**Московский государственный технический   
университет им. Н. Э. Баумана**

Курс «Технологии машинного обучения»  
Отчёт по лабораторной работе №2  
«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»

Выполнил: Проверил:  
Зелинский Д.М., Нардид А.Н.,  
группа ИУ5-61Б каф. ИУ5

Дата: Дата:

Подпись: Подпись:

2023  
Москва

**Цель работы**

Изучение способов предварительной обработки данных для дальнейшего формирования моделей.

**Описание датасета**

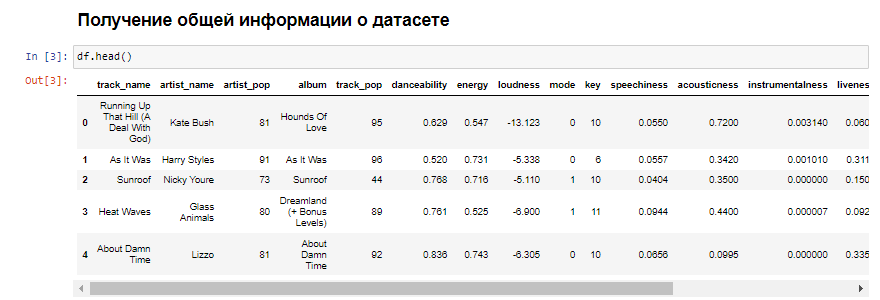
В качестве набора данных мы используем "Популярные песни ТикТока 2022г." https://www.kaggle.com/datasets/sveta151/tiktok-popular-songs-2022

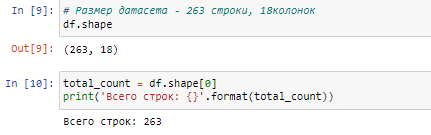
Этот набор данных предоставляет всю важную информацию, которая может потребоваться для дальнейшего анализа, начиная с базовых знаний, таких как название трека и имя исполнителя, заканчивая самыми продвинутыми сведениями, такими как темп, time\_signature и т.д. и т.п.

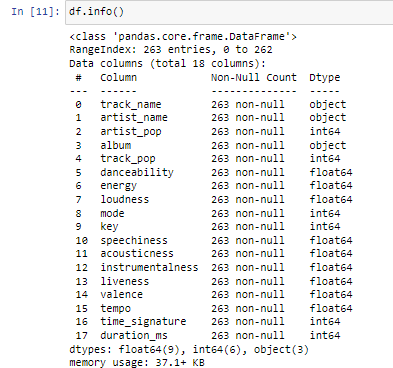
Проанализировав статистику по тикток-песням, можно получить данные об особенностях медиакультуры молодёжи.

