Кластеризация и визуализация каталога Spotify (200 000 треков) с целью получения алгоритма для построения плейлистов

Паремузов Матвей, 332 группа

Цели

1. Разведочный анализ данных
2. Построить непрерывную 2-D проекцию треков
3. Сгруппировать треки методом HDBSCAN
4. Интерпретировать и проверить практическую полезность

1. **Локальные соседи.** Для каждой точки x_i ищем k ближайших соседей $\mathcal{N}_k(i)$ по метрике $d(\cdot,\cdot)$ (часто *cosine* для аудио-признаков).

2. **Граф в $\mathbb{R}^D$.** Из расстояний делаем вероятности соседства

$$p_{ij} = \exp\!\left(-\frac{\max(0,\, d_{ij}-\rho_i)}{\sigma_i}\right), \quad j \in \mathcal{N}_k(i).$$

Параметры ρ_i (минимальный радиус) и σ_i подбираются так, чтобы $\sum_j p_{ij} = \log_2 k$ — тем самым у всех точек одинаковая «энтропия» окружения.

Симметризация (t-норма Лукаса) даёт

$$w_{ij} = p_{ij} + p_{ji} - p_{ij}p_{ji} \in [0,1].$$

3. **Проекция в $\mathbb{R}^d$ и оптимизация.** В низком измерении задаём

$$q_{ij} = \frac{1}{1 + a\|y_i - y_j\|^{2b}} \quad (a, b \text{ — константы из } \texttt{min\_dist}),$$

и минимизируем бинарную кросс-энтропию

$$\mathcal{L} = -\sum_{i<j}\Big[w_{ij}\,\log q_{ij} + (1-w_{ij})\log(1-q_{ij})\Big].$$

Оптимизация — стохастический градиентный спуск с отрицательной выборкой (negative sampling).

2. Ключевые гиперпараметры

`n_neighbors`	k — баланс локальная ↔ глобальная структура.
`min_dist`	минимальная дистанция в проекции: ↓ → кластеры плотнее.
`metric`	метрика в $\mathbb{R}^D$ (euclidean, cosine, …).

3. Почему UMAP подходит для музыкальных данных

- Сохраняет и форму локальных кластеров (жанры/муды), и относительное расположение «островов» — t-SNE этим жертвует.
- Линейная сложность по памяти при $k \ll N$ → обрабатывает *200k треков* на ноутбуке.
- Параметр `min_dist` позволяет «сжимать» плотные зоны, подчёркивая жанровые и эмоциональные кластеры.

Silhouette Score: что измеряет и как считать

1. Формула для одной точки x_i

$$a(i) = \frac{1}{|C_i| - 1} \sum_{x_j \in C_i,\, j \neq i} d(x_i, x_j) \quad \text{(средняя дистанция до своего кластера)}$$

$$b(i) = \min_{C \neq C_i} \frac{1}{|C|} \sum_{x_j \in C} d(x_i, x_j) \quad \text{(лучшая чужая «соседняя» кластер-дистанция)}$$

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}} \in [-1,\, 1]$$

- $s(i) \approx 1$ — точка далеко от чужих, близко к своим (отлично).
- $s(i) \approx 0$ — граница между кластерами.
- $s(i) < 0$ — точка ближе к чужому кластеру, чем к своему (плохое назначение).

2. Глобальный показатель

$$\text{Silhouette}(X, \mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

Числится по всем N точкам или по не-noise точкам, если алгоритм (HDBSCAN) выбрасывает шум.

