Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Отчет по лабораторной работе №1

«Разведочный анализ данных»

по дисциплине «Технологии машинного обучения»

Выполнил:
студент группы ИУ5Ц-84Б
Тихонова Д.Д.
подпись, дата

Проверил:
к.т.н., доц., Ю.Е. Гапанюк
подпись, дата

2025 г.

СОДЕРЖАНИЕ ОТЧЕТА

1. Цель лабораторной работы

Изучение различных методов визуализация данных.

2. Описание задание

- Выбрать набор данных (датасет). Вы можете найти список свободно распространяемых датасетов здесь https://github.com/ugapanyuk/courses_current/wiki/DSLIST.
- Для первой лабораторной работы рекомендуется использовать датасет без пропусков в данных, например из Scikit-learn.
- Пример преобразования датасетов Scikit-learn в Pandas Dataframe можно посмотреть здесь - https://github.com/ugapanyuk/courses_current/blob/main/notebooks/ds/sklearn_datasets.ipynb.

Для лабораторных работ не рекомендуется выбирать датасеты большого размера.

- Создать ноутбук, который содержит следующие разделы:

1. Текстовое описание выбранного Вами набора данных.
2. Основные характеристики датасета.
3. Визуальное исследование датасета.
4. Информация о корреляции признаков.

- Сформировать отчет и разместить его в своем репозитории на github.

Средства и способы визуализации данных можно посмотреть здесь - https://github.com/ugapanyuk/courses_current/wiki/VISUAL.

В качестве опорного примера для выполнения лабораторной работы можно использовать пример - https://github.com/ugapanyuk/courses_current/blob/main/notebooks/eda/eda_visualization.ipynb.

3. Основные характеристики датасета

Название датасета: Most Streamed Spotify Songs 2024 (Самые транслируемые песни Spotify в 2024 году)
Ссылка: https://www.kaggle.com/datasets/nelgiriyewithana/most-streamed-spotify-songs-2024

О датасетах

Этот набор данных представляет собой исчерпывающую подборку самых популярных песен на Spotify в 2024 году. Он содержит подробную информацию о характеристиках каждого трека, его популярности и присутствии на различных музыкальных платформах, что делает его ценным ресурсом для музыкальных аналитиков, энтузиастов и профессионалов отрасли. Датасет состоит из 27 столбцов (после удаления некоторых столбцов) и 4598 строк, где каждая строка представляет собой запись о песне.

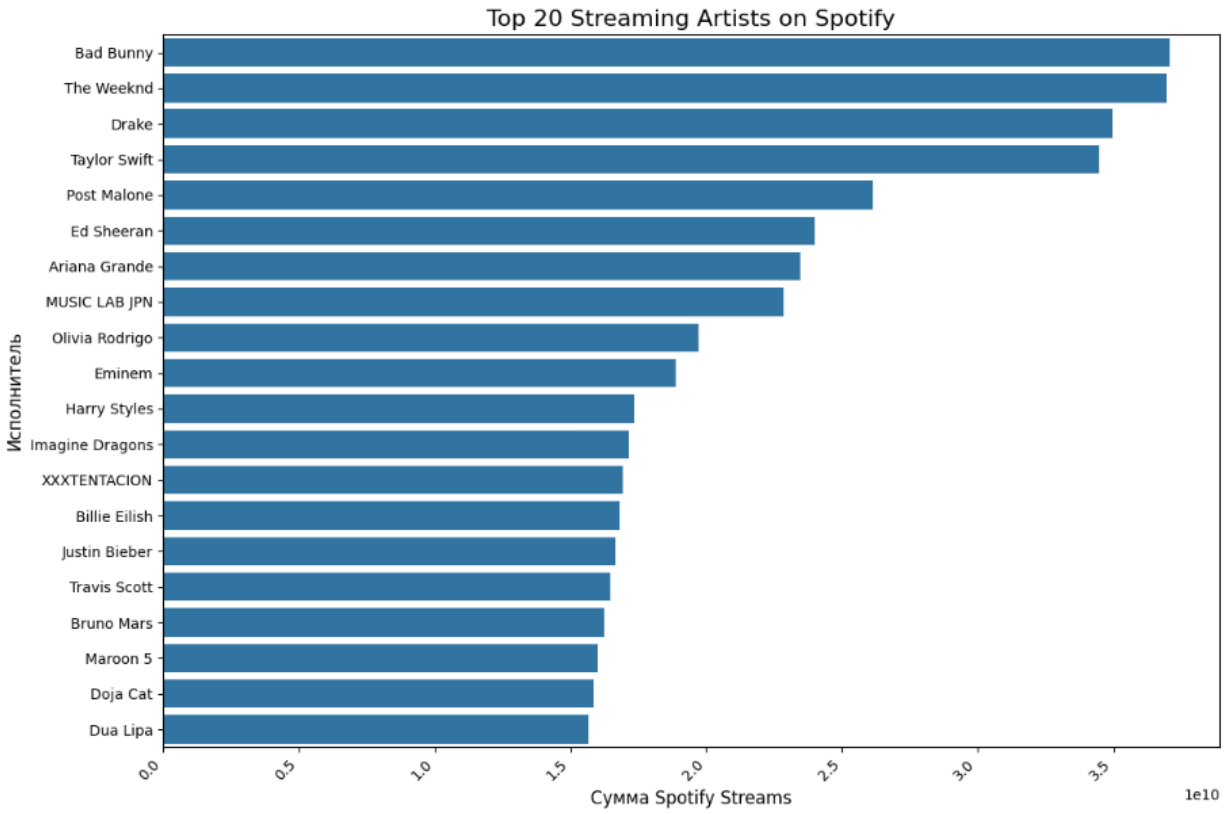
Структура данных:

