SPOTIFY
PREDICTION AND RECOMMENDATION SYSTEM

Bug Killer
Group members:
Yixuan Chen, Yushi Dai,
Zhizhen Xie, Muchen Liang

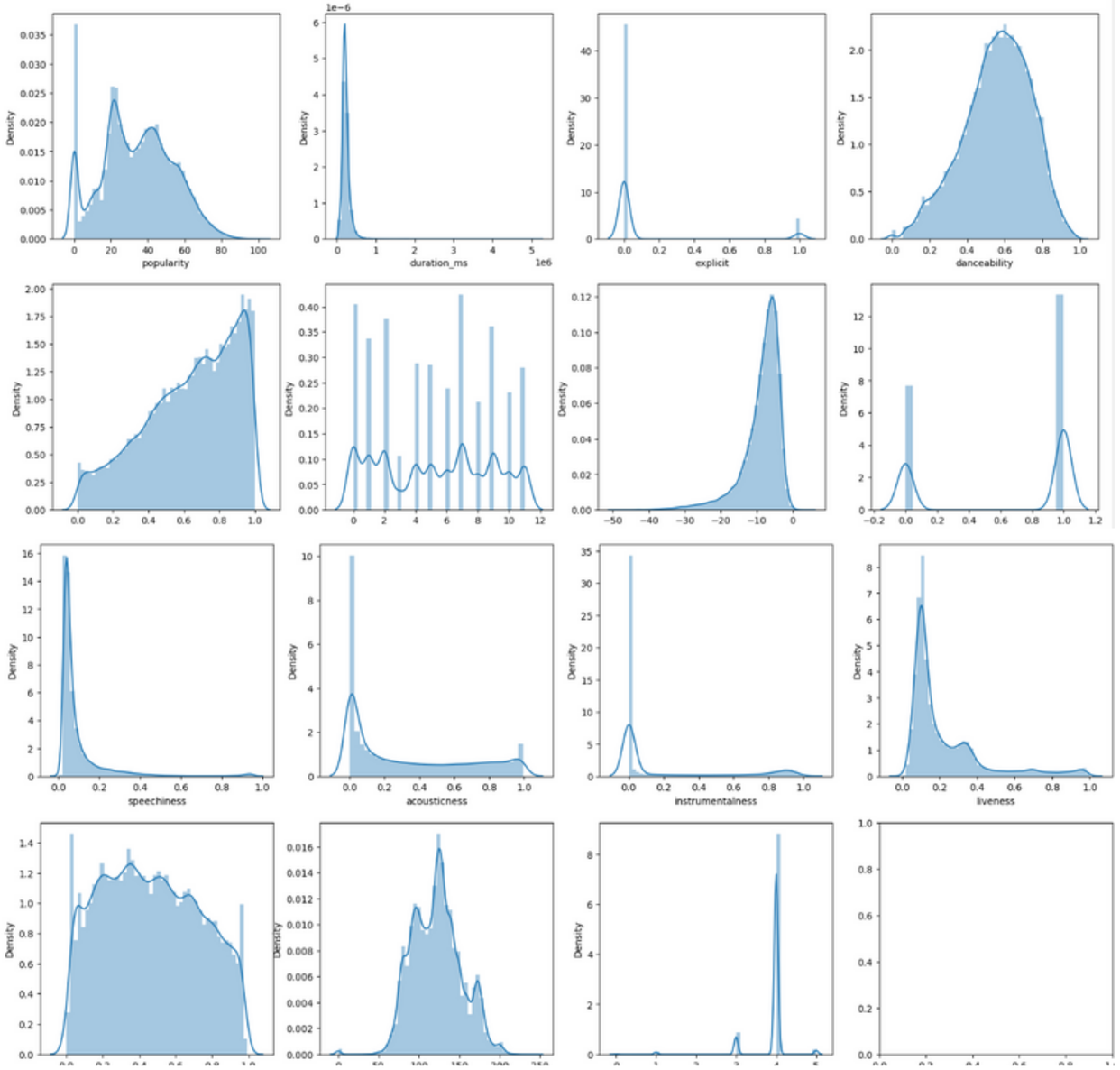
CONTENTS

DATA

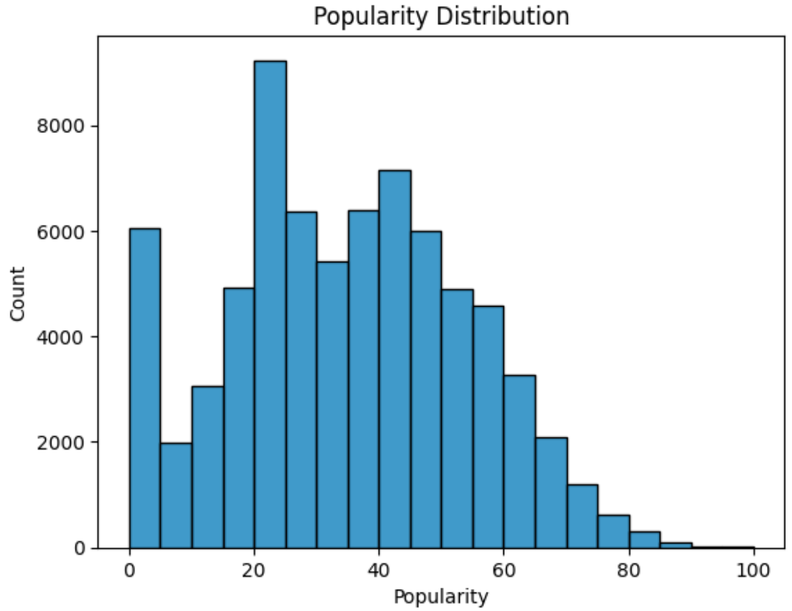
DATASET
