
What kind of recommender system:

Collaborative filtering ➡ memory based ➡ item-item

Techniques used for building it:

- One hot encoding/bag of words/TF-IDF
- Cosine similarity
- Euclidean distance

Recommender System



GOALS & QUESTIONS

- **Question 1:** What would be good features for a RS?
- **Question 2:** What are the challenges for our simple RS in comparison to the conventional RS ?

Feature selection

Q 1: What would be the good features for RS?

Song info features

- Provides logical similarities

- artist
- title
- genre
- country

Songs characteristics features

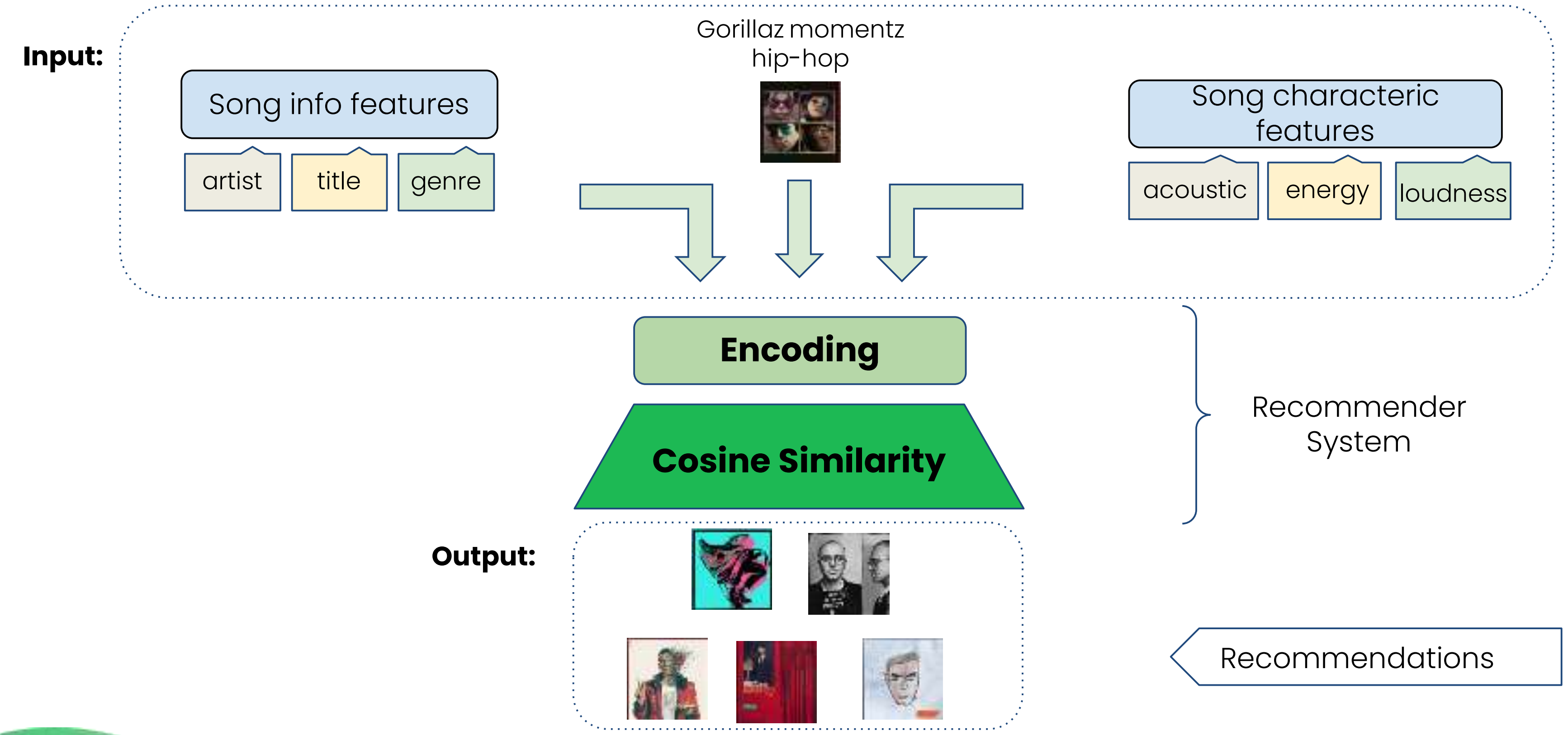
- Defining the characteristics/elements of the songs

- danceability
- acoustics
- loudness
- energy
- mode
- ...

correlation →

genre, popularity

Workflow of RS



Showcase of RS

Input song:

momentz by Gorillaz feat. De la Soul

Gorillaz  Kansas 0.83	Gorillaz feat. Jenny beth  We Got the Power 0.74	Logic  everybody dies 0.72
Logic  keanu reeves 0.72	Salmo  ho paura di uscire 0.71	Eminem  Stepdad 0.71

Challenges

Q 2: What were the challenges in our RS in comparison to a conventional RS?

- Verifying good features
- Target audience
- Validating results/metric
- Lacking features like feedback, click rate, etc
- Keeping it unbiased

6. Learnings and Outlook

Lessons Learned

- Picked a dataset with lower complexity that was required for the tasks

  → **First have the goal in mind, then gather approp. data**

- Should have cleaned data before the prediction task
  - Not the beginning of EDA

- With raw audio data we could have extracted lyrics or trained an LSTM Model for genre prediction

  → **Some music features are not expressive enough**

Teamwork

ORGANIZATION

COMMUNICATION

- semi-fixed meetings
- collaborative milestone planning

PLANNING

- miro (brainstorming)
- notion (important links and notes)

- follow an overall goal
- refine milestones together

CODE SHARING

- github
- tried: google colab, kaggle

Tools used:
- pandas & sklearn pipelines, 3d vis. in plotly, …
- Spotify API for developers

Meet our Team

Abhi

Informatics

İlayda

Informatics

Isi

Informatics

Maxi

RCI

Norma

DEA

THANK YOU