- **Track:** Название песни.
- **Album Name:** Название альбома, к которому принадлежит песня.
- **Artist:** Имя исполнителя (или исполнителей) песни.
- **Release Date:** Дата выпуска песни.
- **ISRC:** Международный стандартный код записи песни.
- **All Time Rank:** Рейтинг песни на основе ее популярности за всё время.
- **Track Score:** Оценка, присвоенная треку на основе различных факторов.
- **Spotify Streams:** Общее количество прослушиваний песни на Spotify.
- **Spotify Playlist Count:** Количество списков воспроизведения Spotify, в которые включена песня.
- **Spotify Playlist Reach:** Охват песни во всех плейлистах Spotify (суммарное количество подписчиков плейлистов).
- **Spotify Popularity:** Показатель популярности песни на Spotify (от 0 до 100).

- **YouTube Views:** Общее количество просмотров официального видео с песней на YouTube.
- **YouTube Likes:** Общее количество лайков на официальном видео с песней на YouTube.
- **TikTok Posts:** Количество сообщений в TikTok с участием песни.
- **TikTok Likes:** Общее количество лайков на публикациях TikTok с песней.
- **TikTok Views:** Общее количество просмотров сообщений TikTok с участием песни.
- **YouTube Playlist Reach:** Охват песни во всех плейлистах YouTube (суммарное количество подписчиков плейлистов).
- **Apple Music Playlist Count:** Количество плейлистов Apple Music, в которые включена песня.
- **AirPlay Spins:** Количество раз, которое песня воспроизводилась на радиостанциях.
- **SiriusXM Spins:** Количество раз, которое песня звучала на SiriusXM.
- **Deezer Playlist Count:** Количество плейлистов Deezer, в которые включена песня.
- **Deezer Playlist Reach:** Охват песни во всех плейлистах Deezer (суммарное количество подписчиков плейлистов).
- **Amazon Playlist Count:** Количество плейлистов Amazon Music, в которые включена песня.
- **Pandora Streams:** Общее количество прослушиваний на Pandora.
- **Pandora Track Stations:** Количество радиостанций Pandora, на которых звучит песня.
- **Shazam Counts:** Общее количество раз, когда песня была распознана с помощью Shazam.
- **Explicit Track:** Указывает, содержит ли песня откровенный контент (1 - да, 0 - нет).

