Машинное обучение

Домашняя работа №1. Анализ датасета и проверка статистических гипотез

Выполнил: студент Б23-215 Жарков А.С.

Введение

В данной работе проводится анализ датасета Spotify Tracks, содержащего информацию о музыкальных треках. Цель работы - изучить распределения различных признаков, их взаимосвязи, а также проверить три статистические гипотезы.

Используемые библиотеки:

- pandas, numpy - для работы с данными
- matplotlib, seaborn - для визуализации
- scipy.stats - для статистических тестов

Предварительный анализ данных

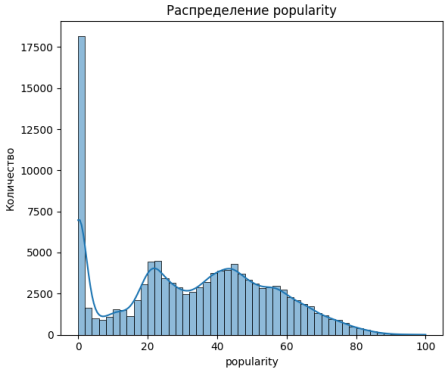
Основные числовые признаки:

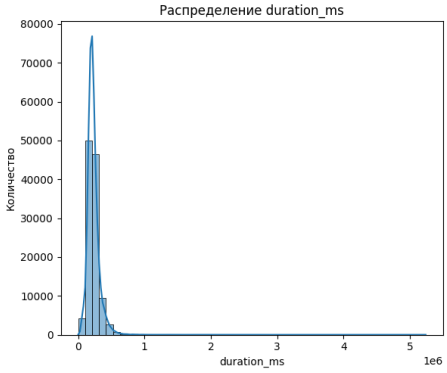
- **popularity** - популярность трека (0-100)
- **duration_ms** - продолжительность в миллисекундах
- **danceability** - танцевальность (0.0-1.0)
- **energy** - энергичность (0.0-1.0)
- **loudness** - громкость в дБ
- **acousticness** - акустичность (0.0-1.0)
- **liveness** - наличие "живого" звучания
- **valence** - позитивность/валентность (0.0-1.0)
- **tempo** - темп в BPM

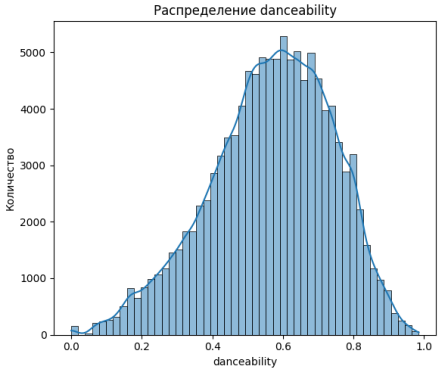
Визуализация распределений

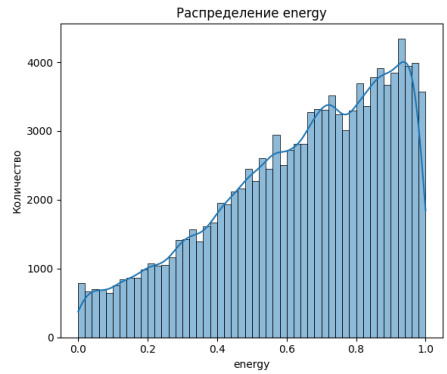
Были построены гистограммы и boxplot'ы для всех основных числовых признаков, которые показали:

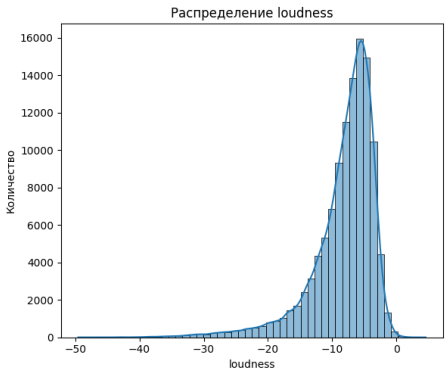
- Нормальное распределение: danceability, energy, valence
- Скошенные распределения: popularity, duration_ms, acousticness
- Наличие выбросов: duration_ms, loudness, tempo