**Выполнение работы**

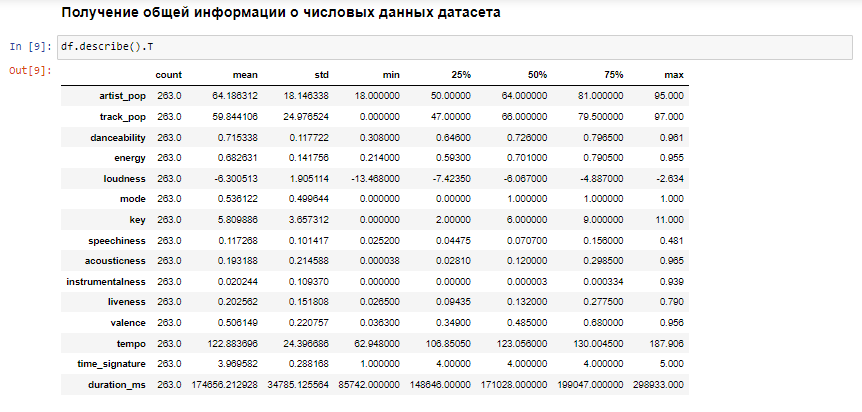


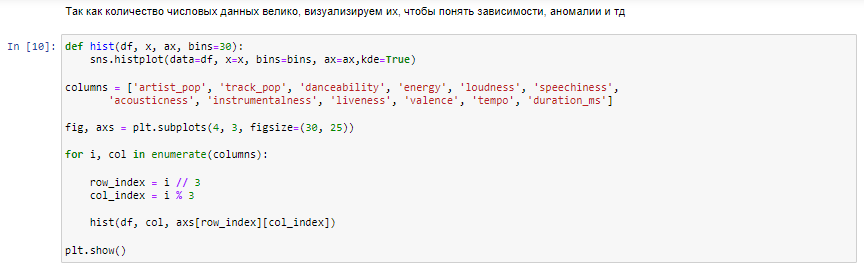


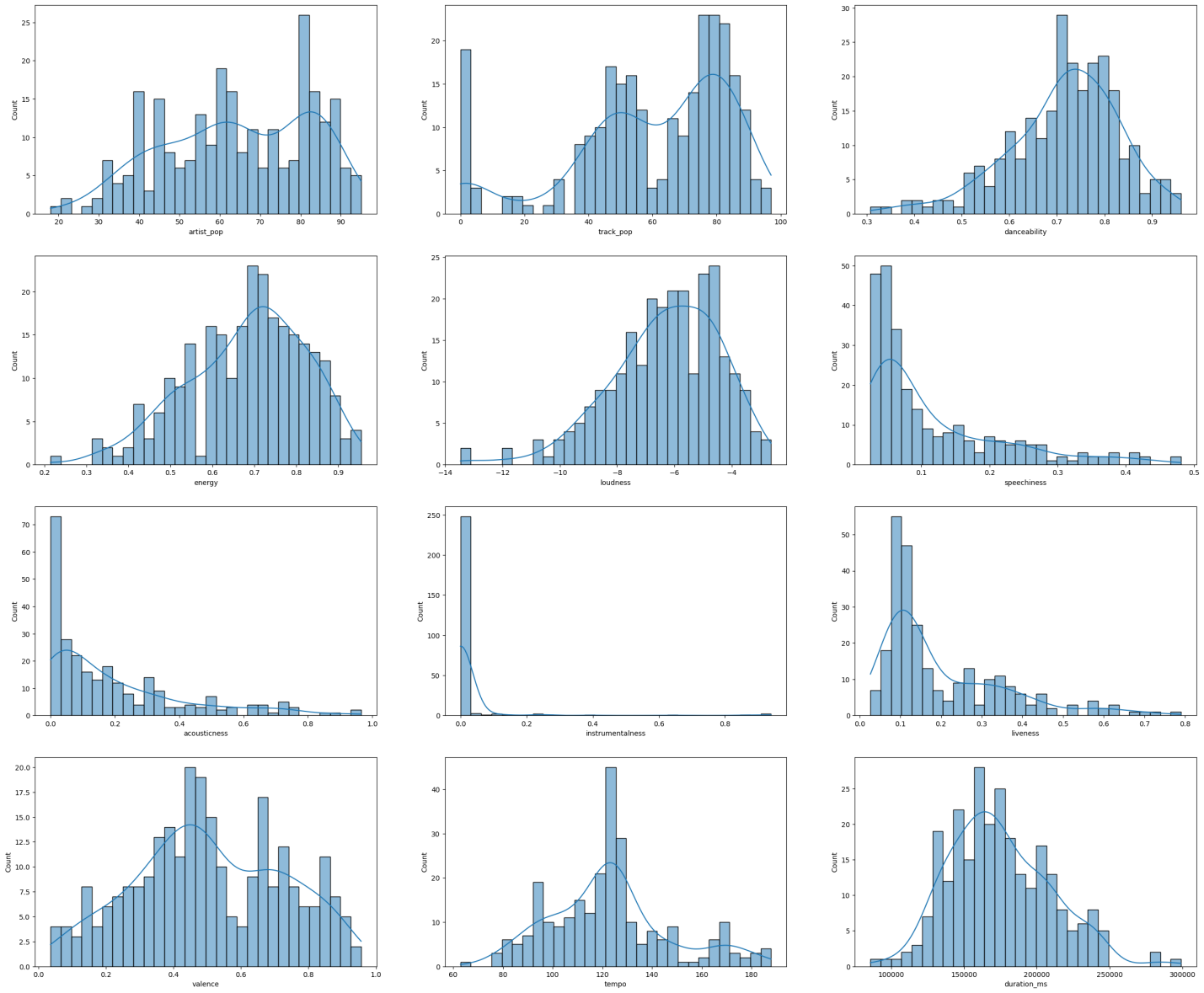








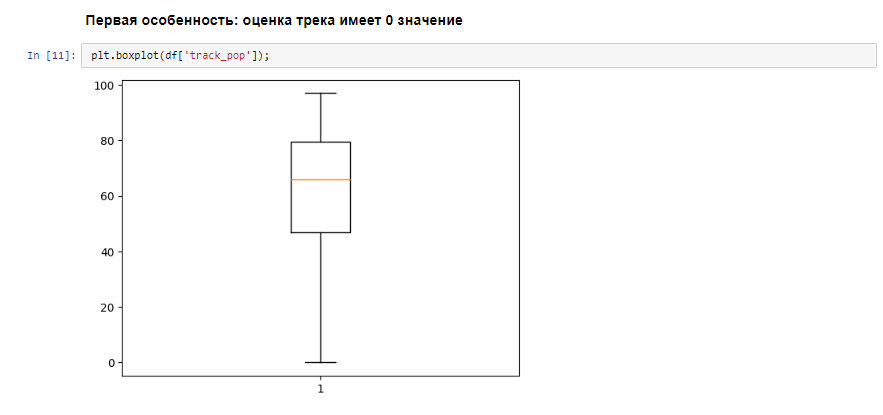




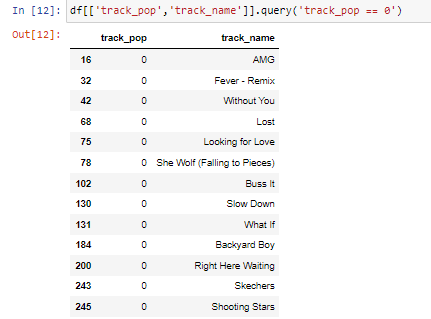
В переменной 'track\_pop', есть скачок на 0. Вероятно, это означает, что для таких треков у нас отсутствуют данные, которые нам придется либо исключить, либо заменить средним значением по столбцу.