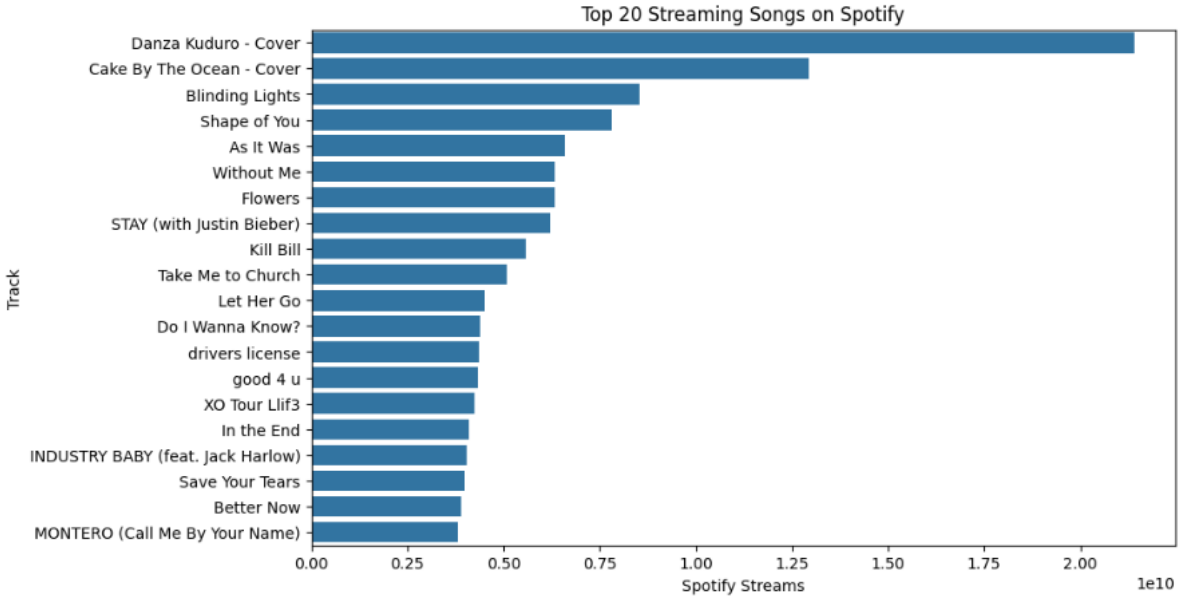
4. Визуальное исследование датасета

4.1. Top 20 исполнителей на Spotify



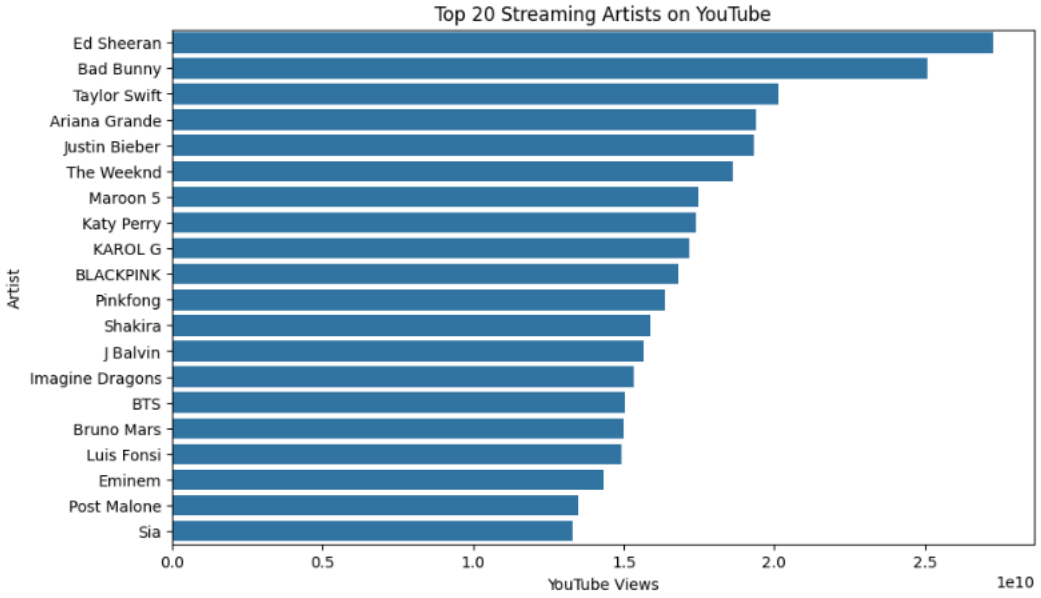
Данный график отображает топ-20 исполнителей по суммарному количеству прослушиваний на Spotify. Видим, что лидирует Bad Bunny, за ним The Weeknd и т.д.

4.2. Top 20 песен на Spotify



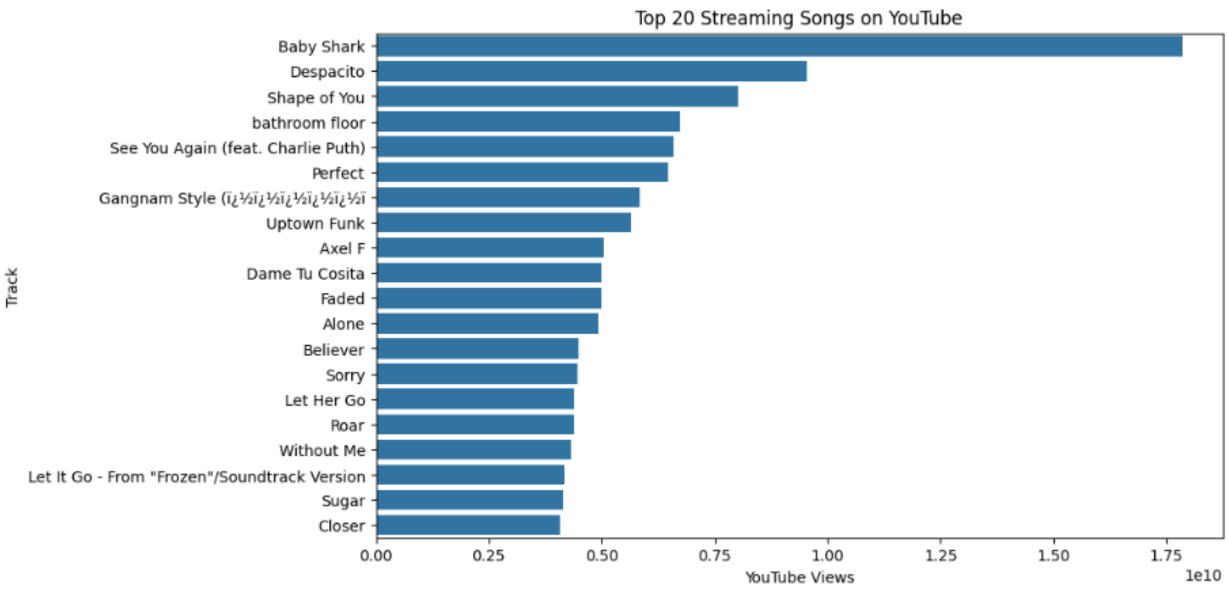
Данный график отображает топ-20 песен по суммарному количеству прослушиваний на Spotify. Самой популярной песней является Danza Kuduro – Cover.