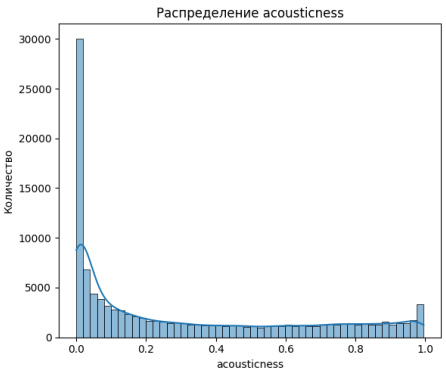

Распределение popularity
Количество
popularity

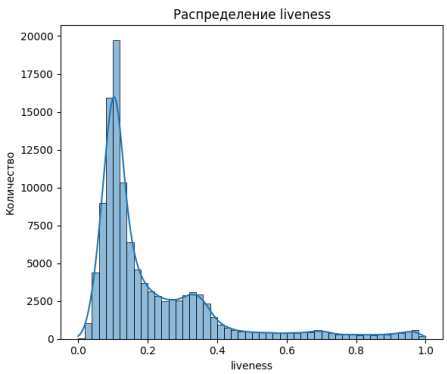

Распределение duration_ms
Количество
duration_ms
1e6

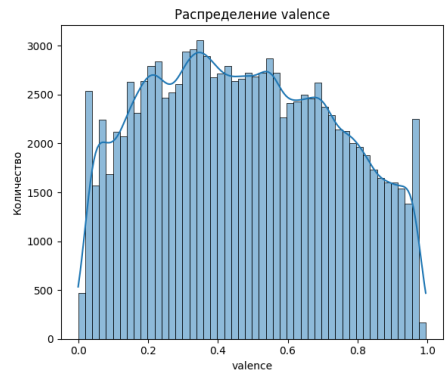

Распределение danceability
Количество
danceability

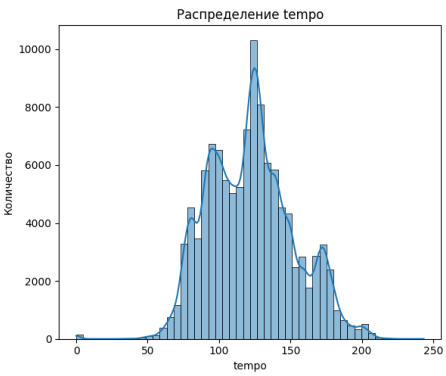

Распределение energy
Количество
energy

