



Высшая школа бизнеса

Бизнес-информатика

Москва
2025

Групповой проект 1

Подготовили:
Нелепова Дарья
Фадееенкова Екатерина
Калашникова Мария
Ившина Александра

Фролова Юлия
Чепорев Никита

Подготовка данных

Очистка и подготовка данных

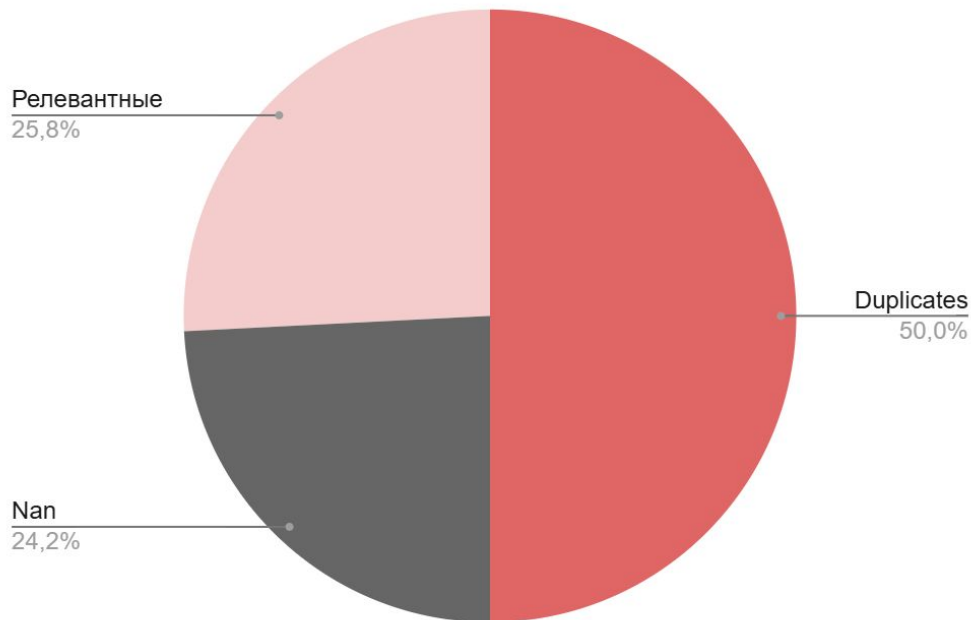
- title
- rating
- ratingLevel
- ratingDescription
- release year
- user rating score
- user rating size



- title
- rating_description
- release_year
- user_rating_score

NaN и duplicates

Изначальная структура данных



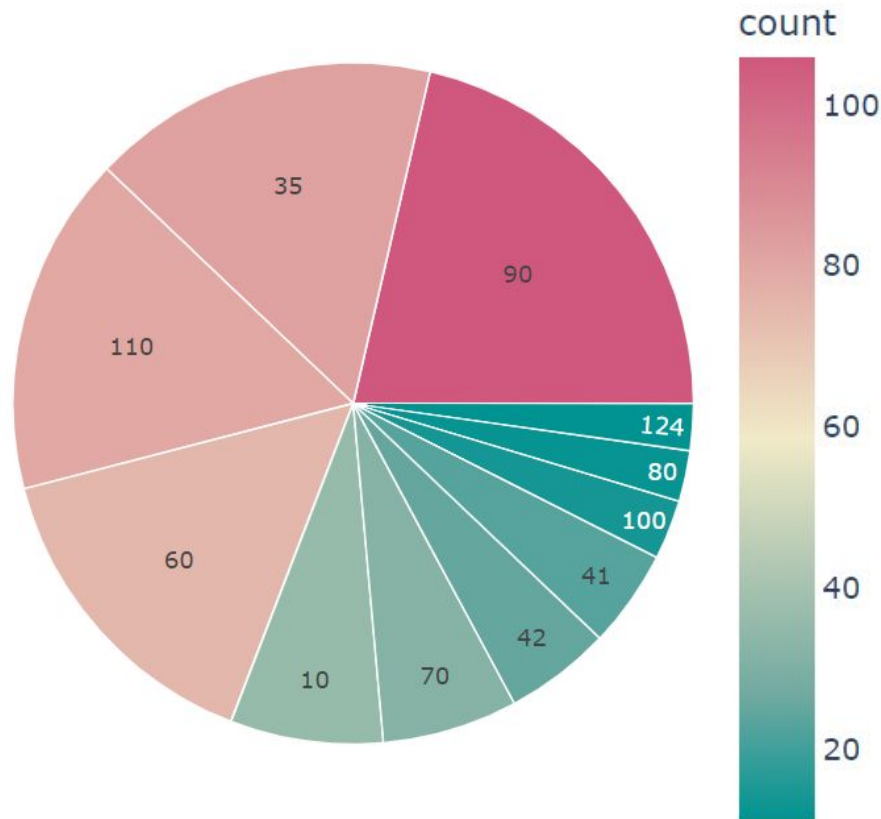
Как решали проблему:

1. Дроп дубликатов
2. Подключение внешних датасетов
3. Работа с двумя датасетами параллельно