2. Глобальный показатель

$$\text{Silhouette}(X, \mathcal{C}) = \frac{1}{N} \sum_{i=1}^{N} s(i)$$

3. Грубая шкала интерпретации

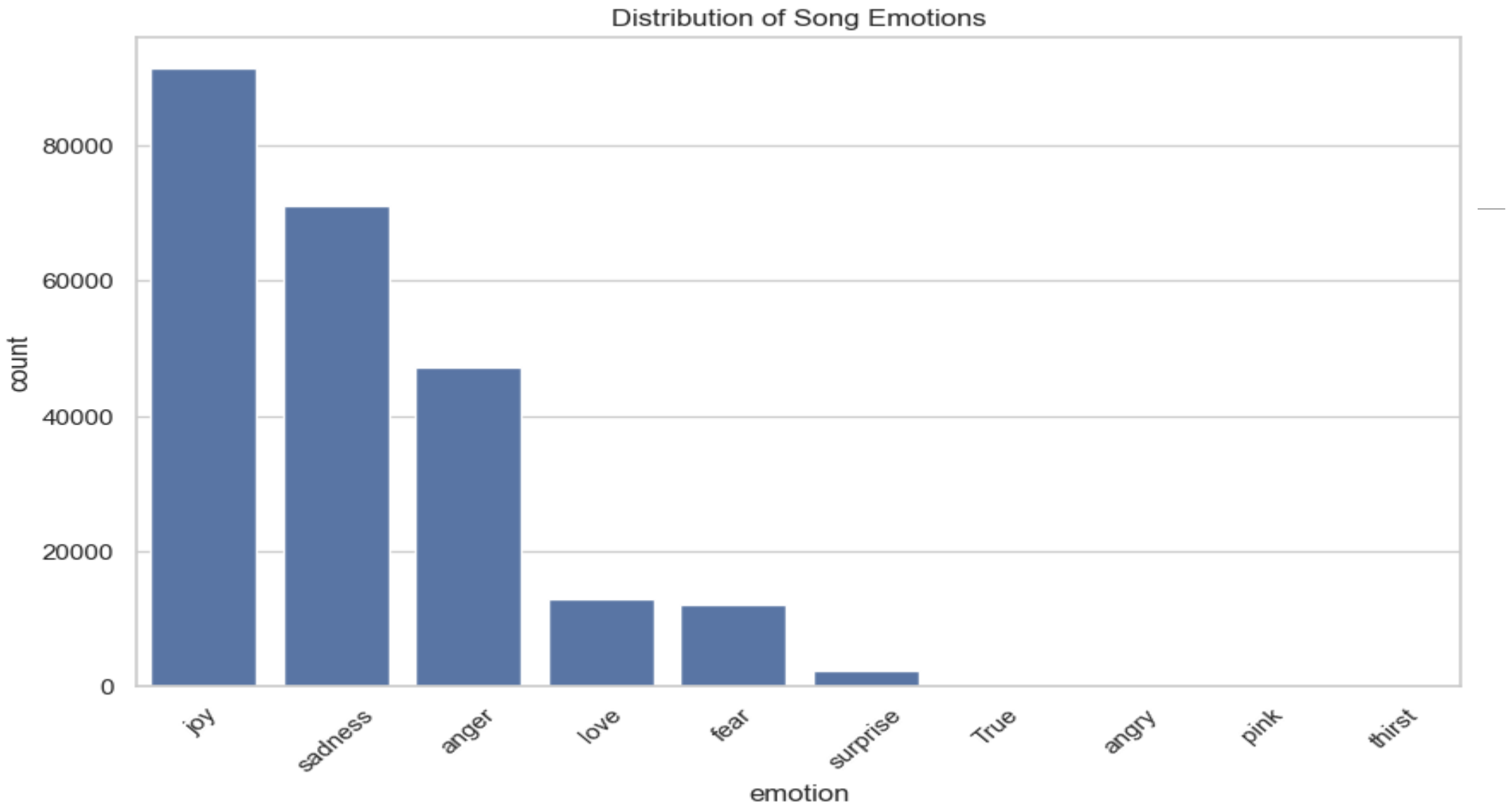
$s < 0$	кластеры перекрываются — пере-или недокластизация.
$0 \ldots 0.20$	слабое разделение, «пятна» без чётких границ.
$0.20 \ldots 0.50$	умеренное качество (норма для реальных данных).
> 0.50	высокое качество, редко встречается вне синтетических данных.