4.3. Top 20 исполнителей на YouTube



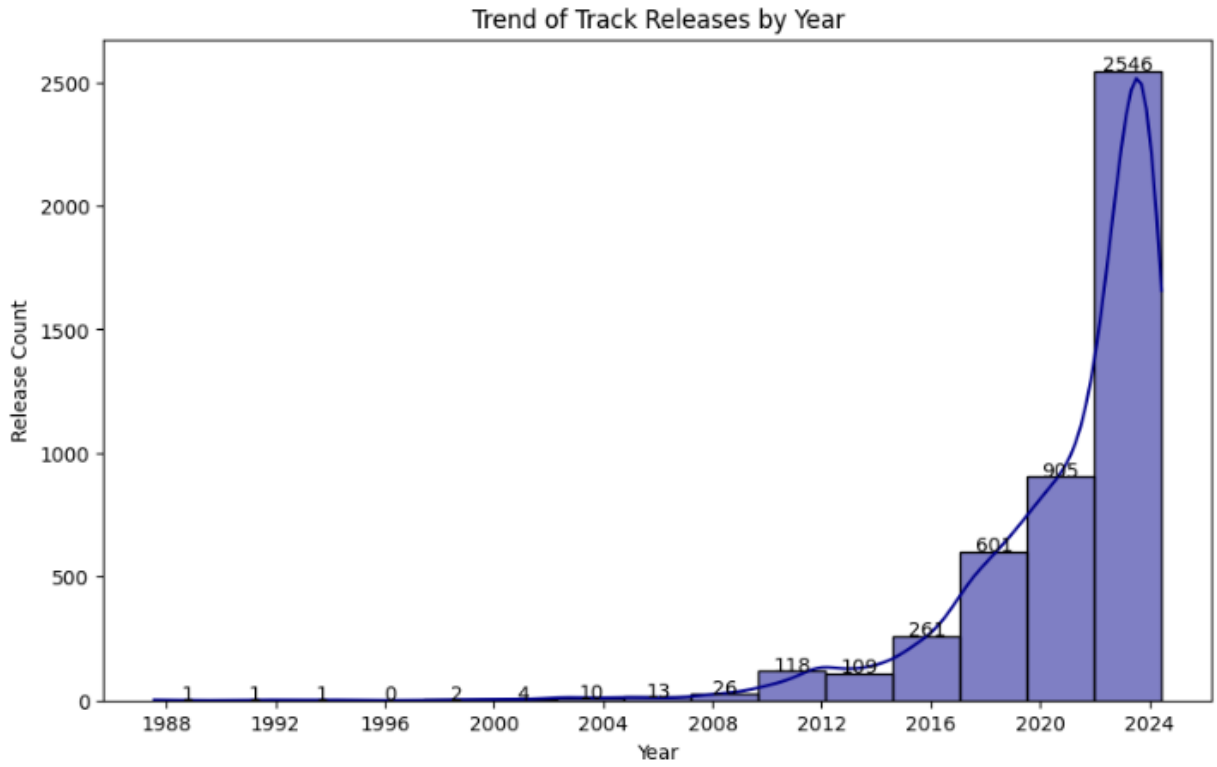
Данный график отображает топ-20 исполнителей по суммарному количеству просмотров на YouTube. Мы видим, что Ed Sheeran является самым популярным исполнителем на YouTube, The Bad Bunny - на 2м месте и так далее.