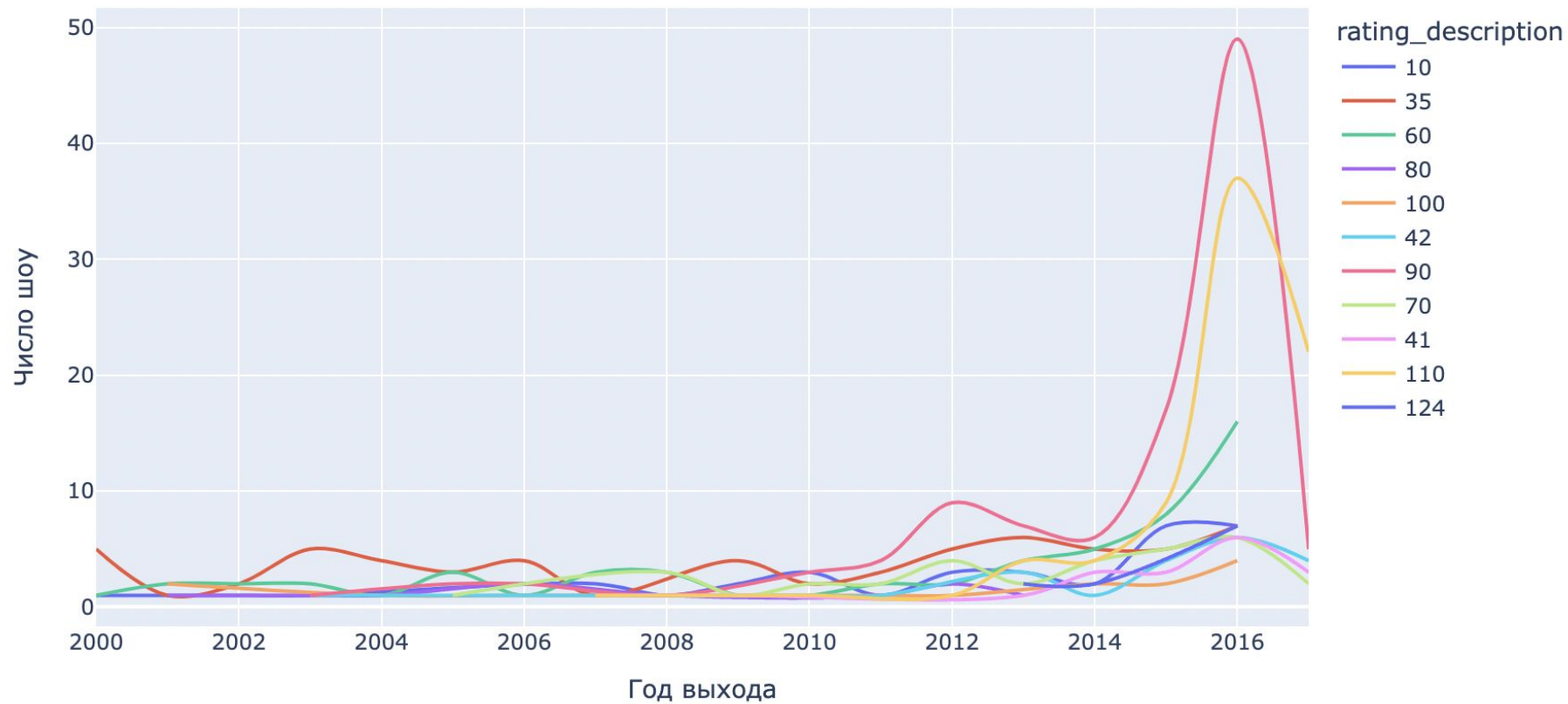
EDA

Распределение шоу по рейтинговым категориям

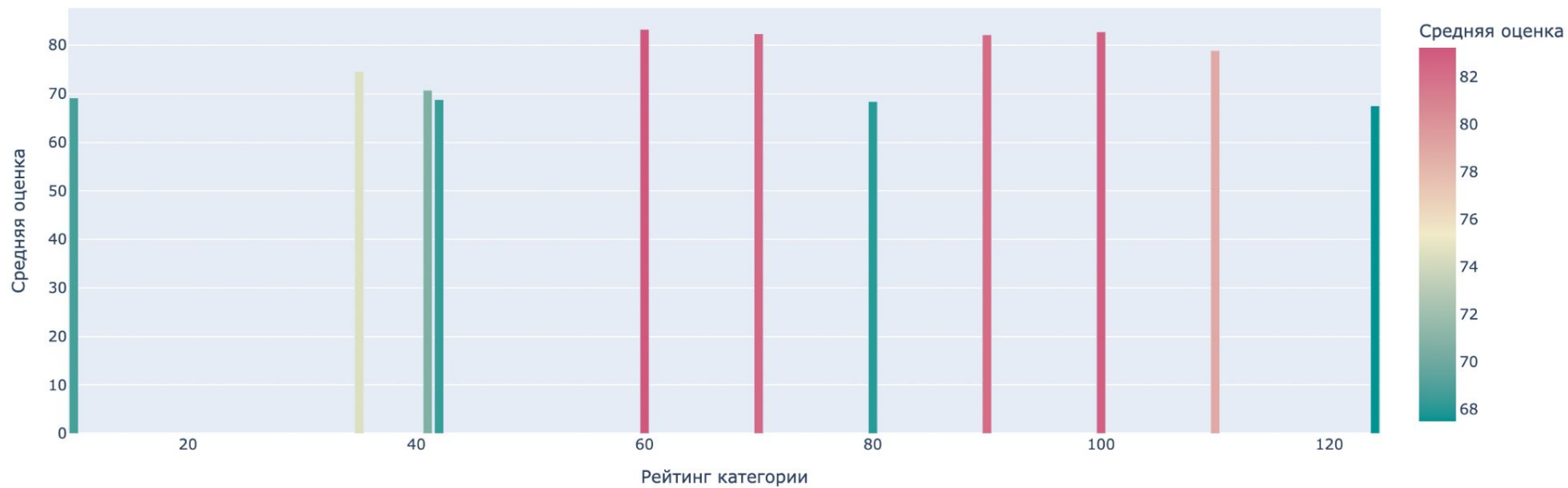
90: Parents strongly cautioned. May be unsuitable for children under 14.
35: Suitable for all ages.
110: For mature audiences. May not be suitable for children under 17.
60: Parental guidance suggested. May not be suitable for children.
10: Suitable for all ages.
70: Parental guidance suggested. May not be suitable for children.
42: Suitable for children ages 7 and older. Content may be mild.
41: Suitable for children ages 7 and older.
100: Strong violence, sexual content and adult language.
80: Crude and sexual humor, language and some drug use.
124: This movie has not been rated. Intended for adult audiences.



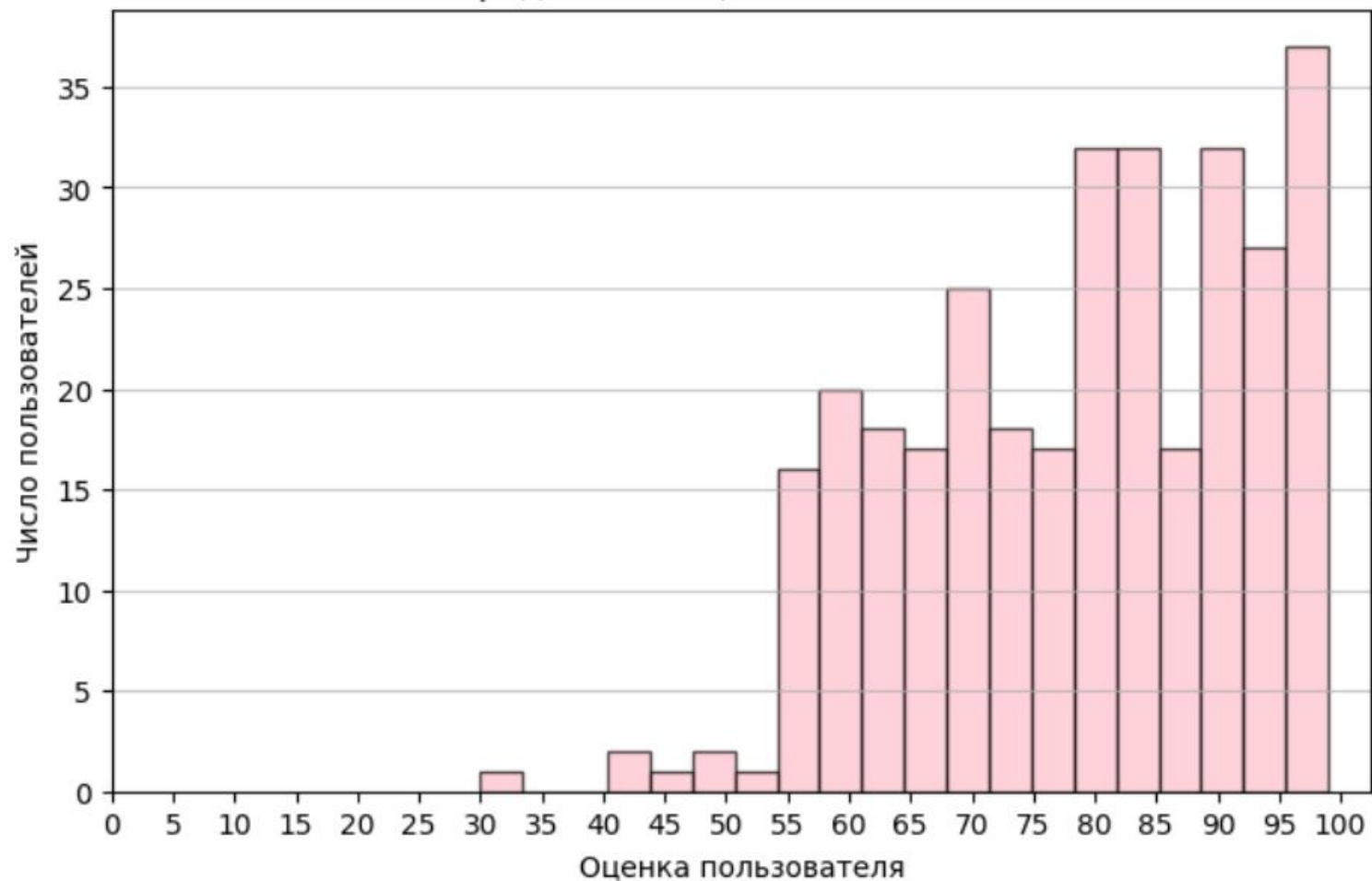
Изменение количества шоу по рейтинговым категориям (с 2000 года)



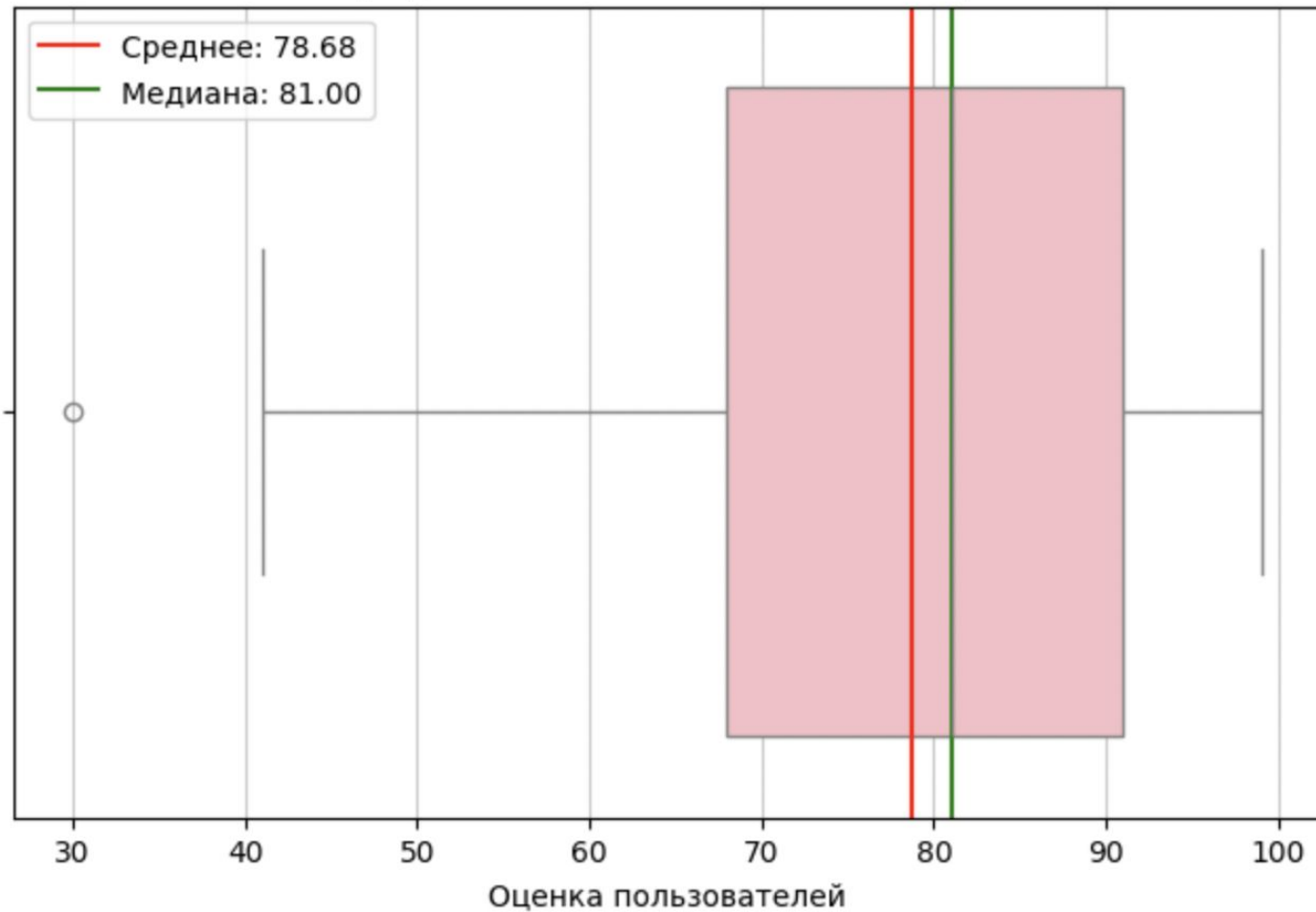
Средний рейтинг пользователей для каждой категории