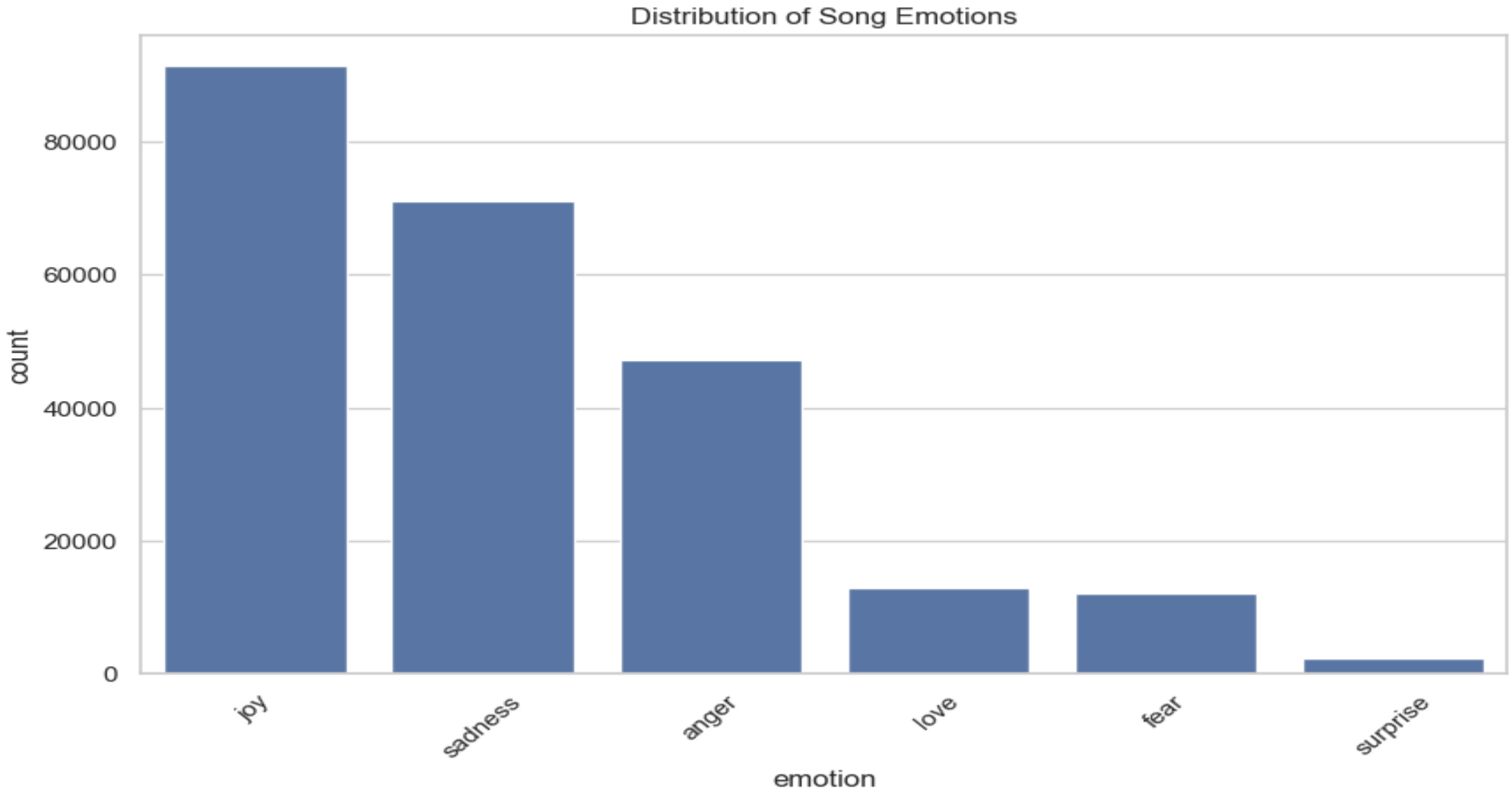
Анализ данных:

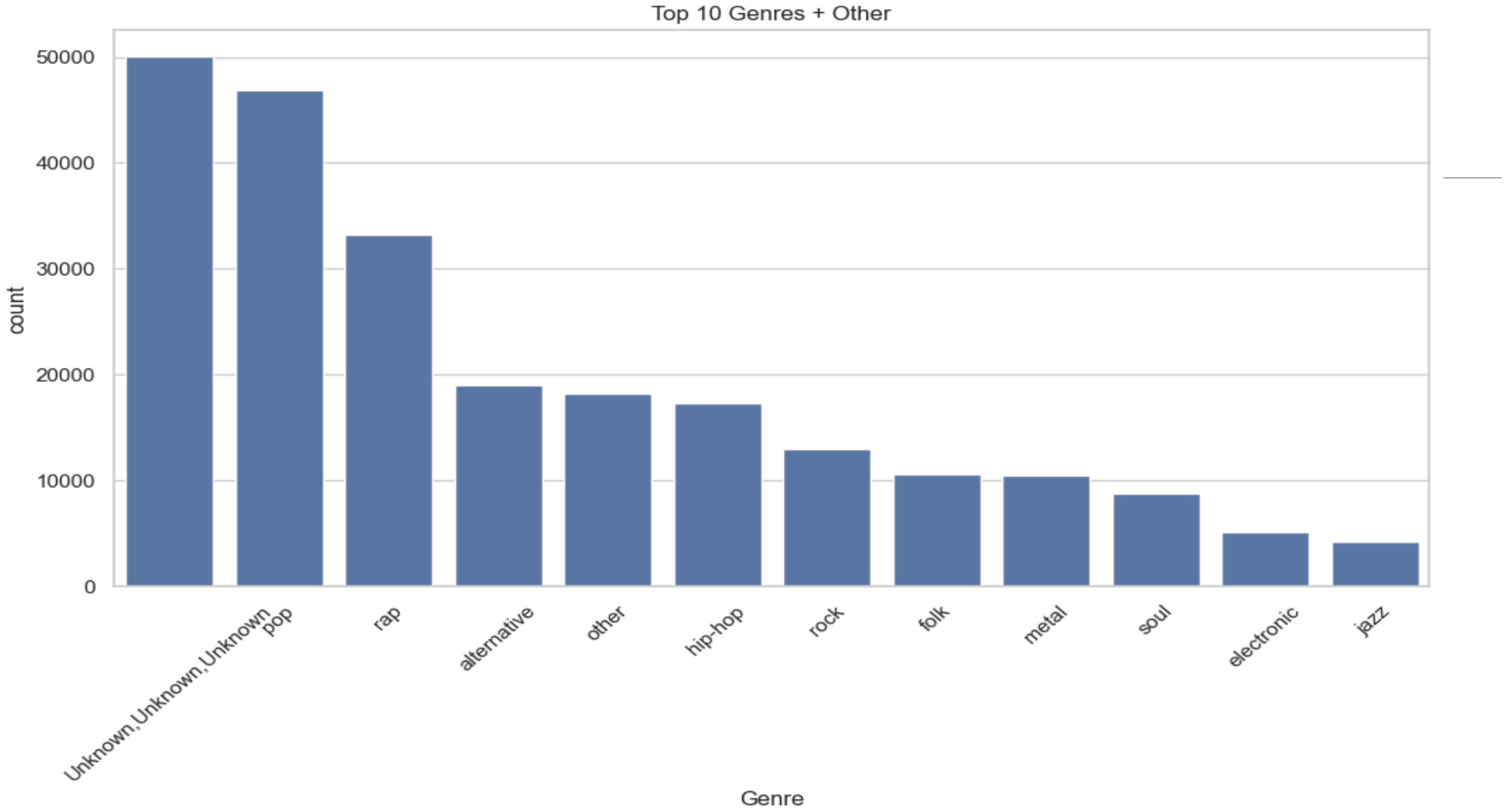
Признаки по трекам:
- Исполнитель
- Название трека
- Мера изменчивости звуковых характеристик песни
- Жанр
- Год выпуска
- Тональность
- Темп (в Уд\мин)
- Средняя громкость песни
- Наличие ненормативной лексики