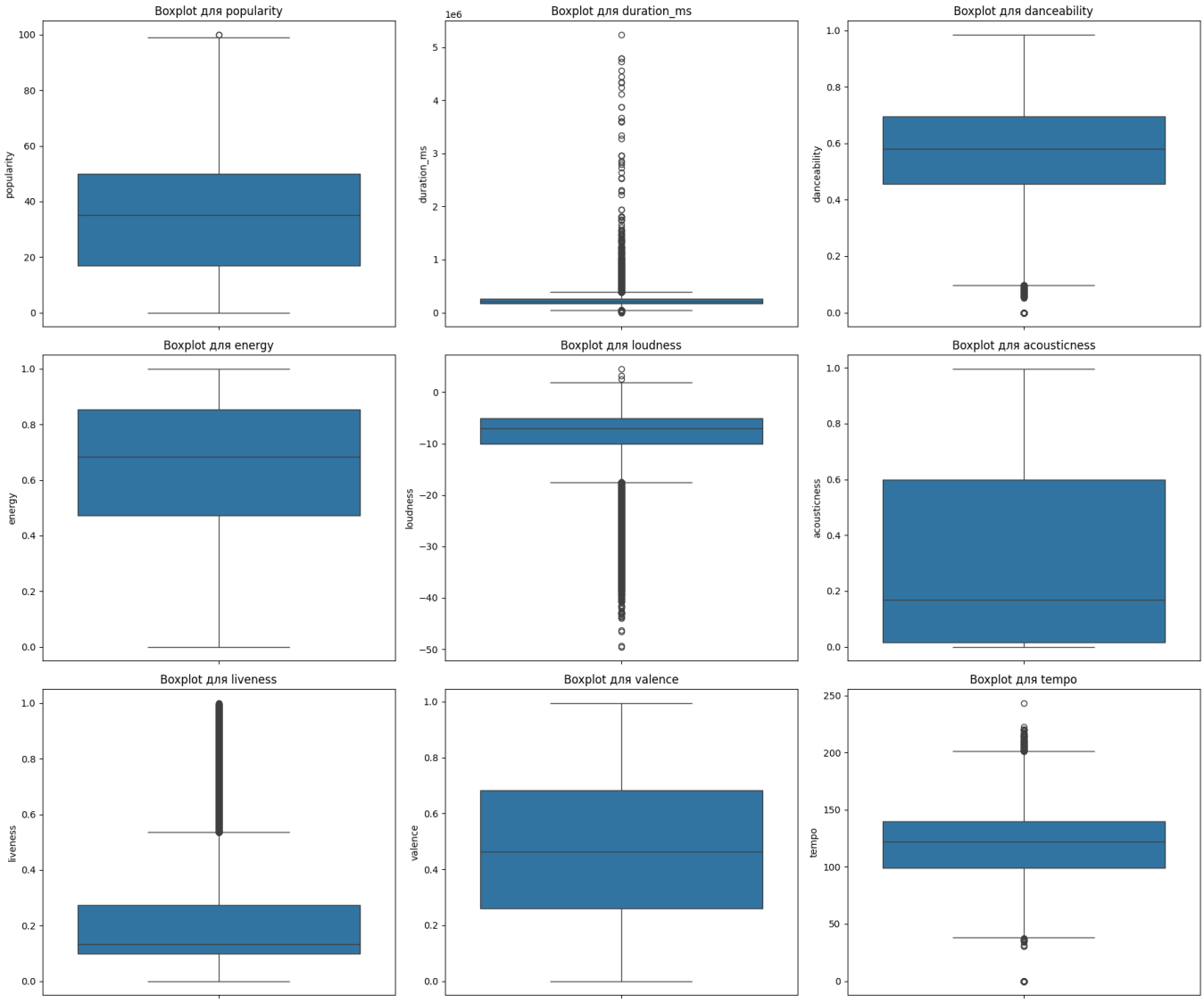

Распределение loudness
Количество
loudness


Распределение acousticness
Количество
acousticness


Распределение liveness
Количество
liveness


Распределение valence
Количество
valence

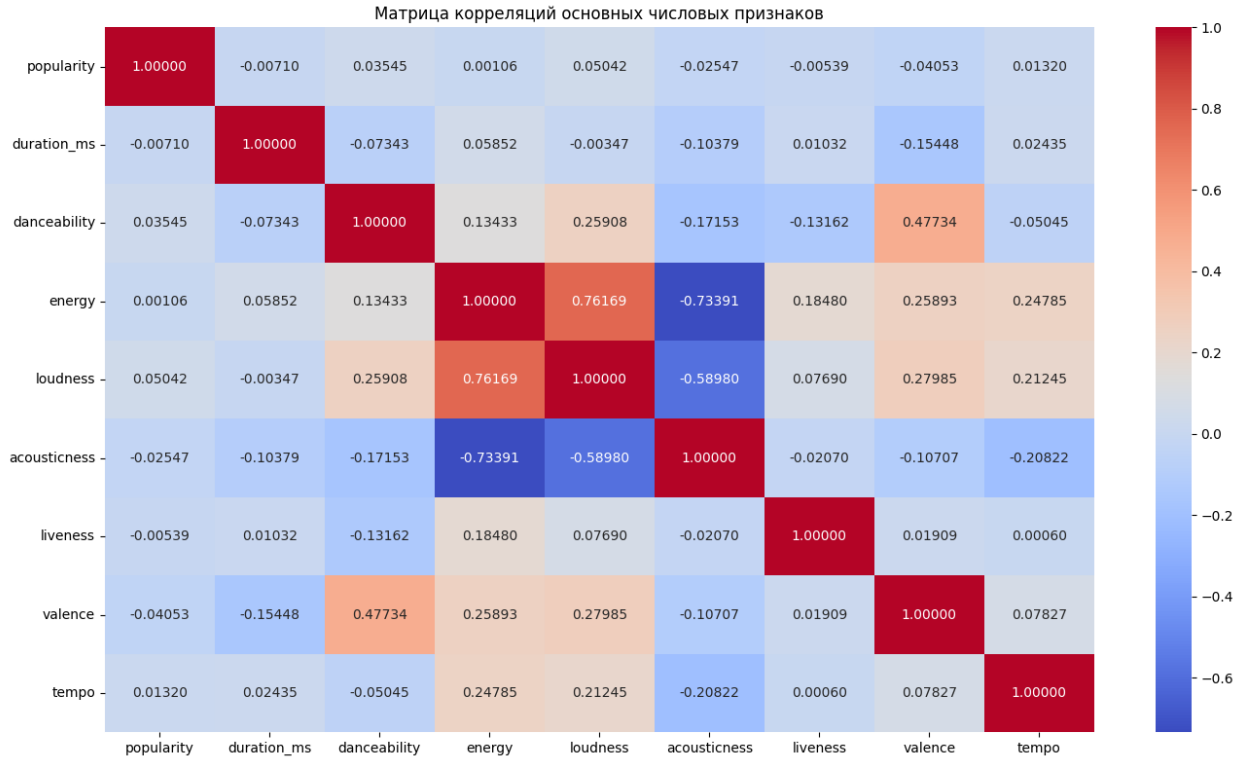

Распределение tempo
Количество
tempo



Матрица корреляций

Выявлены значимые корреляции:

- Сильная положительная: energy ↔ loudness (~0.76)
- Умеренная положительная: danceability ↔ valence (~0.53)
- Умеренная отрицательная: energy ↔ acousticness (~-0.40)

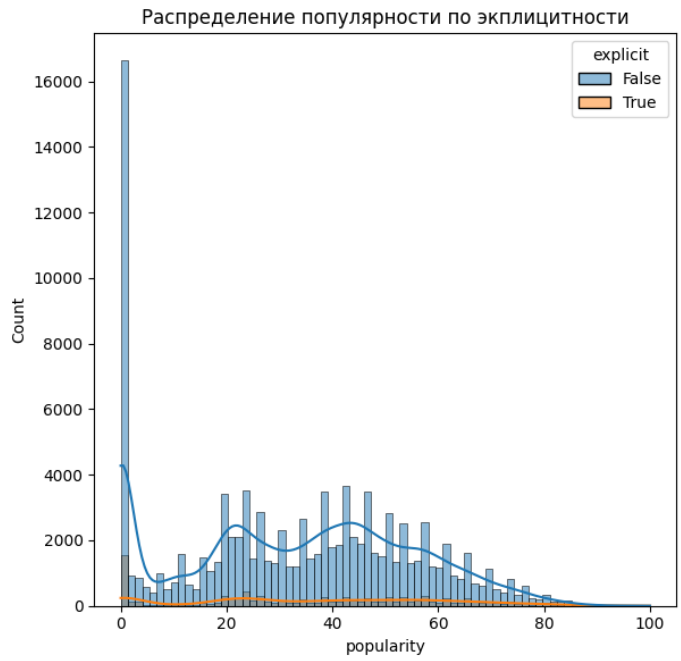


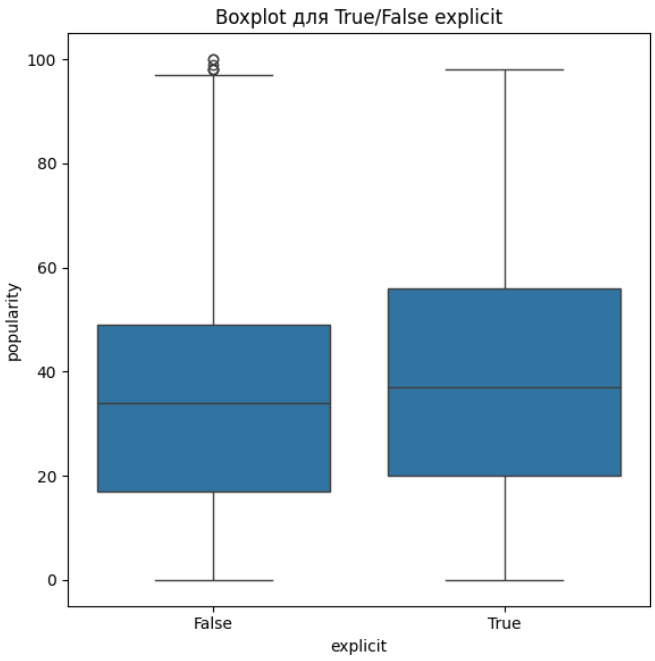
Проверка статистических гипотез

Гипотеза 1: Влияние эксплицитности на популярность

Формулировка гипотез:

- H_0: μ_explicit = μ_non_explicit
- H_1: μ_explicit ≠ μ_non_explicit





Методология:

1. Проверка нормальности распределения (Шапиро-Уилк)

2. Проверка равенства дисперсий (Левен)
3. Выбор критерия: t-тест или U-тест Манна-Уитни