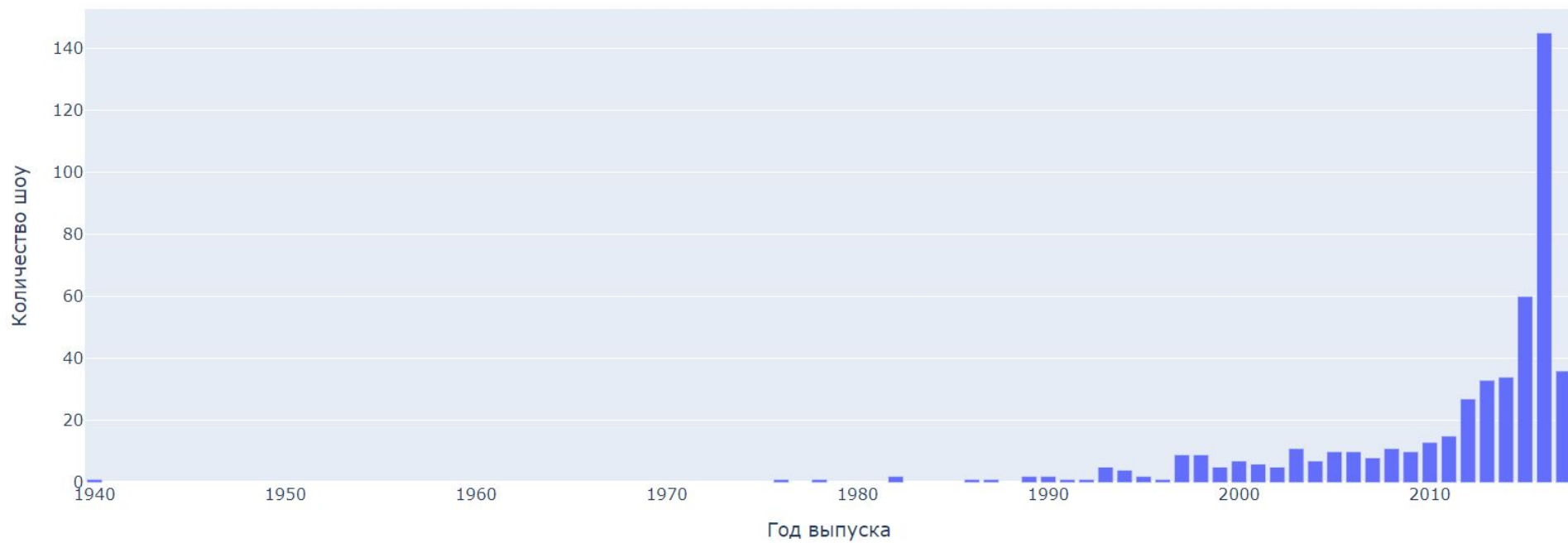
Распределение оценок пользователей



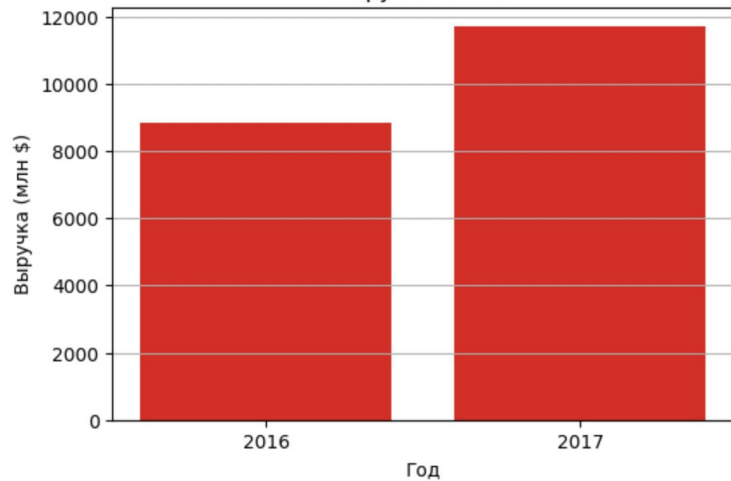
Распределение пользовательских оценок



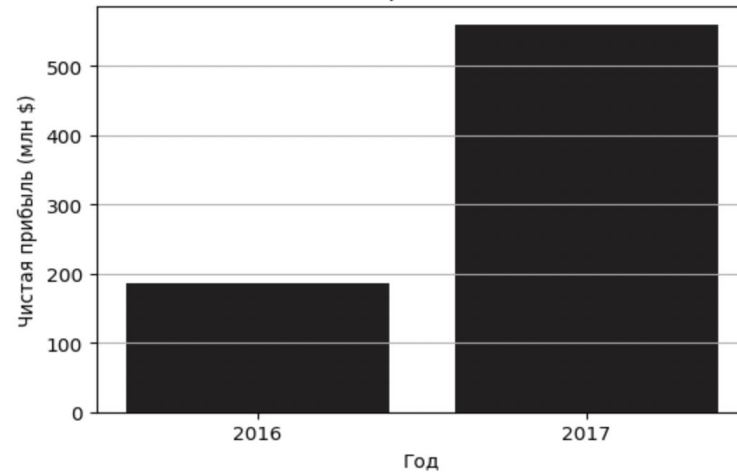
Распределение количества запущенных шоу по годам



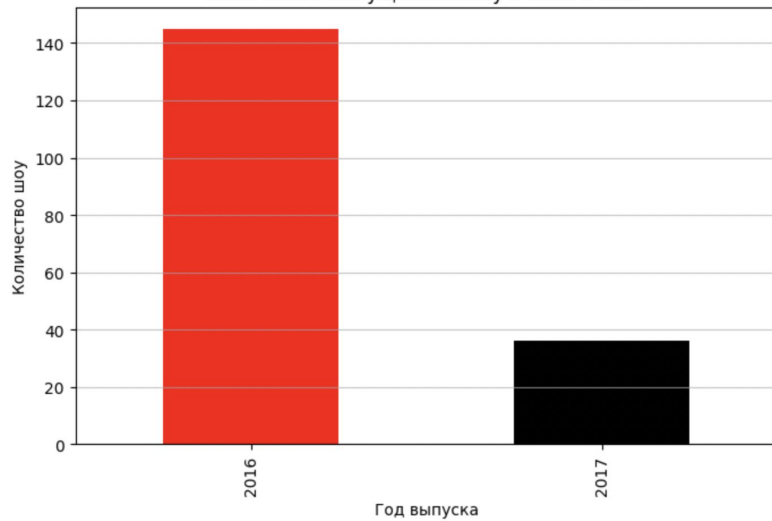
Выручка Netflix



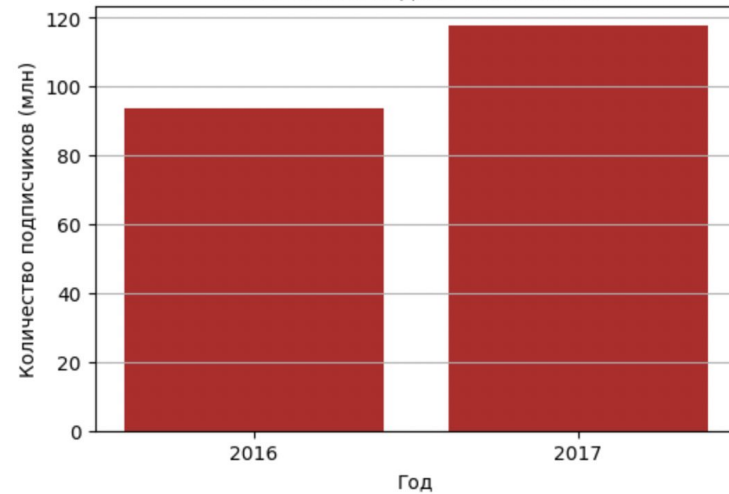
Чистая прибыль Netflix



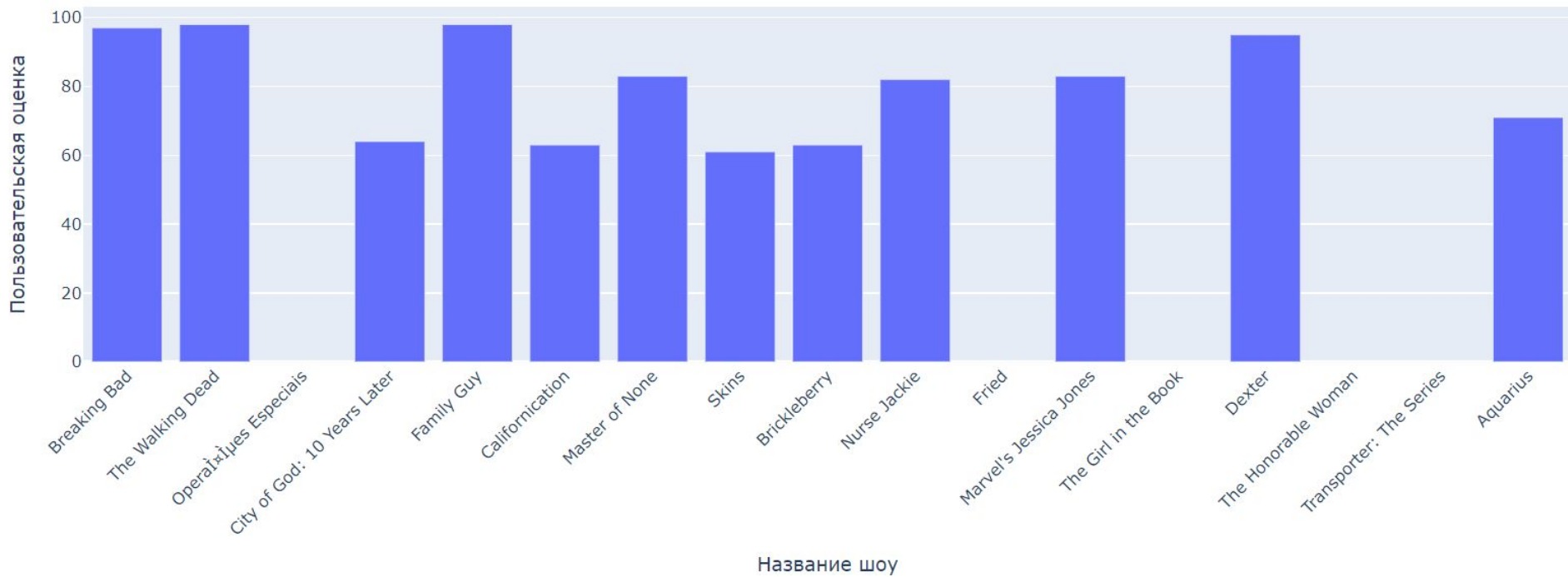
Количество выпущенных шоу в 2016 и 2017



Количество подписчиков Netflix



Сравнение 'Californication' с шоу такого же рейтинга (2013-2015)