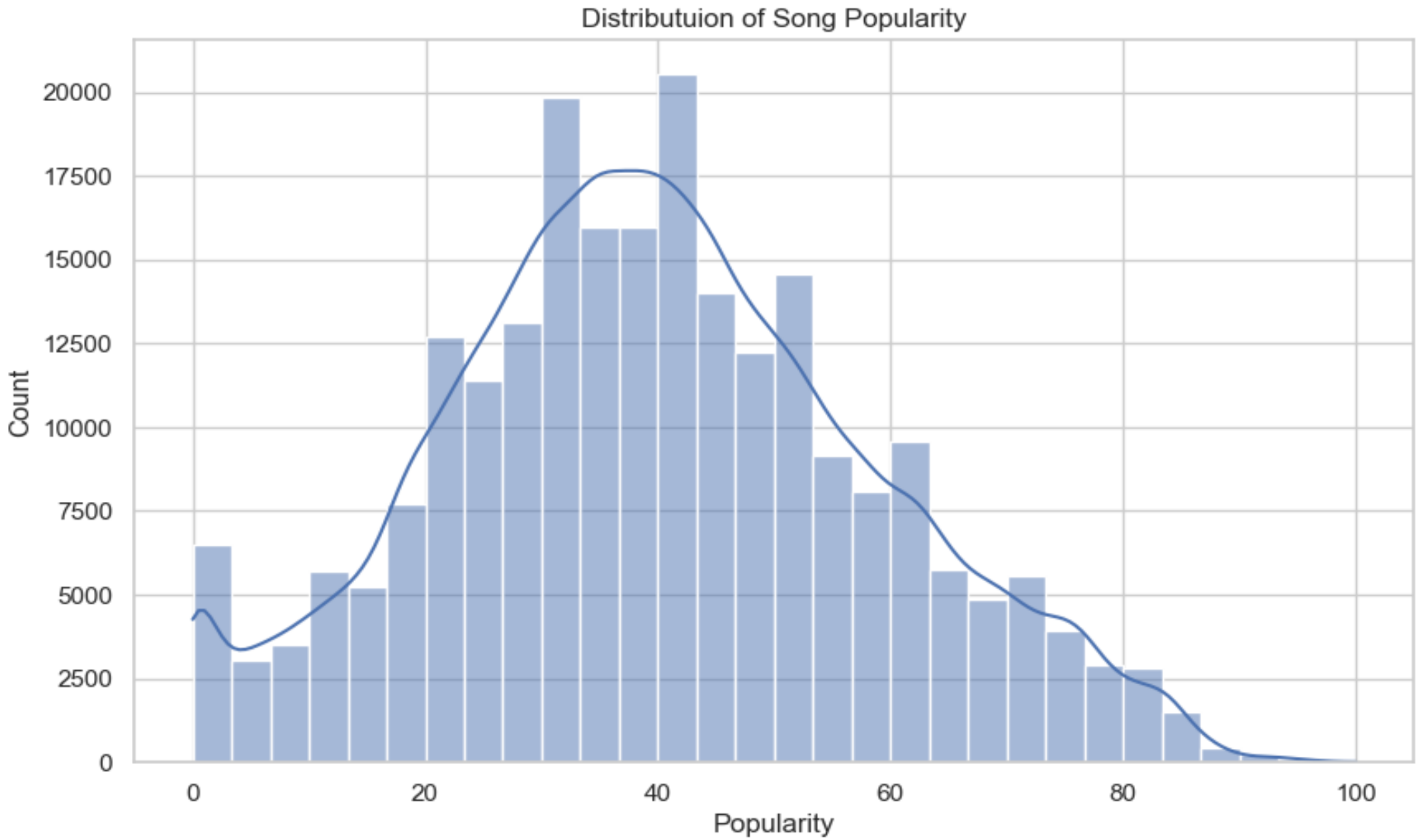
Также в данном датасете присутствуют признаки, полученные с помощью нейросети T5 от компании Google. Вот эти признаки:

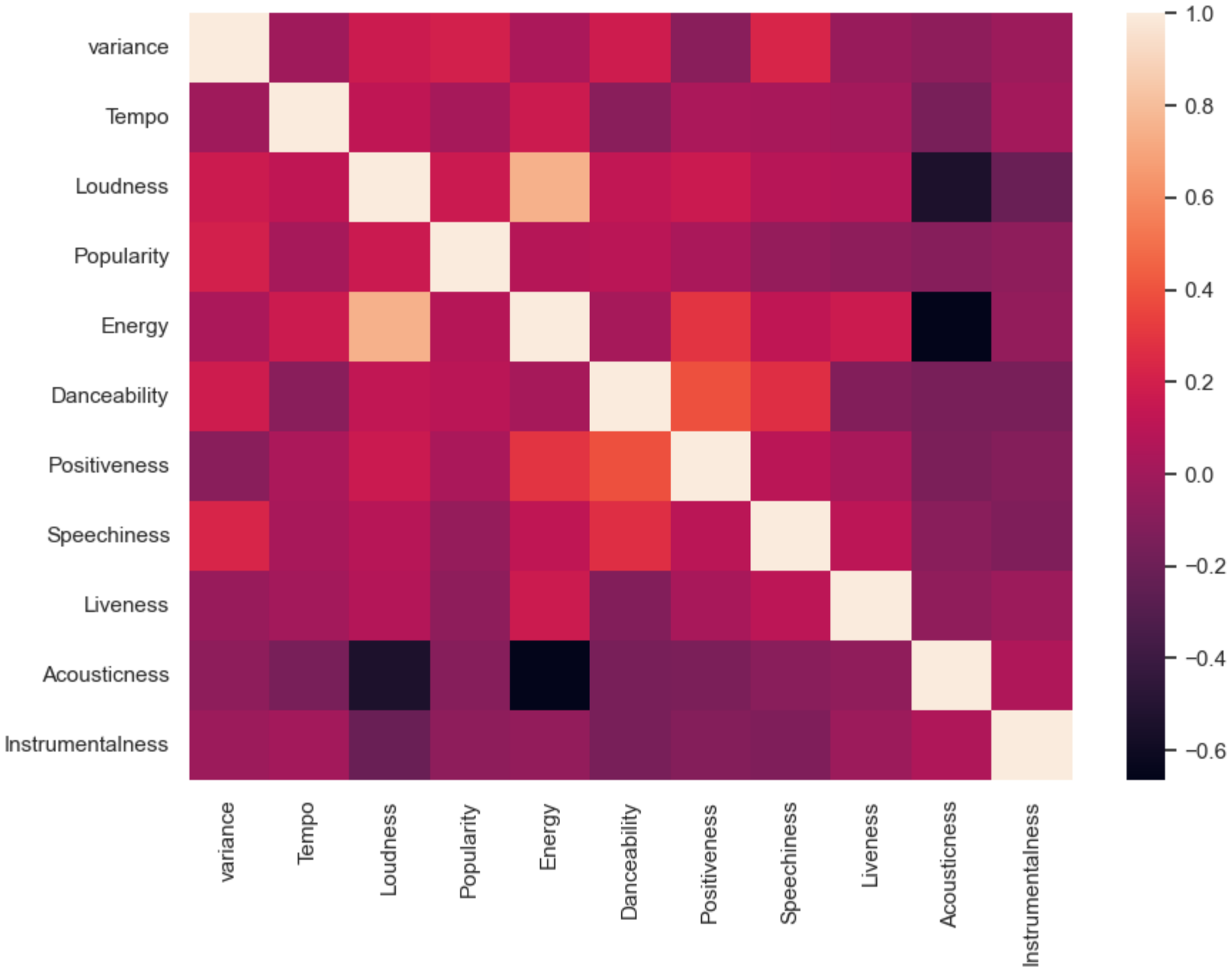
- Основная эмоция песни
- Насколько вероятно трек является танцевальным (0-100)
- Энергичность трека
- Позитивность
- Речетативность
- Вероятность записи трека быть живым выступлением
- Акустическое качество
- Вероятность того, что трек инструментальный