Spotify Tracks Dataset from Kaggle

DATA CLEANING
Preprocessing the data to remove duplicates and missing values

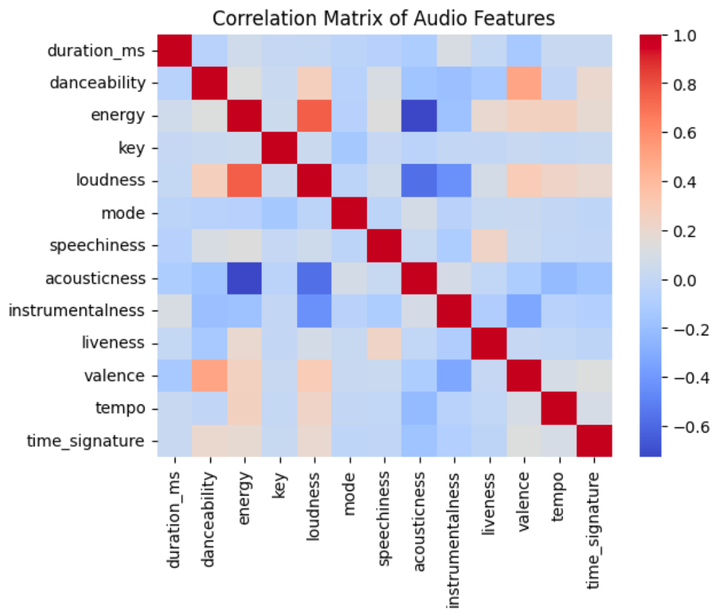
DATA TRANSFORMATION
Binary Encoding categorical variables 'explicit'