Сравнение Californication с оценками по всем шоу из датасета:

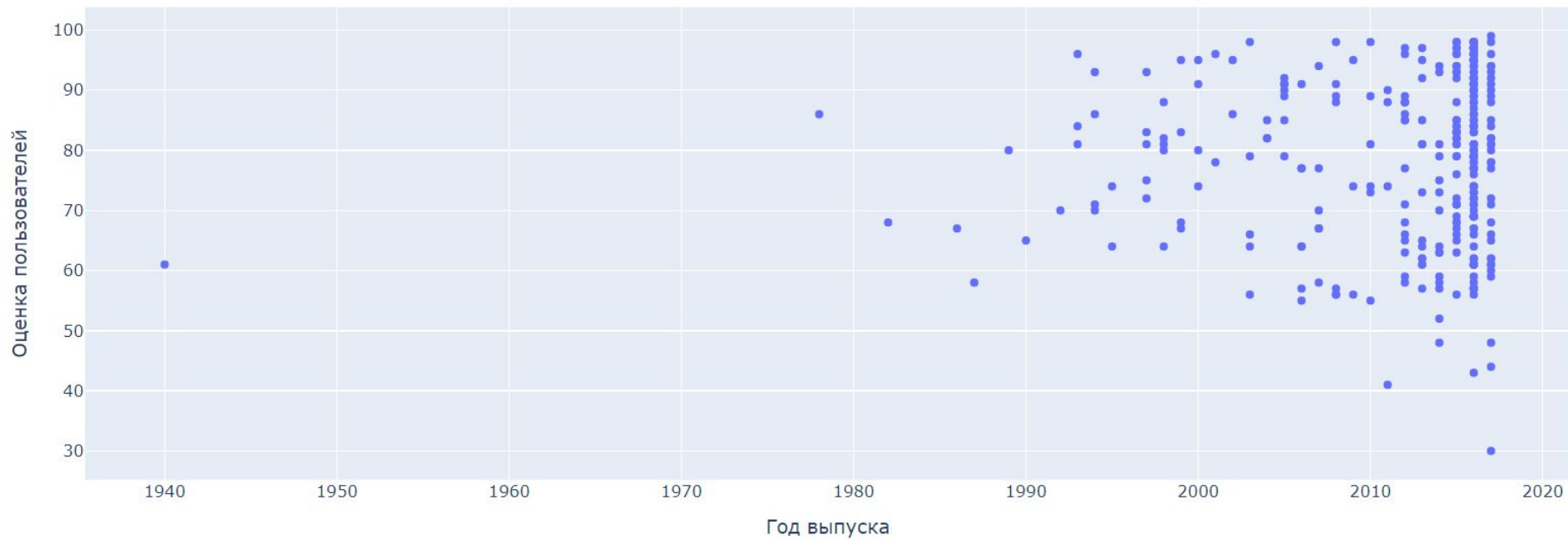
Параметр	Значение по Californication	Среднее по датасету	Медиана по датасету	Мин. значение в датасете	Макс. значение в датасете
0 Средняя оценка пользователей	63	78.679365	81.0	30.0	99.0

Важные признаки, отсутствующие в данных, но влияющие на успешность шоу

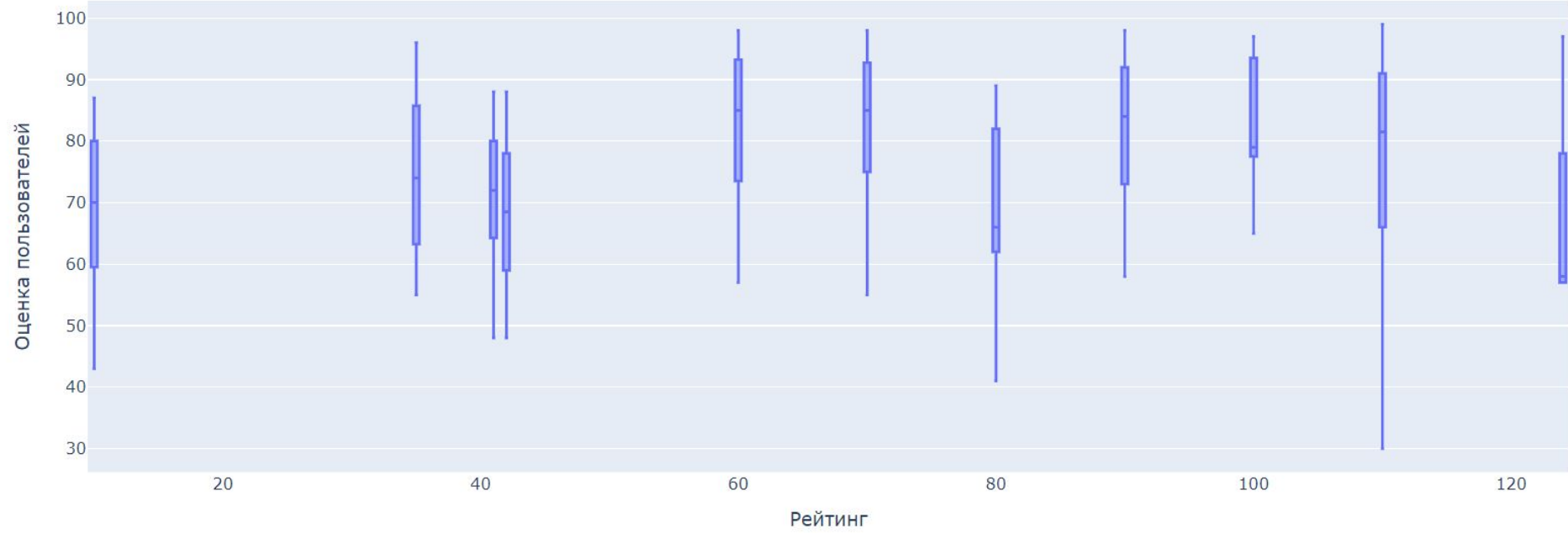


Признак	Тип данных	Влияние на успешность шоу
Бюджет производства	int (млн \$)	Высокий бюджет позволяет делать масштабные съемки, использовать спецэффекты и привлекать известных актеров.
Длительность съемок	int (дни)	Чем дольше снимается сериал, тем более он может быть проработан, но задержки могут указывать на проблемы в продакшене.
Режиссер	str	Известные режиссеры привлекают аудиторию и задают стиль проекта.
Команда	int (кол-во Людей)	Чем больше людей участвует в создании сериала, тем выше вероятность качественного продакшена и сложного сценария.