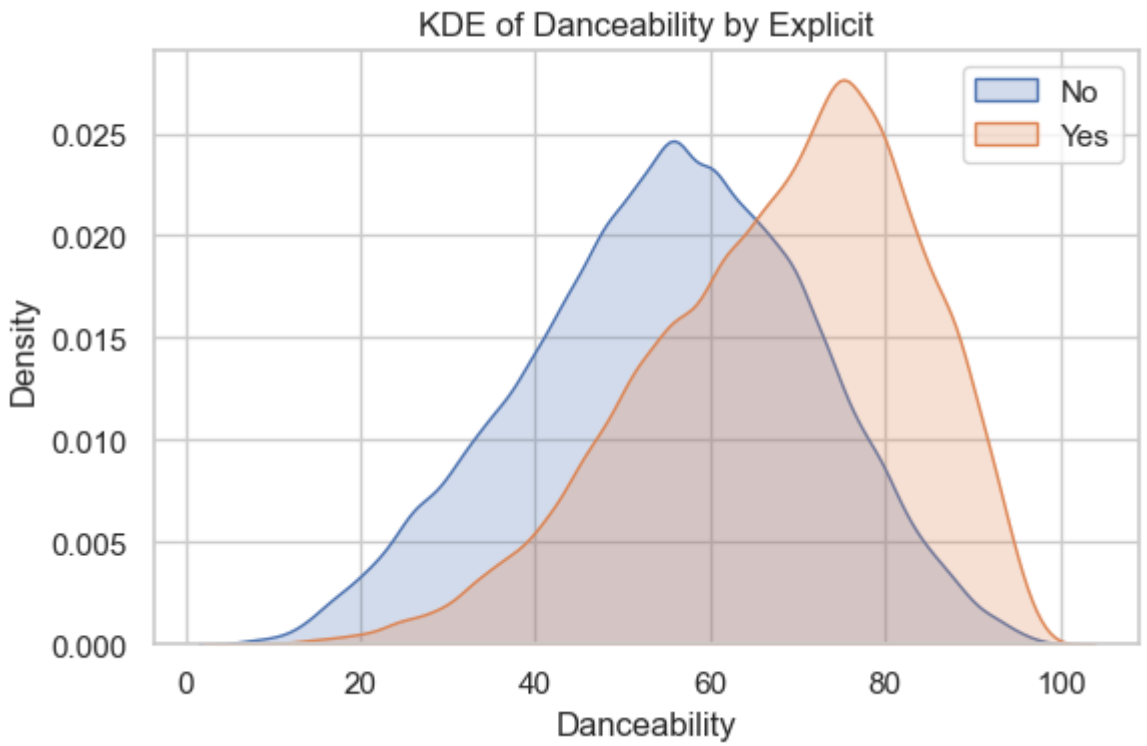

Distribution of Song Emotions
80000
60000
40000
20000
0
count
joy
sadness
anger
love
fear
surprise
True
angry
pink
thirst
emotion

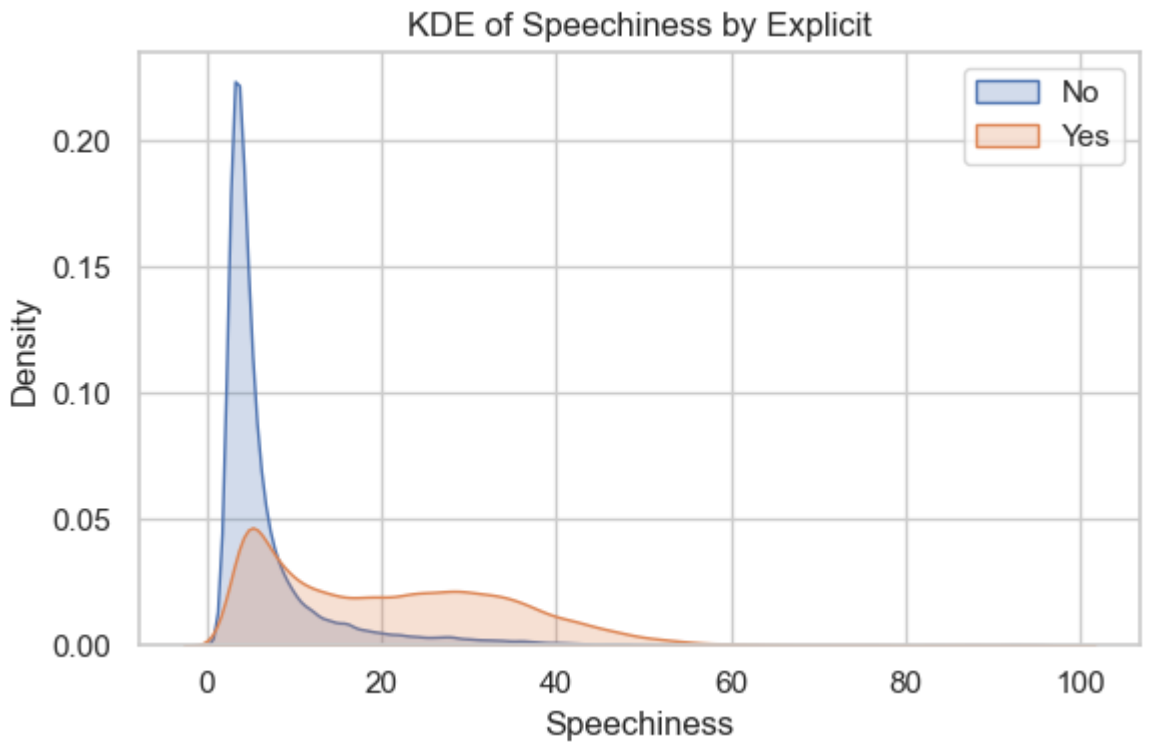

Distribution of Song Emotions
80000
60000
40000
20000
0
count
joy
sadness
anger
love
fear
surprise
emotion