"Громкость" - единственная функция с отрицательными значениями

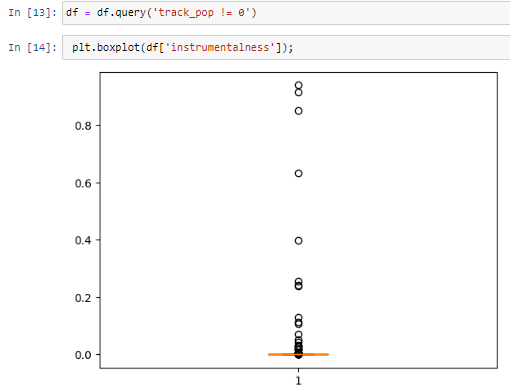
"Инструментальность" имеет почти все значения в 0: из-за этого данный столбец не несет особой смысловой нагрузки, исключим его в дальшейем.

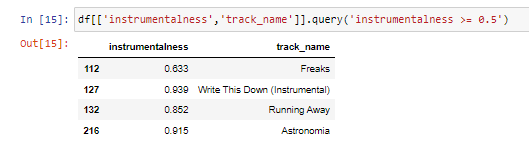


Видим, что Q1 и Q3 находятся в разрезе между 45 и 75 баллами, поэтому будем считать значения равные 0 - выбросами, посмотрим, какой их процент в датафрейме.