4.4. Top 20 песен на YouTube



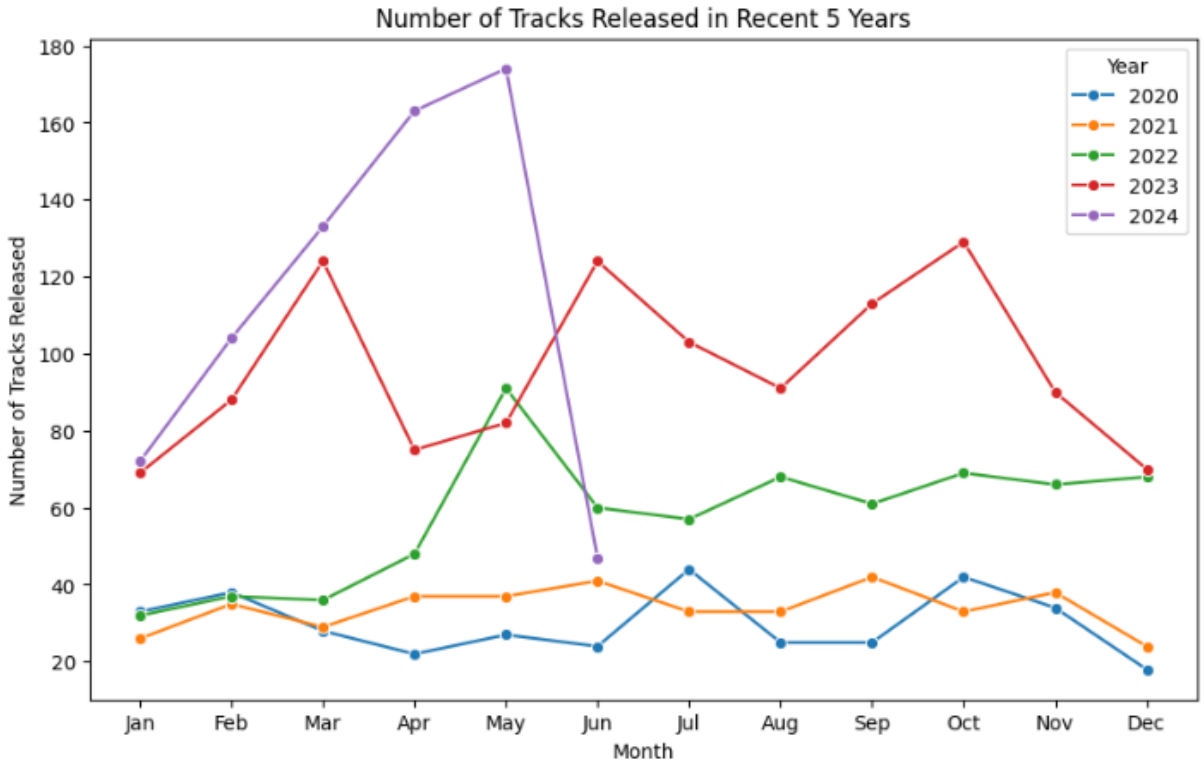
Данный график отображает топ-20 песен по суммарному количеству просмотров на YouTube. Самой популярной песней является Baby Shark.

4.5. Гистограмма распределения релизов по годам



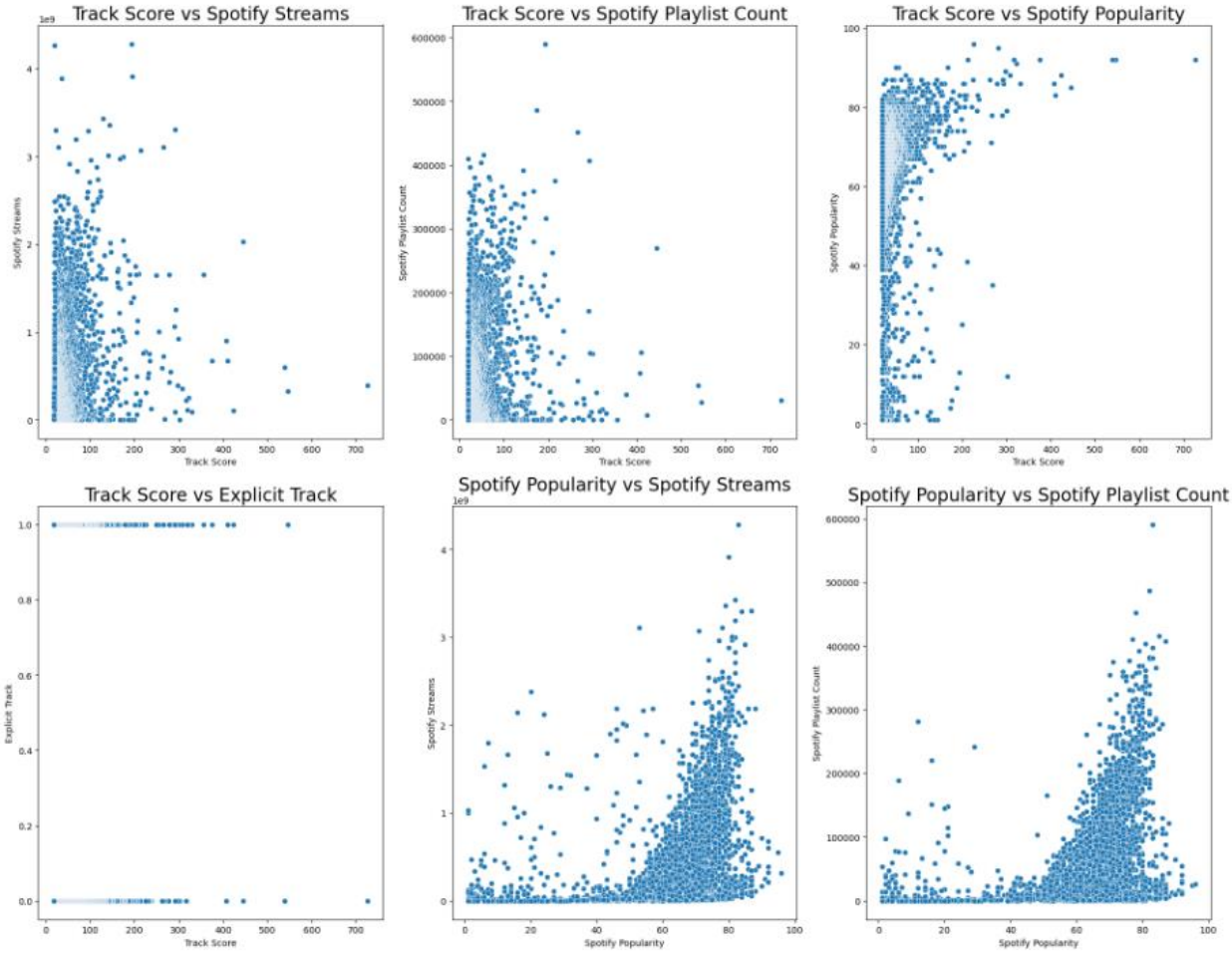
Гистограмма показывает распределение релизов треков по годам. Видно, что наибольшее количество треков было выпущено в районе 2022-2024.