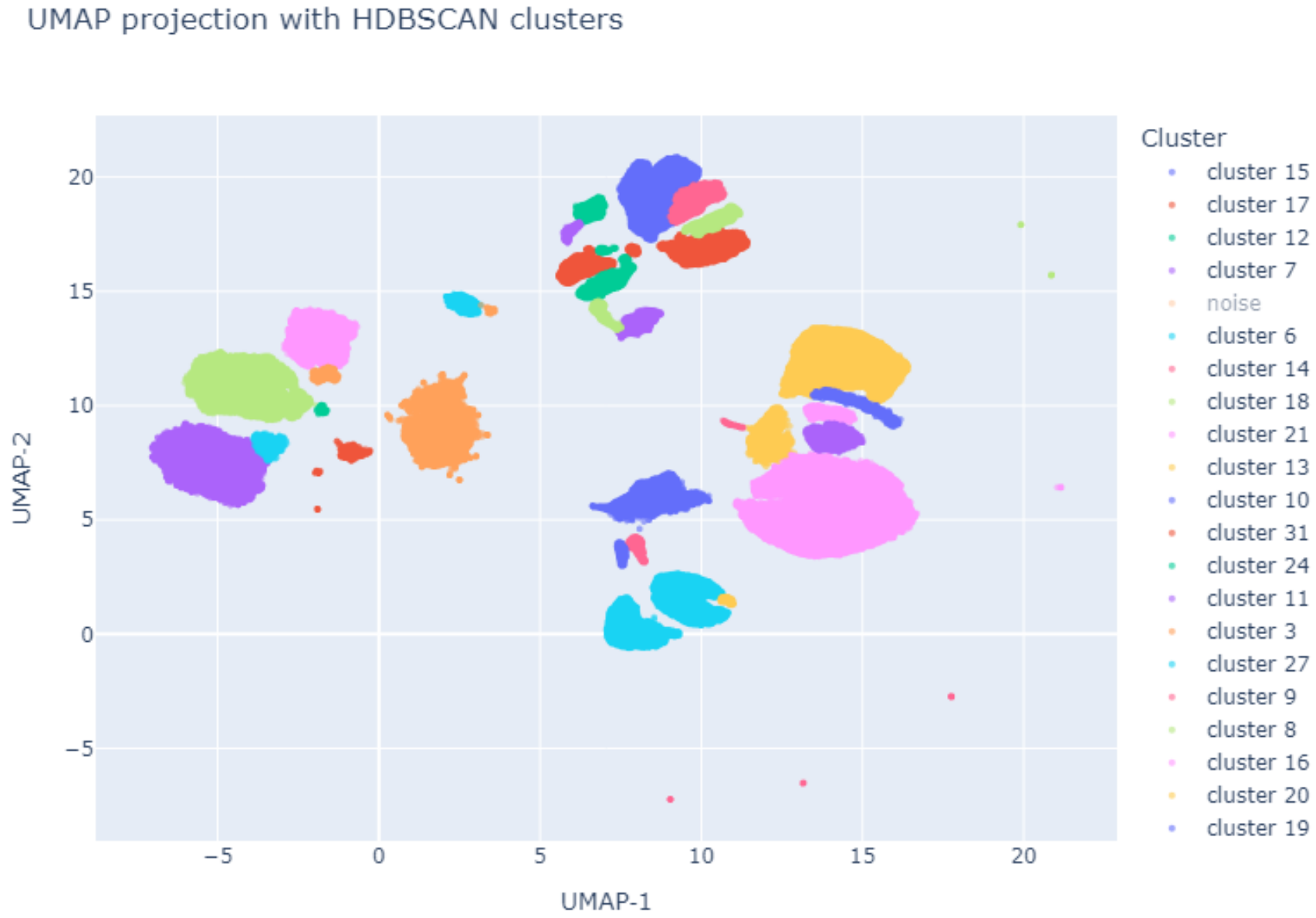

Top 10 Genres + Other
count
50000
40000
30000
20000
10000
0
Unknown,Unknown,Unknown
pop
rap
alternative
other
hip-hop
rock
folk
metal
soul
electronic
jazz
Genre


Distributuion of Song Popularity
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
0
20
40
60
80
100
Popularity


variance
Tempo
Loudness
Popularity
Energy
Danceability
Positiveness
Speechiness
Liveness
Acousticness
Instrumentalness
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6


KDE of Danceability by Explicit
No
Yes
Density
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
100
Danceability


KDE of Speechiness by Explicit
No
Yes
Density
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
Speechiness


UMAP projection with HDBSCAN clusters
UMAP-2
UMAP-1
20
15
10
5
0
−5
−5
0
5
10
15
20
Cluster
cluster 15
cluster 17
cluster 12
cluster 7
noise
cluster 6
cluster 14
cluster 18
cluster 21
cluster 13
cluster 10
cluster 31
cluster 24
cluster 11
cluster 3
cluster 27
cluster 9
cluster 8
cluster 16
cluster 20
cluster 19