Зависимость оценок от года выпуска






Зависимость оценок от рейтинга



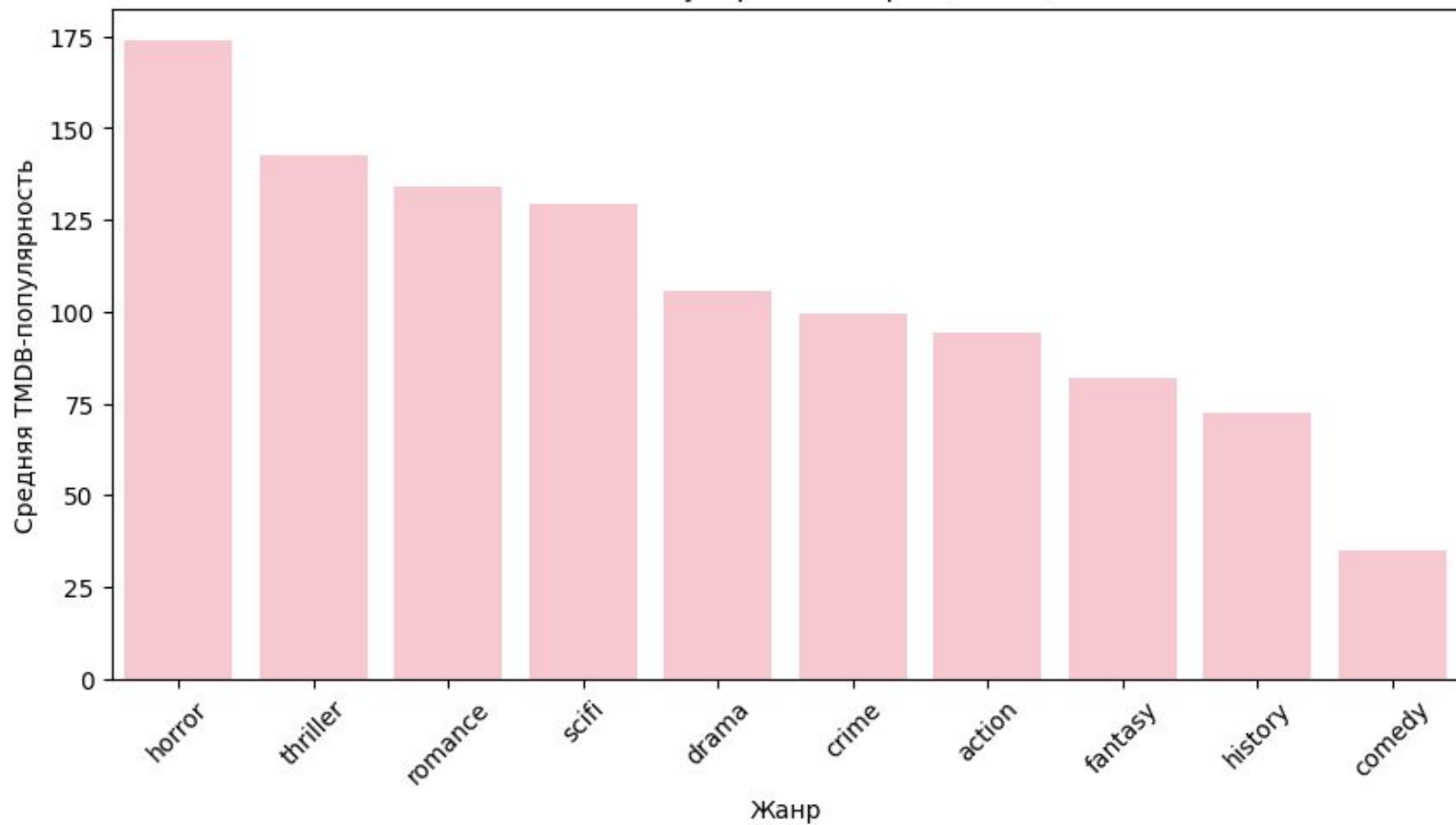
Источник данных и метод сбора информации

Данные охватывают шоу с 1940 по 2017 год

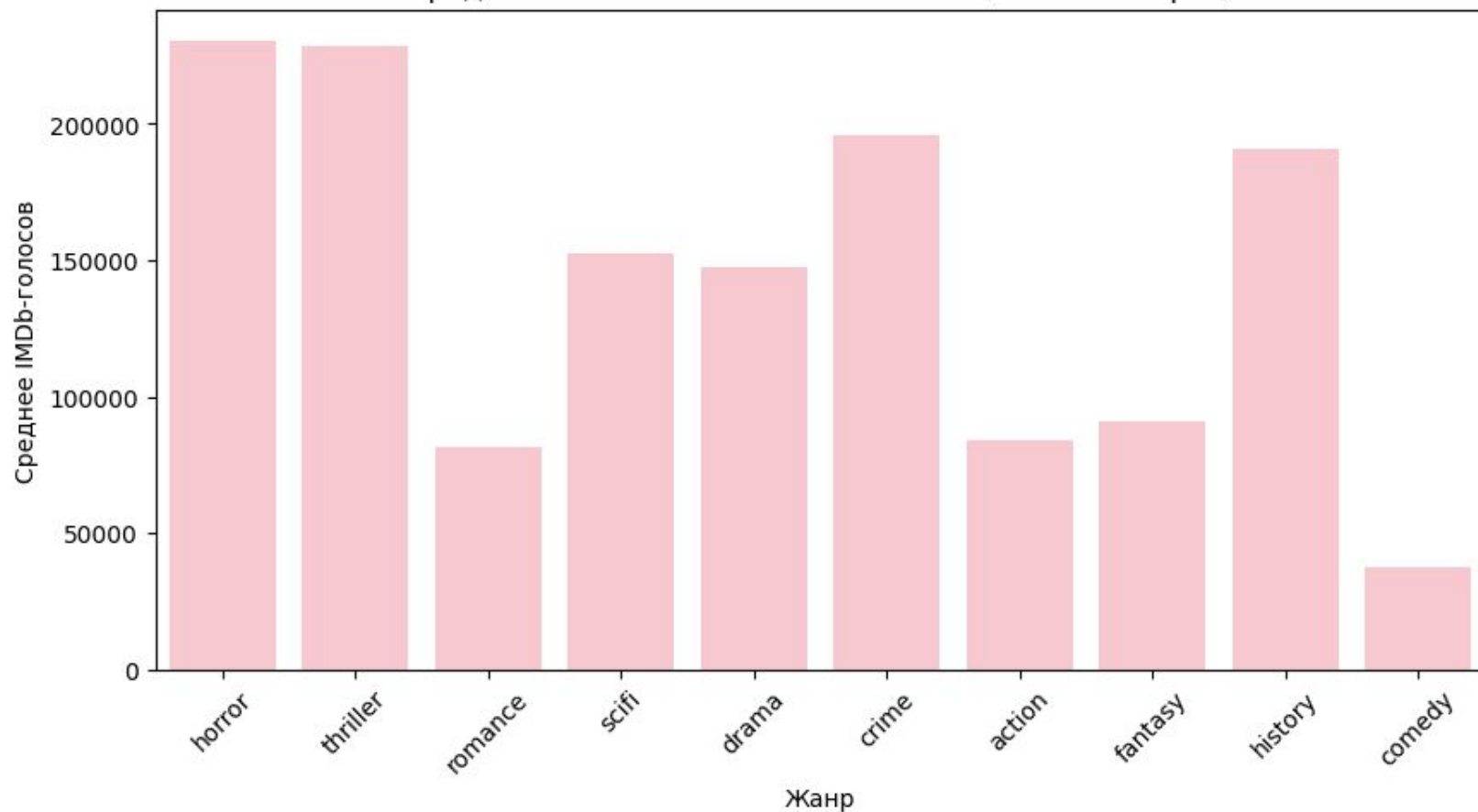
Источник	
 Netflix (платформа)	Прямые оценки пользователей (до 2017 года 5-звездочная система, после – лайк/дизлайк).
 IMDb (Internet Movie Database)	Рейтинги пользователей и метаданные о шоу, особенно для проектов до 2007 года.
 Автоматический парсинг	Данные могли быть собраны автоматически с IMDb, Netflix, Metacritic и других платформ.

Исследование контента по жанрам

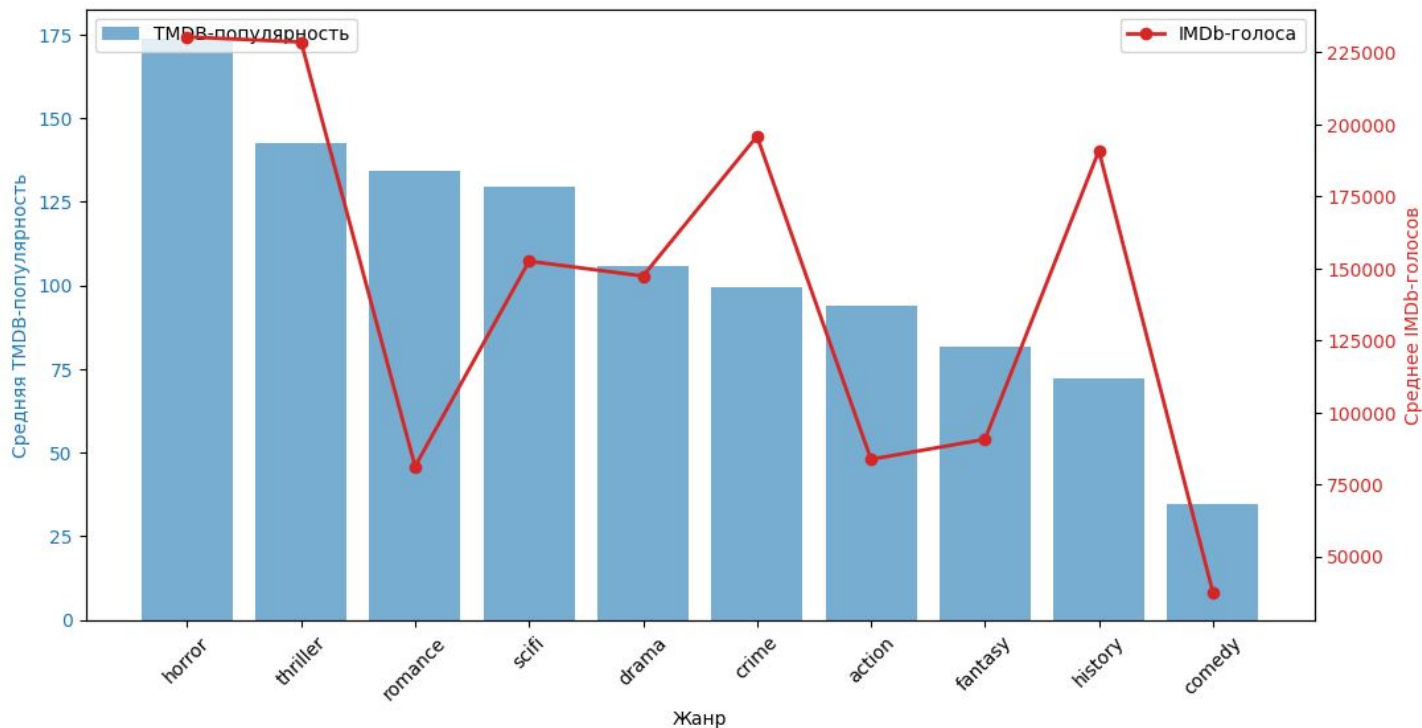
Топ-10 популярных жанров (TMDB)



Среднее количество голосов на IMDb (ТОП-10 жанров)