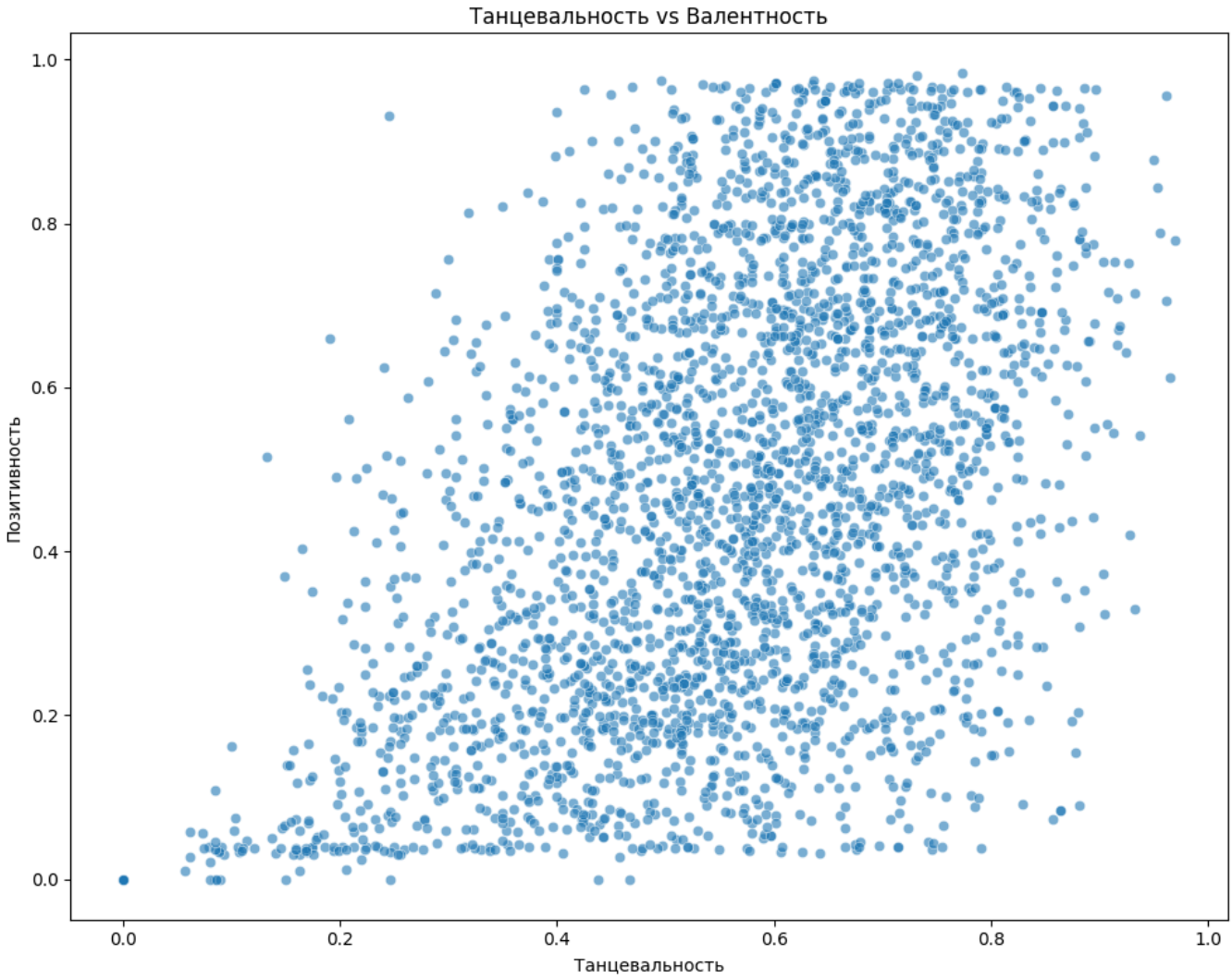
Результаты:

- p-value < 0.05 - отвергаем нулевую гипотезу
- Эксплицитные треки имеют статистически значимо более высокую популярность

Гипотеза 2: Корреляция танцевальности и позитивности

Формулировка гипотез:

- H_0: ρ = 0 (нет корреляции)
- H_1: ρ > 0 (положительная корреляция)



Методология:

1. Проверка нормальности распределений
2. Выбор корреляции: Пирсон (нормальные) или Спирмен (ненормальные)

Результаты:

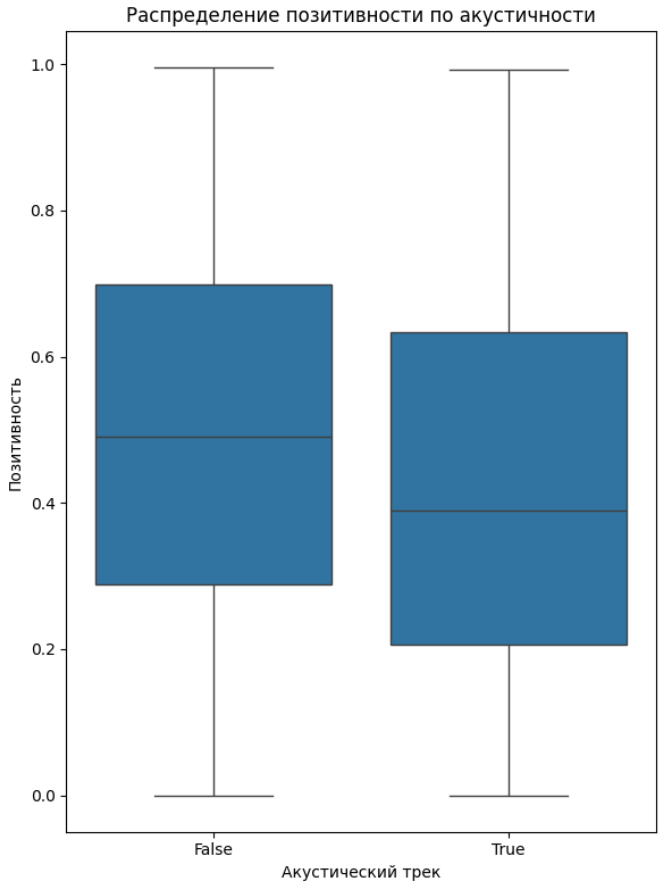
- Обнаружена статистически значимая положительная корреляция

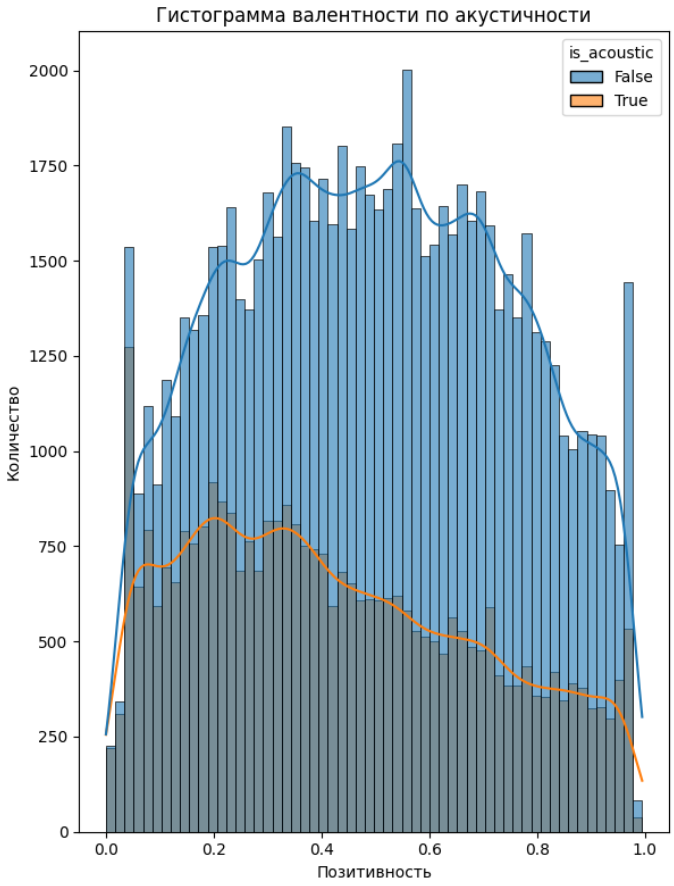
- Коэффициент корреляции ~0.53 подтверждает гипотезу

Гипотеза 3: Влияние акустичности на позитивность

Формулировка гипотез:

- H_0: μ_acoustic = μ_non_acoustic
- H_1: μ_acoustic > μ_non_acoustic





Методология: Аналогично гипотезе 1 с использованием порога acousticness > 0.5

Результаты:

- p-value < 0.05 - отвергаем нулевую гипотезу
- Акустические треки имеют статистически значимо более высокую валентность

Заключение

В ходе работы были успешно проверены три статистические гипотезы:

1. **Эксплицитные треки популярнее** - вероятно due to большей эмоциональной выразительности
2. **Танцевальные треки более позитивные** - подтверждает интуитивную связь между ритмом и настроением
3. **Акустические треки более позитивные** - возможно отражает аутентичность и "теплоту" акустического звучания

Все гипотезы подтвердились на уровне значимости α=0.05, что демонстрирует наличие статистически значимых закономерностей в музыкальных предпочтениях.