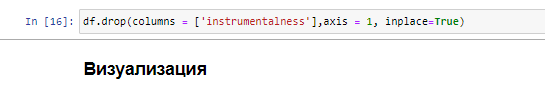


13 треков имеет рейтинг равный 0 - это не большой процент. Эти данные можно удалить.

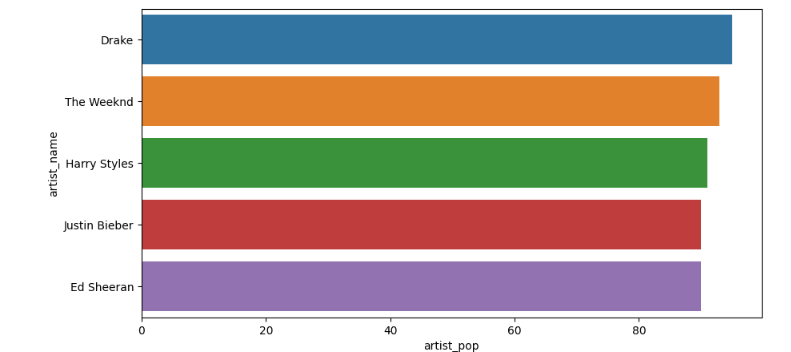




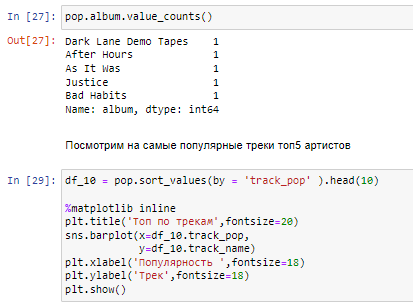
Да, действительно, видим, что 90% нашего датафрейма - вокальные треки, лишь 4 их них имеют порог выше 0.5, что говорит об инструментальности данных треков. Поэтому удалим данный столбец, тк выдвинуть гипотезы и анализировать данные с помощью него не получится.

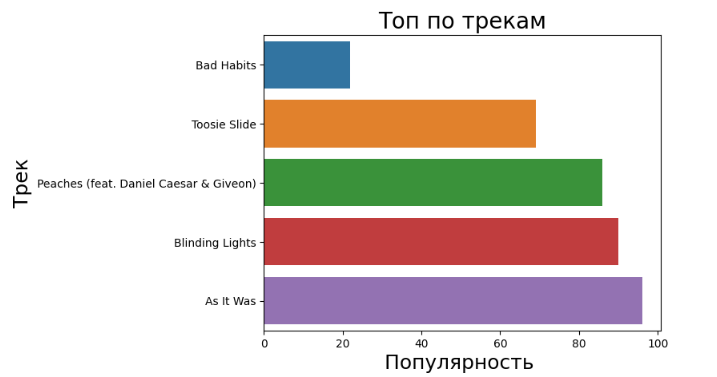




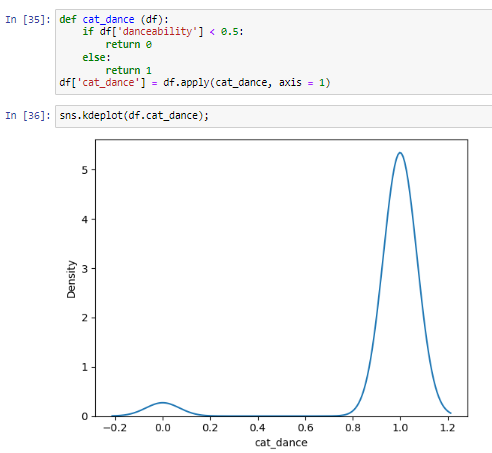


Примерно у всех артистов одинаковый уровень оценки > 80.