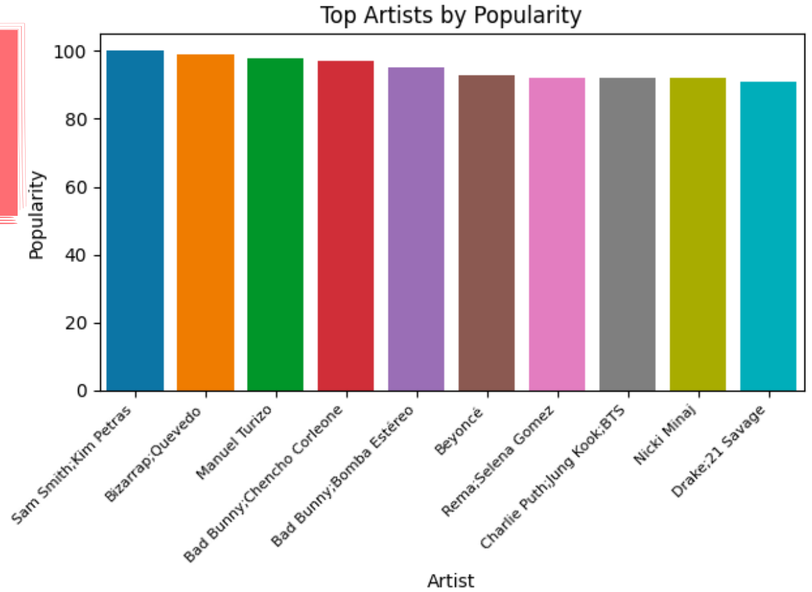
Popularity Distribution



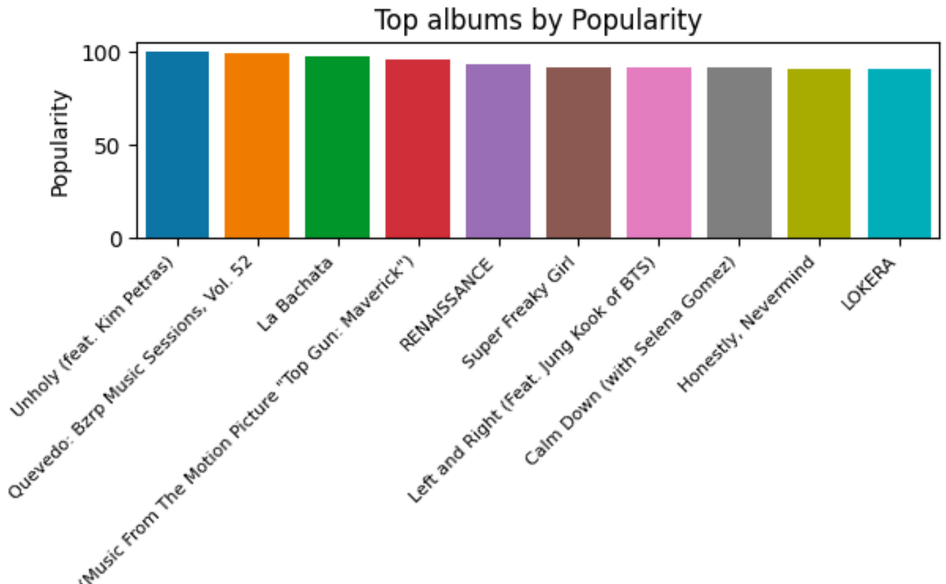
Correlation Matrix



Top Artist by Popularity



Top Album By Popularity



MACHINE LEARNING MODELS

PREDICTION SYSTEM

Set Popularity labels

POPULARITY<50 → 0

POPULARITY>50&<75 → 1

POPULARITY>75&<100 → 2

XG Boost

Extreme Gradient Boosting is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library

Accuracy: 0.766
F-1 Score: 0.682

Disadvantage:
overfitting
computationally expensive

DECISION TREE & RANDOM FOREST

Decision Tree

Accuracy:
0.78(for 0);0.78(for 1)
0.98(for 2);0.77(overall)
Precision:
0.87(for 0); 0.48(for 1);
0.11(for 2);0.78(overall)

F-1 score: 0.77

Random Forest(tuned)

Accuracy:
0.82(for 0);0.82(for 1)
0.99(for 2);0.814(overall)
Precision:
0.81(for 0); 0.81(for 1);
0.33(for 2);0.809(overall)