4.6. Количество треков, выпущенных в последние 5 лет по месяцам



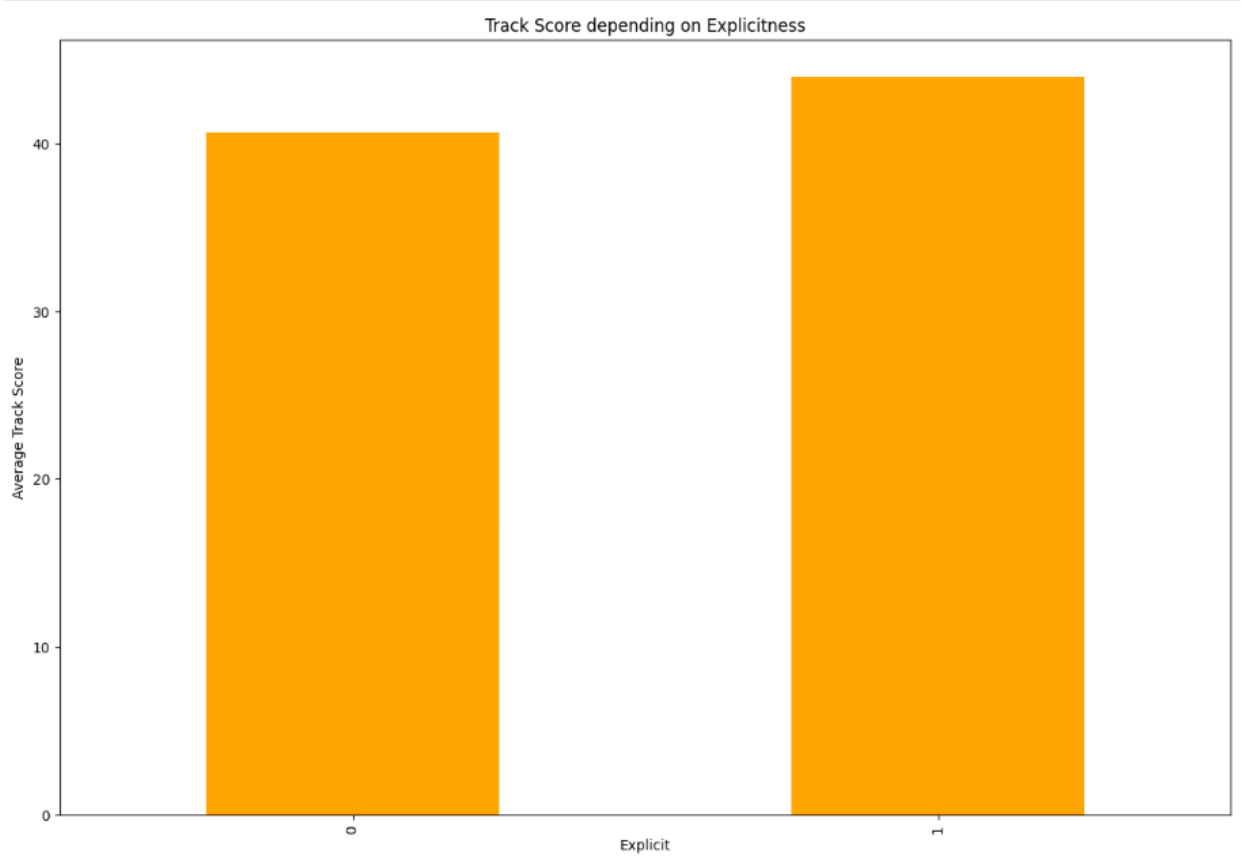
Линейный график показывает количество треков, выпущенных в последние 5 лет по месяцам. Видно, что весной наблюдается пик релизов, а в декабре - спад. Сравнение линий разных цветов позволяет увидеть, как эти тенденции меняются год от года.