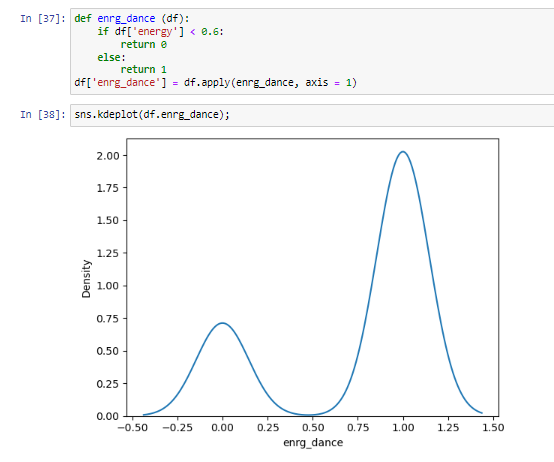
Проведём статистический анализ наших данных. Сначала рассмотрим статистику зажигательности треков.



Статистика зажигательности треков

Большинство треков – танцевальные.

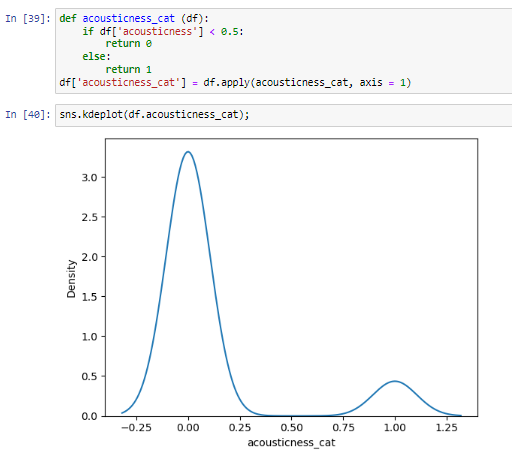
Теперь рассмотрим статистику энергичности треков



Статистика энергичности треков

Большинство треков – энергичные.

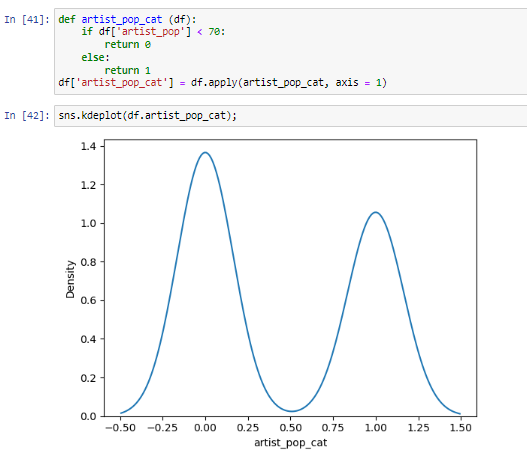
Теперь рассмотрим статистику акустики треков



Статистика акустики треков

Большинство треков – электронные.

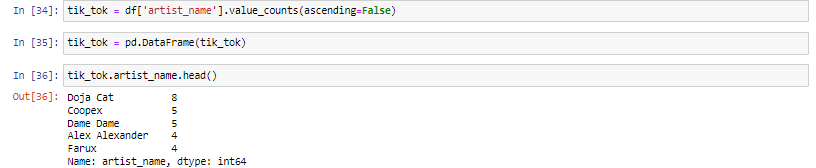
Теперь рассмотрим статистику популярности исполнителей