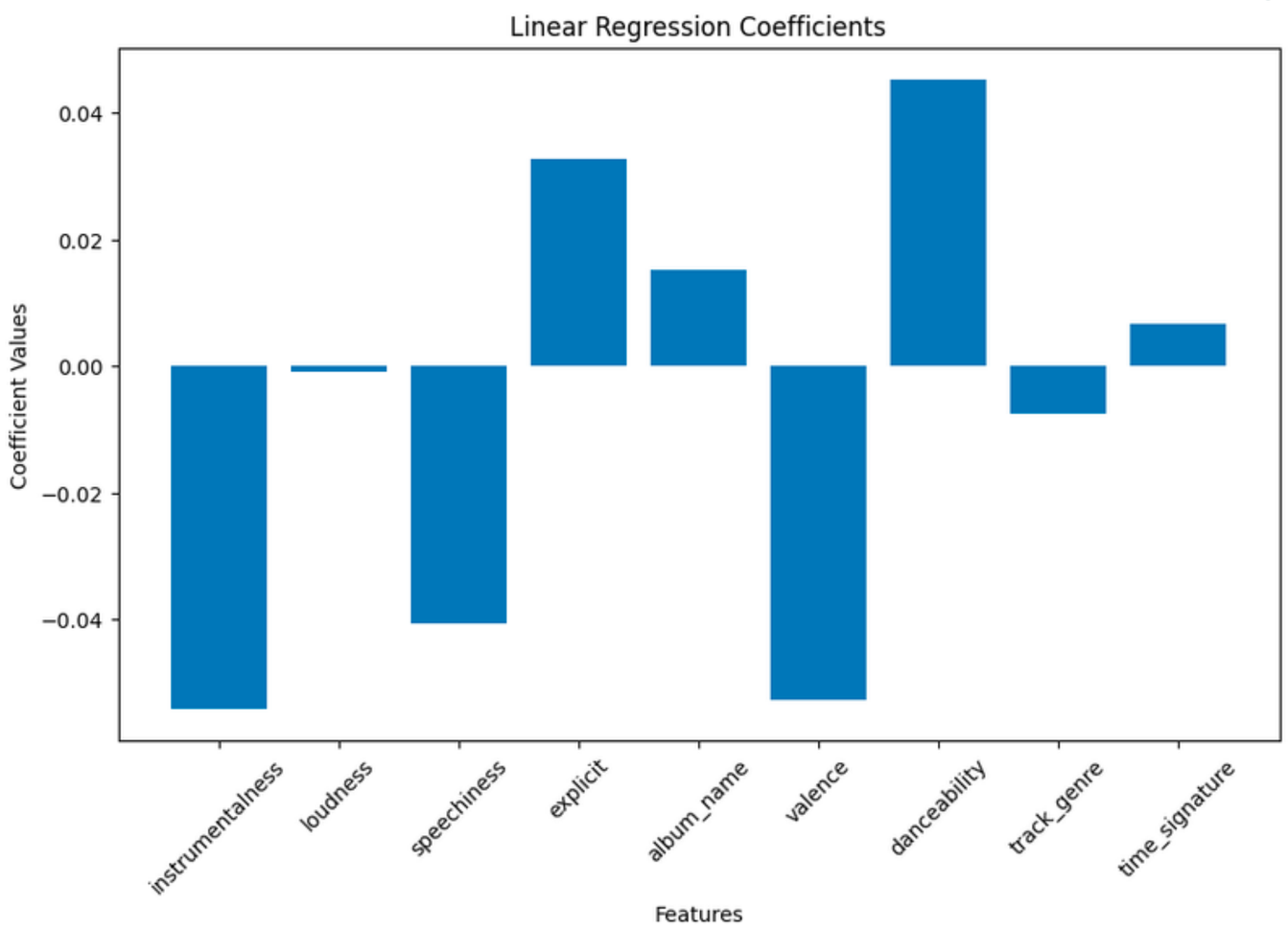
F-1 score: 0.81

Linear Regression

1 : Basic
RMSE: 22.003
R2: 0.0259

2: Lasso
RMSE: 22.0085
R2: 0.0254

3: Remove outliers
RMSE: 22.0469
R2: 0.0279



Logistic linear Regression

Label the popularity

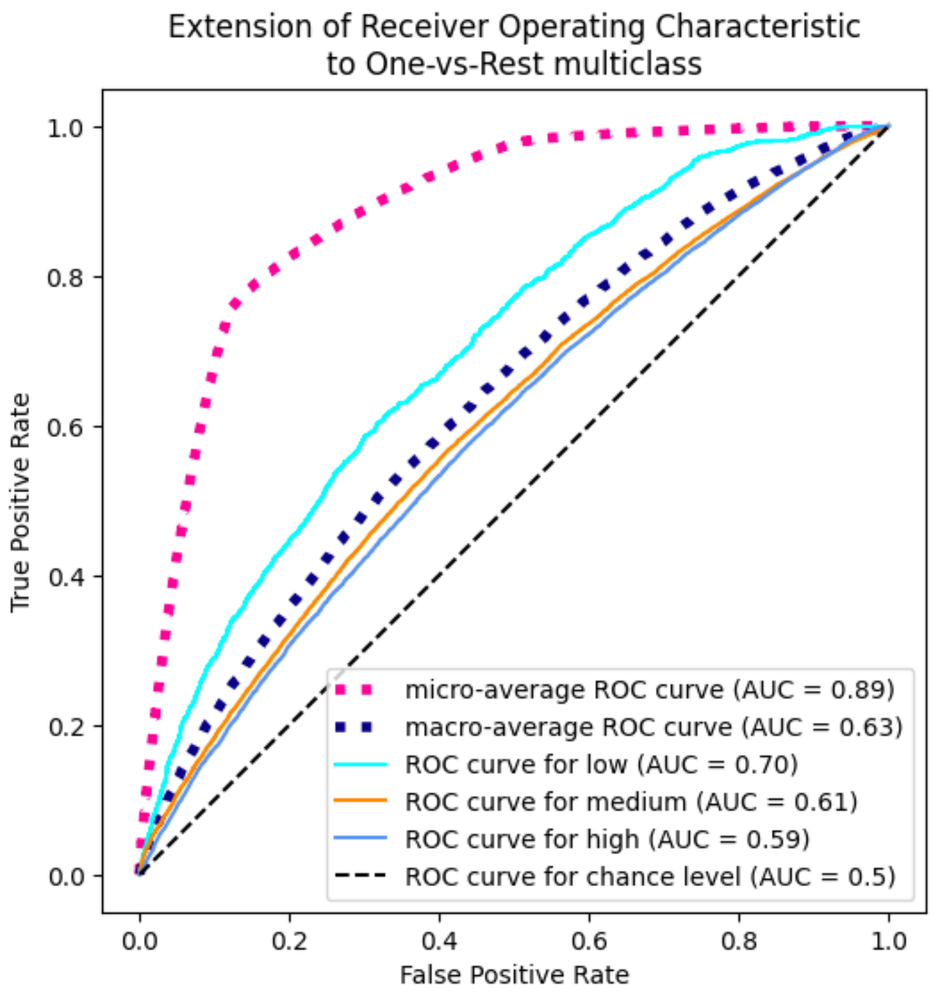