Топ-10 популярных жанров: TMDb vs IMDb

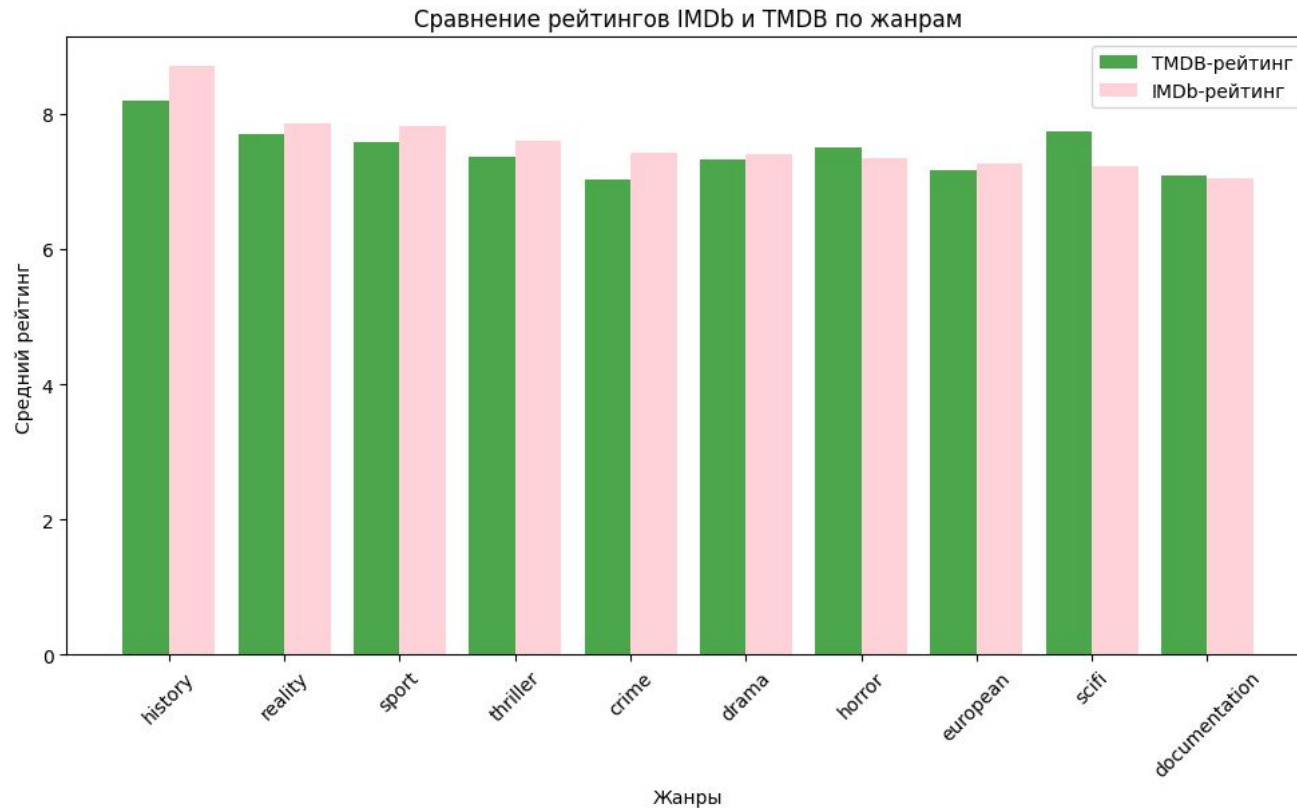


Ужасы и триллеры — основной драйвер обсуждений.

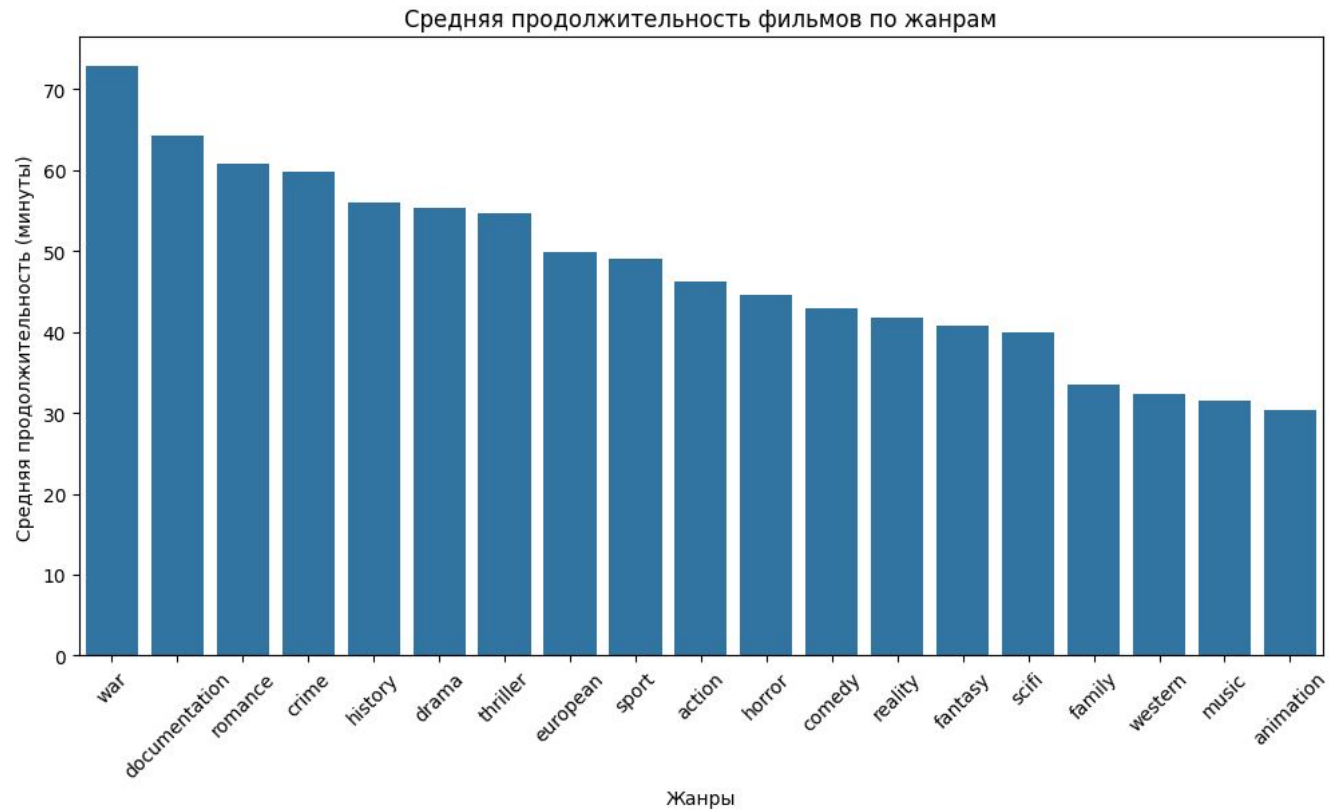
Значит, **инвестировать в эксклюзивные хорроры и триллеры — хорошая идея.**

Фантастика и драмы привлекают много зрителей, но с чуть меньшей вовлеченностью.

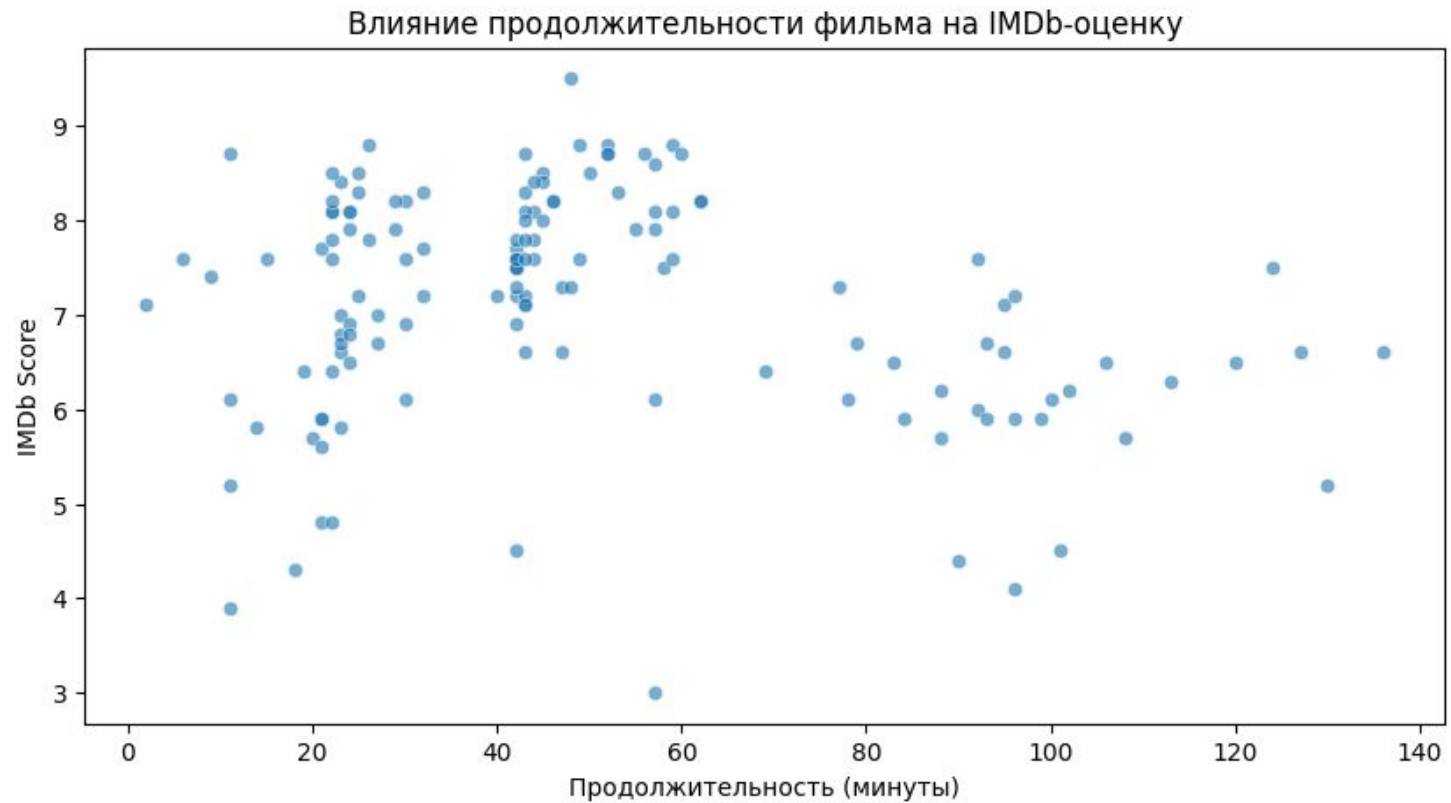
Их **стоит активно продвигать**, но не ожидать взрывной активности