4.7. Диаграммы рассеяния для изучения взаимосвязей между признаками



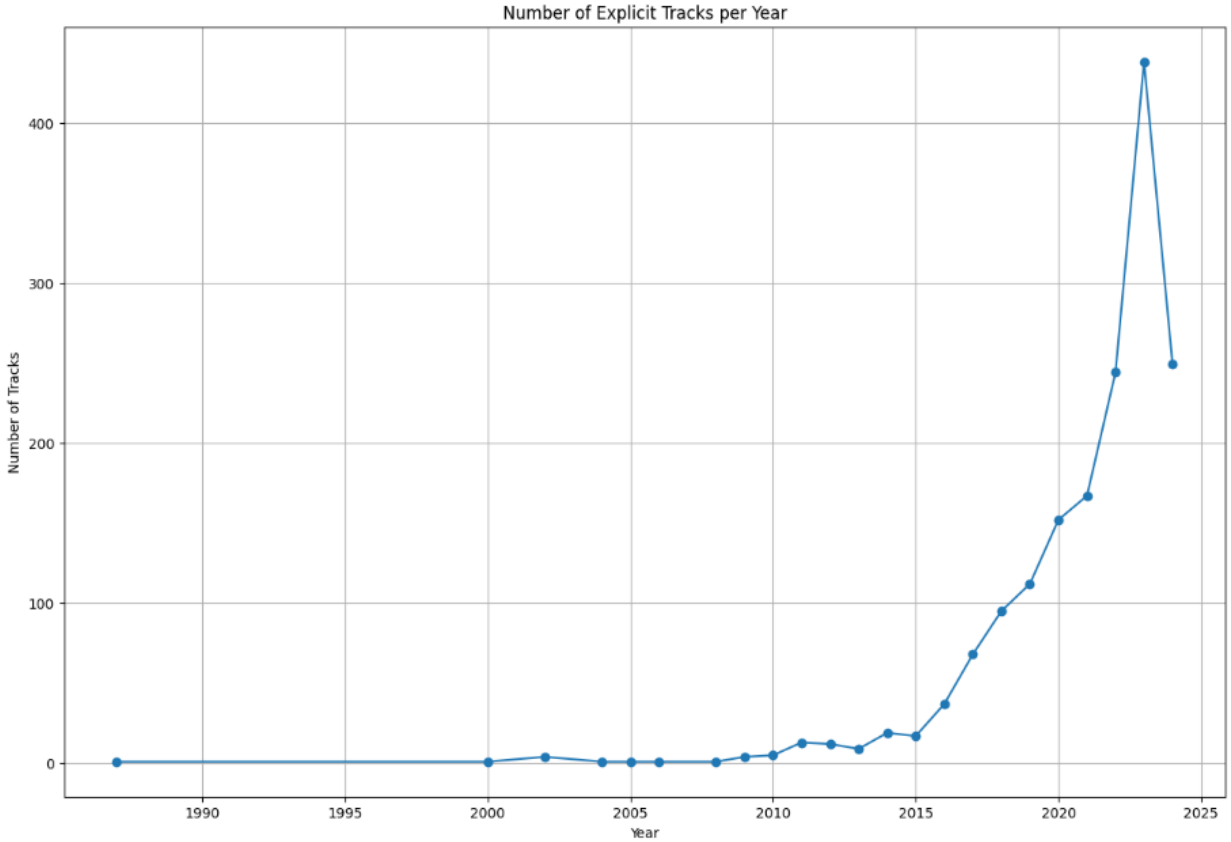
Набор диаграмм рассеяния позволяет оценить взаимосвязи между различными признаками. Например, на графике Track Score vs Spotify Streams можно увидеть, есть ли тенденция к увеличению количества прослушиваний с ростом оценки трека. Аналогично анализируются и другие графики. В частности, наблюдается положительная корреляция между “Spotify Popularity” и “Spotify Streams”, а также между “Spotify Popularity” и “Spotify Playlist Count”.

4.8. Средняя оценка трека в зависимости от наличия откровенного контента



Столбчатая диаграмма показывает среднюю оценку треков в зависимости от наличия откровенного контента. Видно, что средняя оценка треков с откровенным контентом ниже средней оценки треков без откровенного контента.