[0,50]: 0 (low)

[50,75]: 1 (medium)

[75,100]: 2 (high)

Accuracy: 0.7563

F1 Score: 0.6515



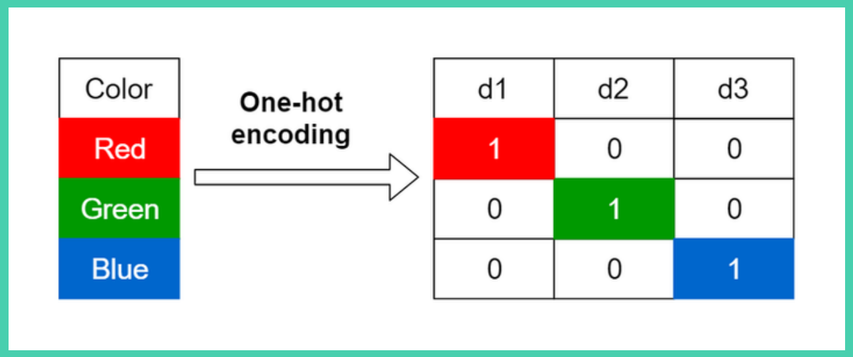
CONTENT-BASED RECOMMENDATION SYSTEM

FEATURE GENERATION

1. track_id, artists, album_name
2. One-hot Encoding
3. Feature Scaling and Normalization