Статистика популярности артистов

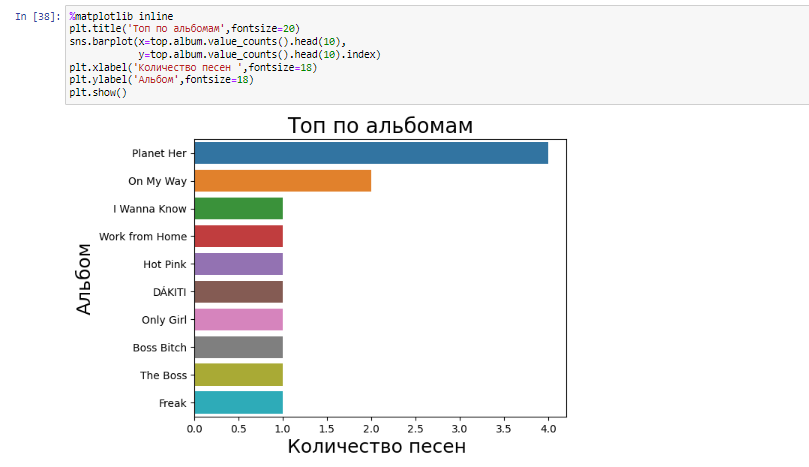
Количество популярных и менее популярных исполнителей примерно равно.

Проверим количество песен у исполнителей и выведем тех, у кого большее количество известных треков

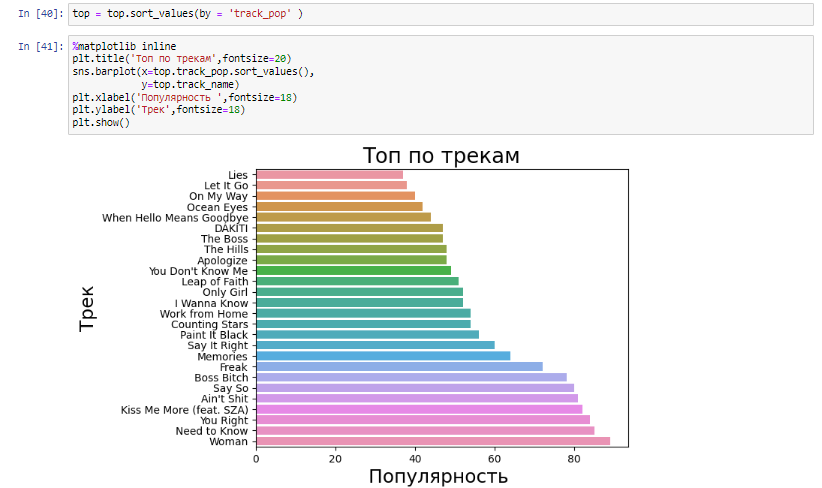


Здесь мы видим, что артисты с самым большим количеством популярных песен не входят в топ лучших. Самое большое число популярных песен наблюдается у исполнителей Doja Cat, Coopex и Dame Dame.