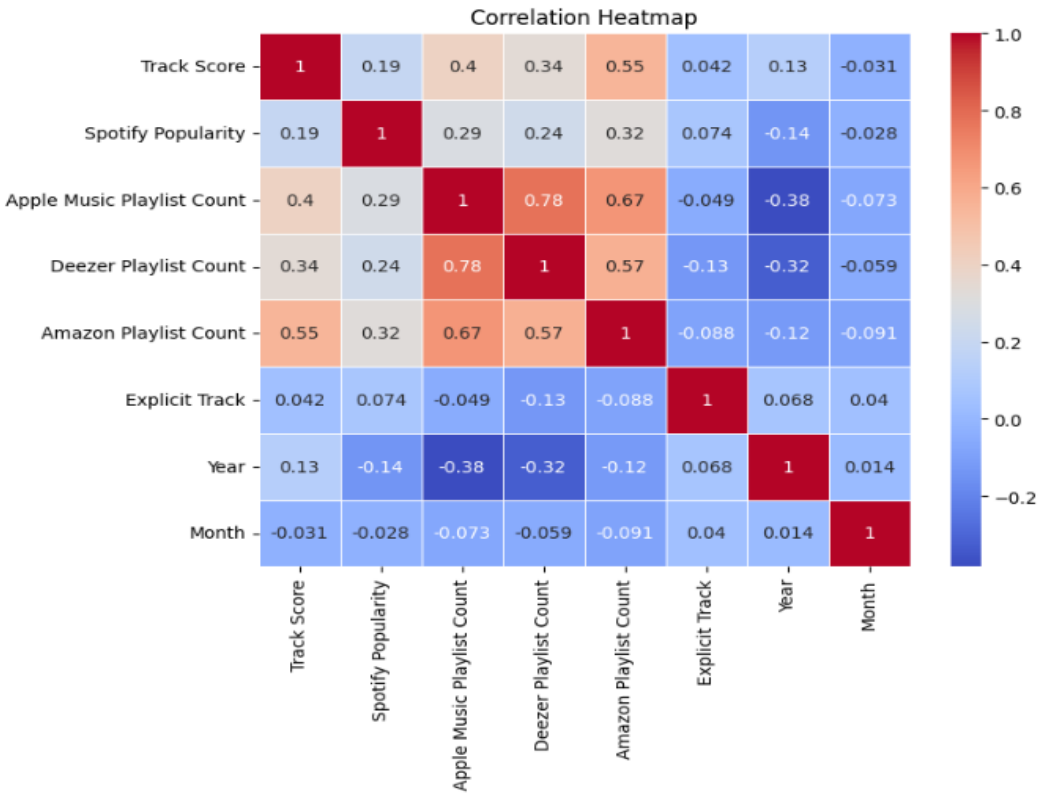
4.9. Количество треков с откровенным контентом по годам



Линейный график показывает количество треков с откровенным контентом по годам. Видно, что количество таких треков увеличивается с течением времени. В 2024 наблюдается пик в количестве треков с откровенным контентом.

5. Информация о корреляции признаков

Для анализа взаимосвязей между числовыми признаками была построена корреляционная матрица:



Отрицательная корреляция между “Year” и “Apple Music Playlist Count” (-0.38) и “Year” и “Deezer Playlist Count” (-0.32) говорит о том, что со временем (с увеличением года выпуска песни) количество плейлистов Apple Music и Deezer, в которые добавляется песня, имеет тенденцию уменьшаться. Положительная корреляция между “Track Score” и “Amazon Playlist Count” (0.55) говорит о том, что чем выше оценка трека, тем больше вероятность, что он будет добавлен в плейлисты Amazon Music. Сильная положительная корреляция между “Apple Music Playlist Count” и “Deezer Playlist Count” (0.78), “Apple Music Playlist Count” и “Amazon Playlist Count” (0.67), и “Deezer Playlist Count” и “Amazon Playlist Count” (0.57) указывает на то, что если песня часто добавляется в плейлисты на одной платформе (например, Apple Music), то она с большой вероятностью будет добавлена и в плейлисты на других платформах (Deezer и Amazon Music). Очень слабая корреляция между “Year” и “Track Score” (0.13) говорит о том, что год выпуска песни практически не влияет на ее оценку.

6. Итог

6.1. Анализ данных

- Обнаружена отрицательная корреляция между годом выпуска песни ("Year") и количеством плейлистов на Apple Music и Deezer ("Apple Music Playlist Count" и "Deezer Playlist Count"). Это может свидетельствовать о том, что новые треки с меньшей вероятностью добавляются в плейлисты на этих платформах по сравнению со старыми. Возможно, это связано с изменением алгоритмов или с изменением музыкальных предпочтений слушателей.
- Выявлена положительная корреляция между оценкой трека ("Track Score") и количеством плейлистов на Amazon Music ("Amazon Playlist Count"). Это может указывать на то, что алгоритмы Amazon Music отдают предпочтение трекам с более высокой оценкой при формировании плейлистов.
- Подтверждена сильная взаимосвязь между количеством плейлистов на разных платформах ("Apple Music Playlist Count", "Deezer Playlist Count" и "Amazon Playlist Count"). Это говорит о том, что существует общая тенденция к тому, что песни, популярные на одной платформе, также становятся популярными и на других.
- Анализ показал, что год выпуска песни ("Year") практически не влияет на её оценку ("Track Score"). Это может говорить о том, что оценка трека больше зависит от его музыкальных качеств, чем от времени его выпуска.