Feature Scaling

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

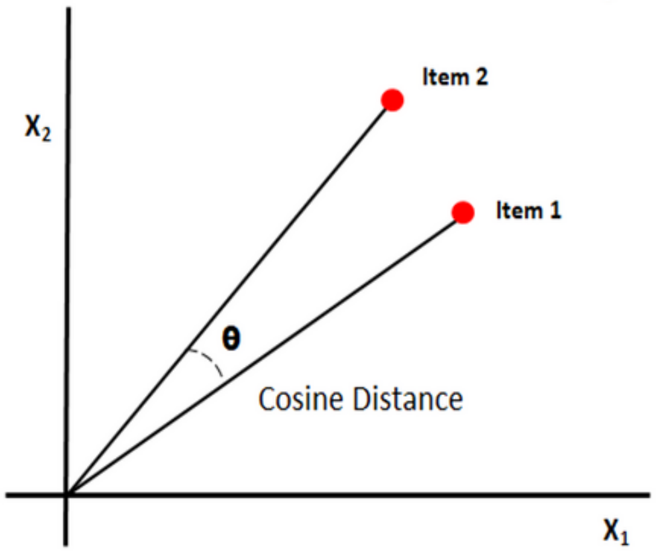


	popularity	duration_ms	explicit	danceability	energy	key	loudness	mode	speechiness	acousticness	...	genre_encode|spanish	genre_encode|study	genre_encode|swedish	genre_encode|synth-pop	genre_encode|tango	genre_encode|techno	genre_encode|trance
0	0.247312	0.053714	0.0	0.557594	0.699	0.818182	0.809185	1.0	0.037226	0.021386	...	0	0	0	0	0	0	0
1	0.333333	0.055569	0.0	0.587156	0.709	0.181818	0.786535	1.0	0.034098	0.005161	...	0	0	0	0	0	0	0
2	0.559140	0.023110	1.0	0.766565	0.485	0.454545	0.682418	1.0	0.045847	0.256024	...	0	0	0	0	0	0	0
3	0.580645	0.046717	0.0	0.556575	0.411	0.181818	0.711883	1.0	0.049531	0.703815	...	0	0	0	0	0	0	0
4	0.268817	0.080803	0.0	0.544343	0.949	0.363636	0.823466	1.0	0.069343	0.001024	...	0	0	0	0	0	0	0

5 rows × 128 columns

BUILDING RECOMMENDER SYSTEM USING COSINE SIMILARITY

Cosine Distance/Similarity

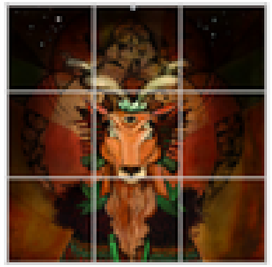


$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

	artist	name	id
0	Speedometer	You've Made Me So Very Happy featuring Ria Currie	5gq3egwoV2pToWU5goXEvi
1	Newen Afrobeat	Qué Sabemos	16zRyplwDUve1JKipYBdEt
2	Marty Robbins	The Red Hills of Utah	7xbKrdp6kwLXzANE580G3b
3	Photon Kid	Voraz	6Nra580NaTHAIgEPBt7ryj
4	The Backseat Lovers	Maple Syrup	4MXE6VCvTsQitHWrAxj7Kg
5	Lack Of Afro	You Could Do Better	1k7UlCzJwIo9Nw3019cTBK
6	Thee Commons	Juaneco Y La Negra	0QaFDdirIxKcsVk62hbKDo
7	Toosii	Favorite Song	1SRw5p2IVAi7RGIHEmZg66
8	Sarah Téibo	Like a Child - Remix	0EKNvyq6JQXvaslzqnjZQf
9	Criolo	Língua Felina – Deluxe Edition	49ieDVc3fyaSV6QUiusAWG

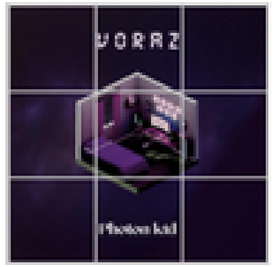
CONNECT TO SPOTIFY API AND VISUALIZE RECOMMENDATION PLAYLIST

e Me So Very Happy featuring Ria Currie

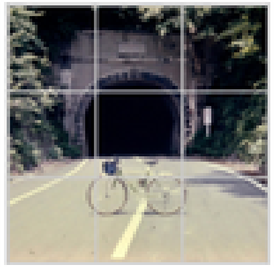


Qué Sabemos

The Red Hills of Utah



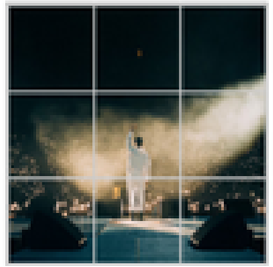
Voraz



Maple Syrup

You Could Do Better

Juaneco Y La Negra



Favorite Song

Like a Child - Remix

Língua Felina – Deluxe Edit

Thank you for listening