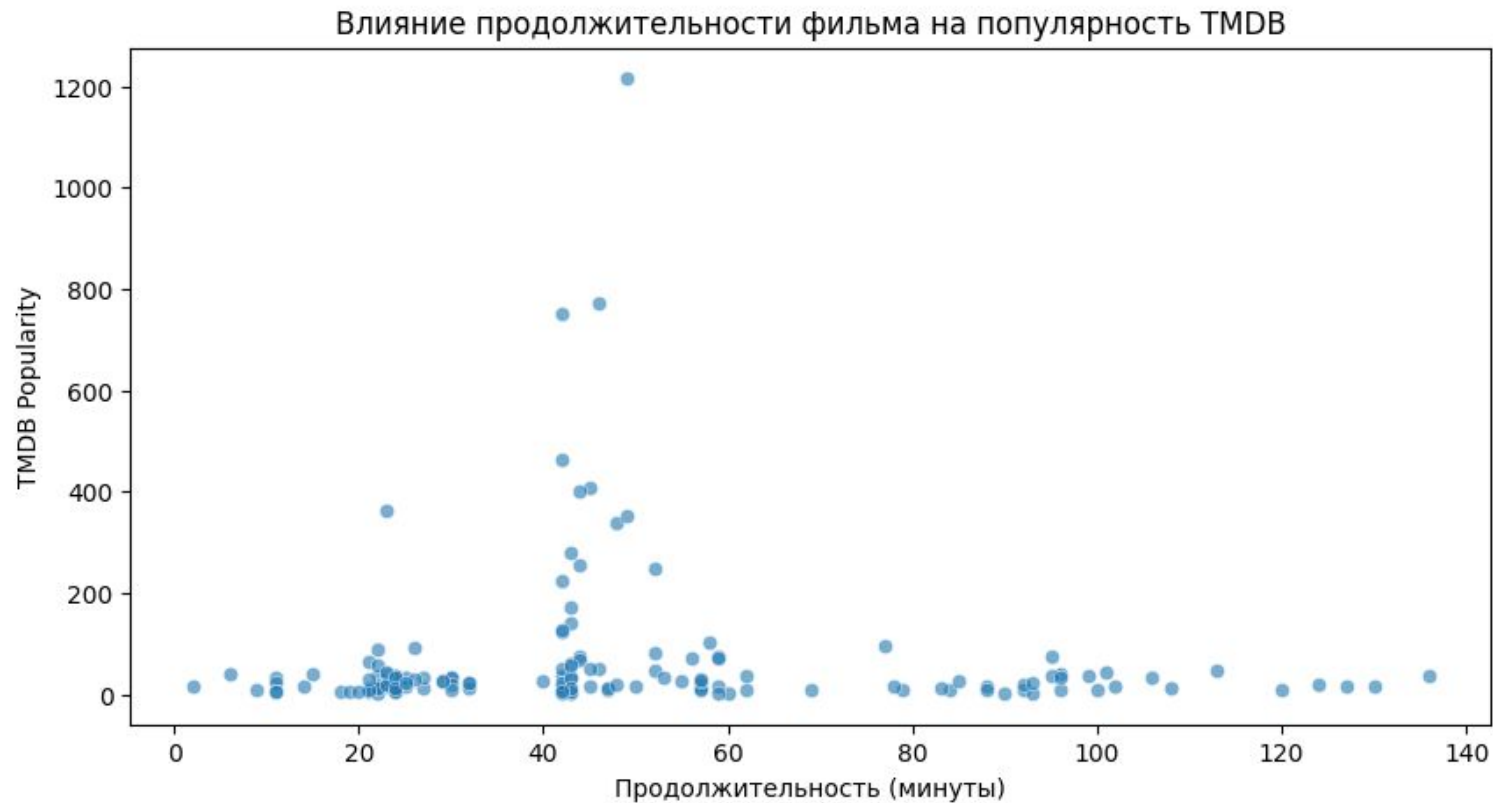
Если цель — гарантированный хит (из предыдущего графика) с высокими оценками (текущий график),
стоит инвестировать в триллеры и криминальные фильмы.
Это отличный выбор для массового успеха и высоких рейтингов.



Документальные (documentation) и военные (war) имеют большую продолжительность.
Комедии, ужасы и боевики обычно короче.



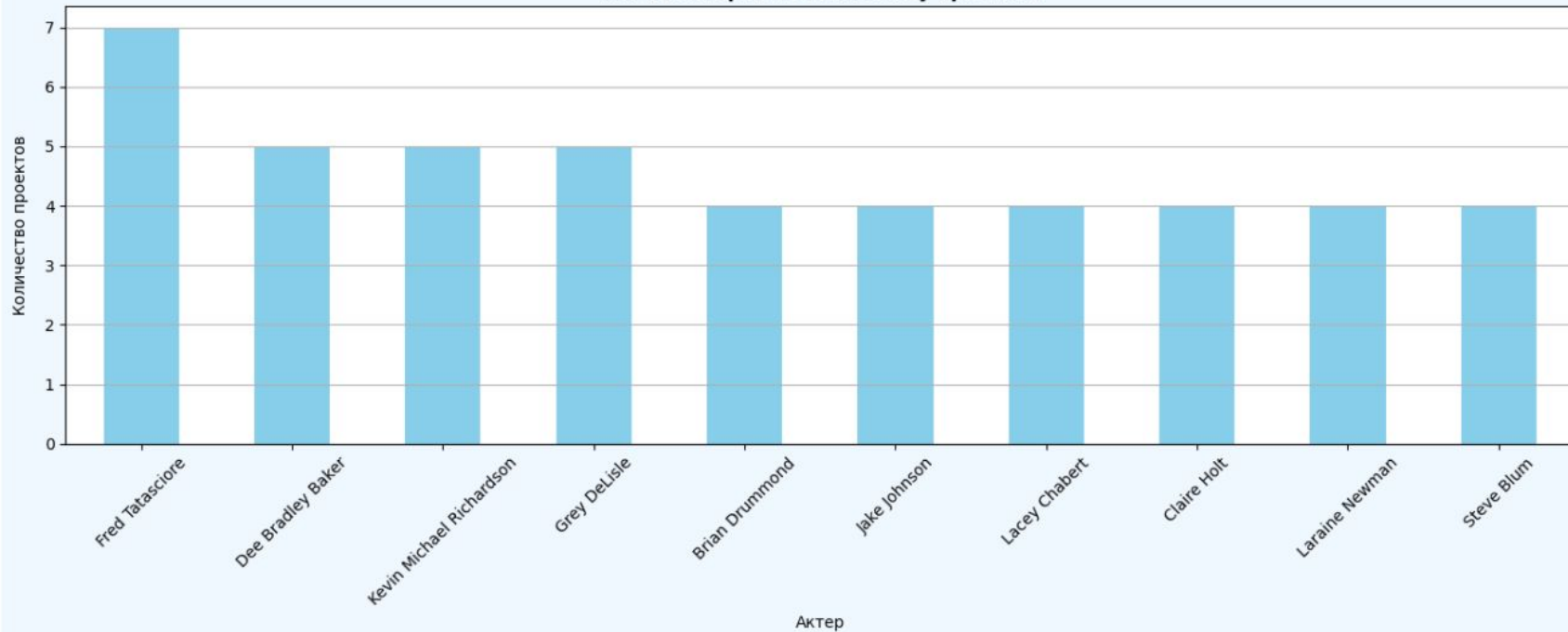
Длина фильма **слабо влияет** на IMDb-оценку (корреляция -0.22)