Выявим самые популярные альбомы

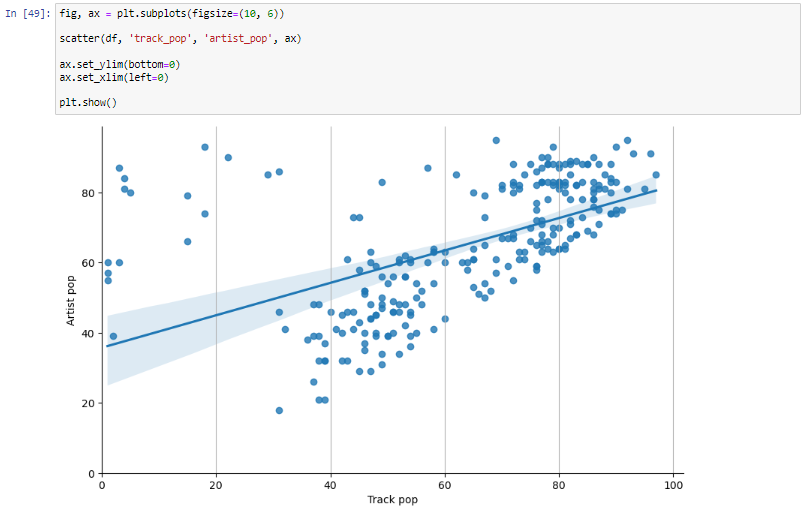


Выявим самые популярные треки

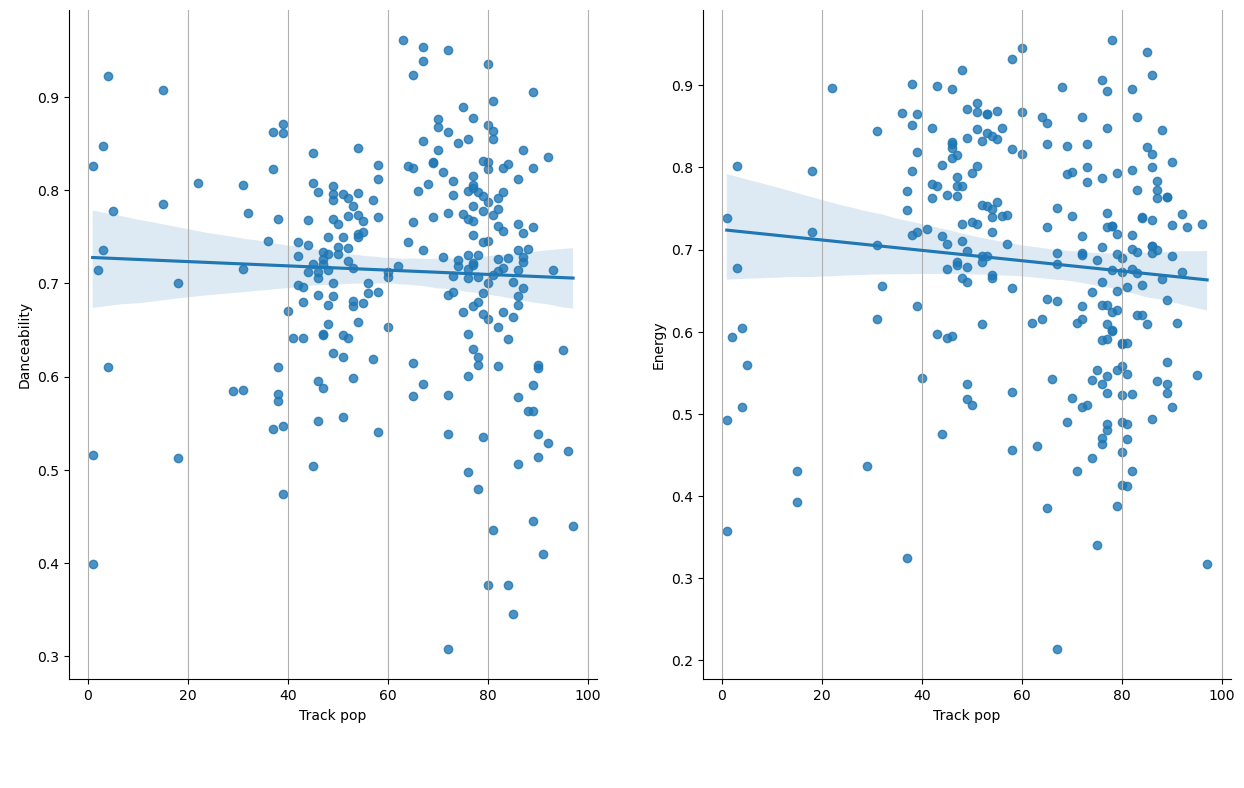


Анализируя полученные графики, можно увидеть, что самым популярным альбомом является альбом «Planet Her» за авторством Doja Cat, и самая известная песня под названием «Woman» входит в этот альбом.

Определим зависимость популярности трека от рейтинга артиста.



Теперь рассмотрим зависимости популярности трека от зажигательности и энергичности.



Исходя из проведённого анализа, видно, что популярность трека не зависит от его зажигательности и энергичности, но зависит от известности артиста.

Построим тепловую карту, чтобы посмотреть корреляцию между переменными