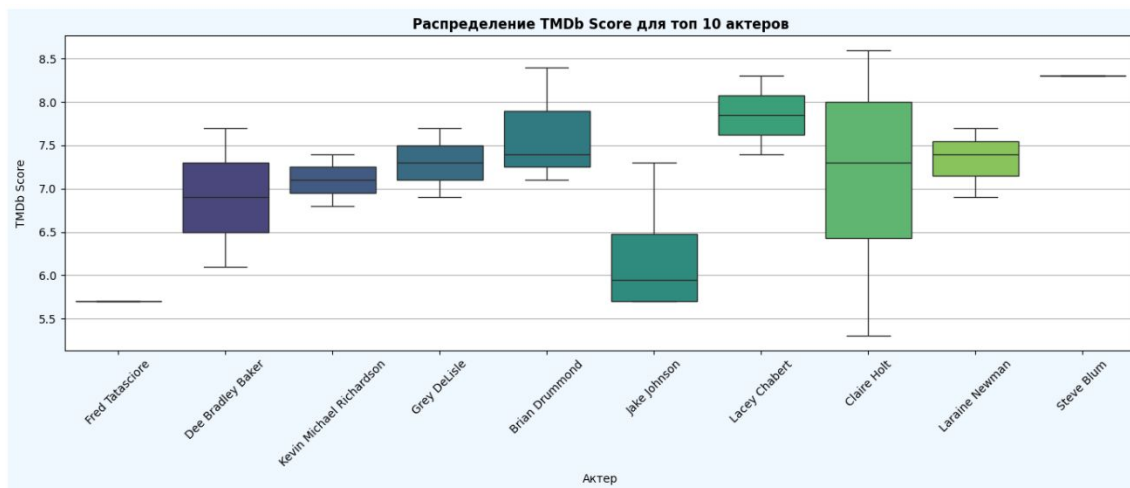
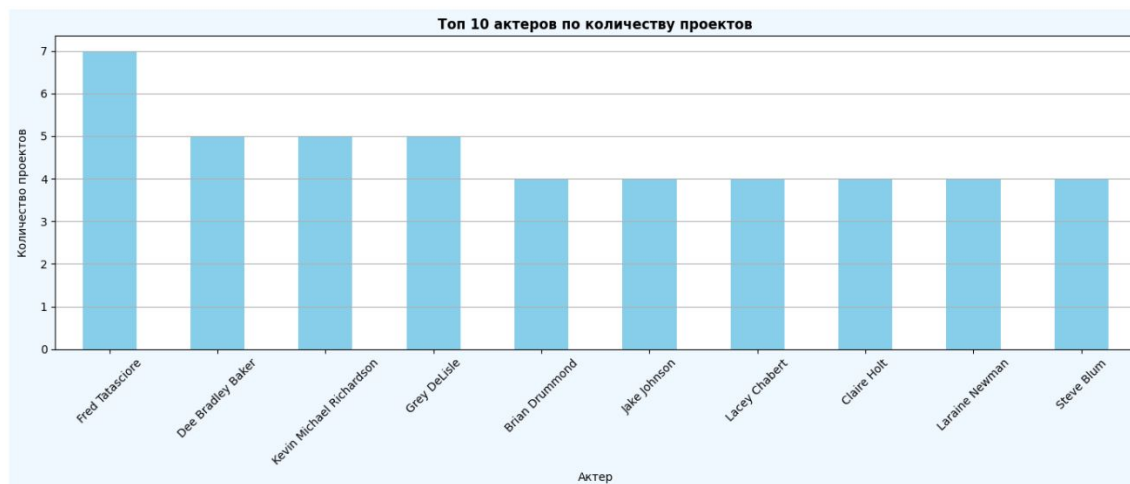


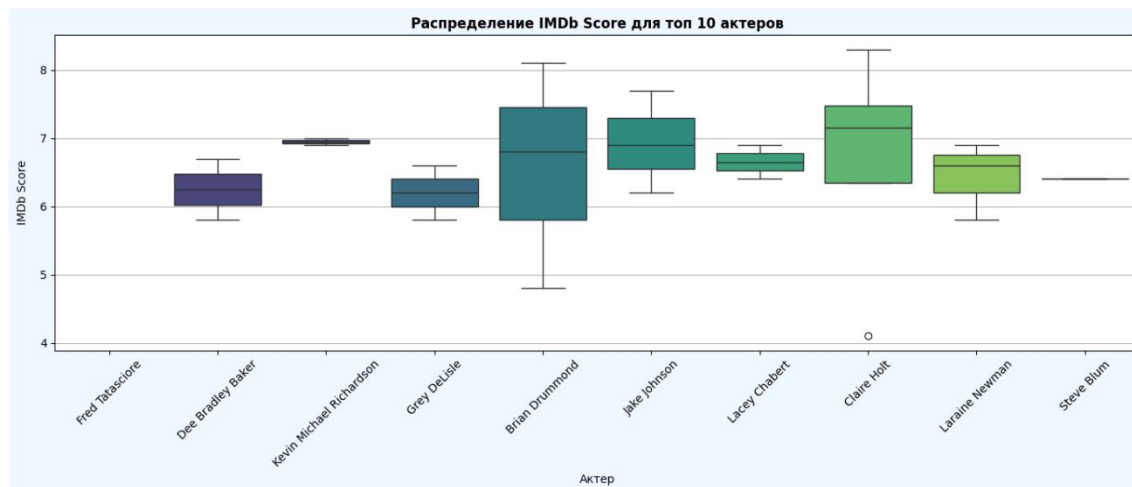
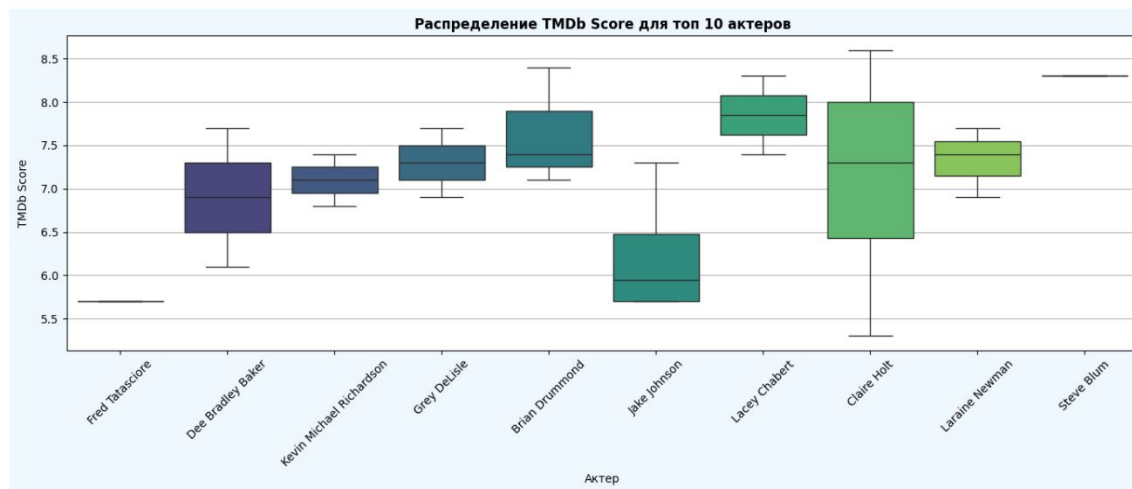
Длина фильма **практически не влияет** на популярность TMDB (корреляция -0.045)

Исследование актерских составов

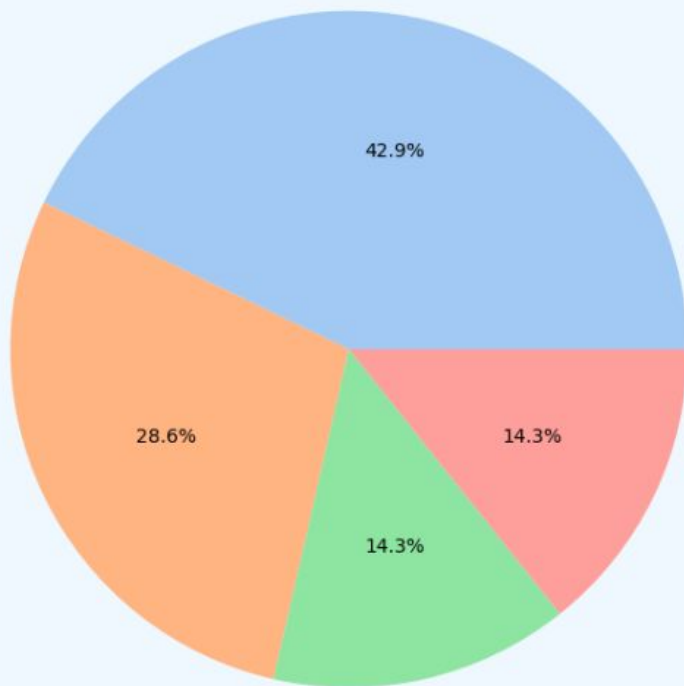
Топ 10 актеров по количеству проектов







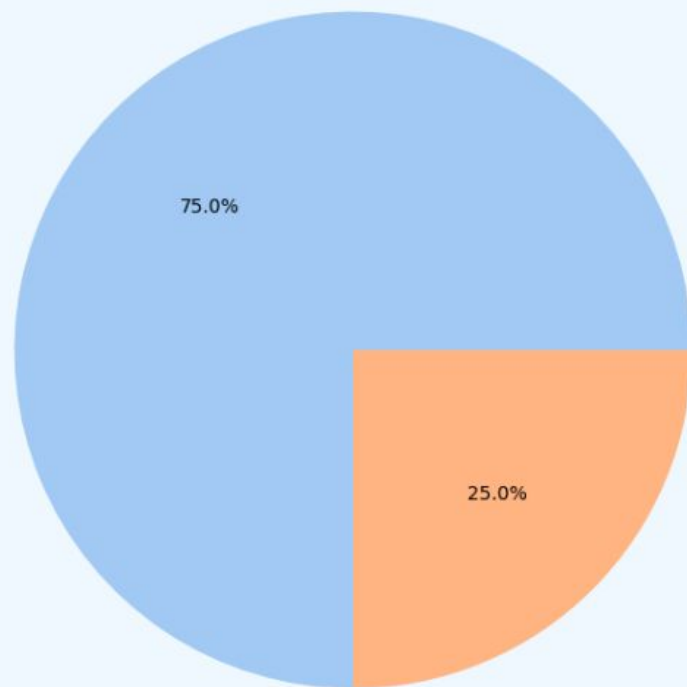
Распределение rating_description для Fred Tatasciore



Rating Descriptions

- Parental guidance suggested. May not be suitable for children.
- Suitable for children ages 7 and older. Content may be mild.
- Suitable for all ages.
- Suitable for children ages 7 and older.

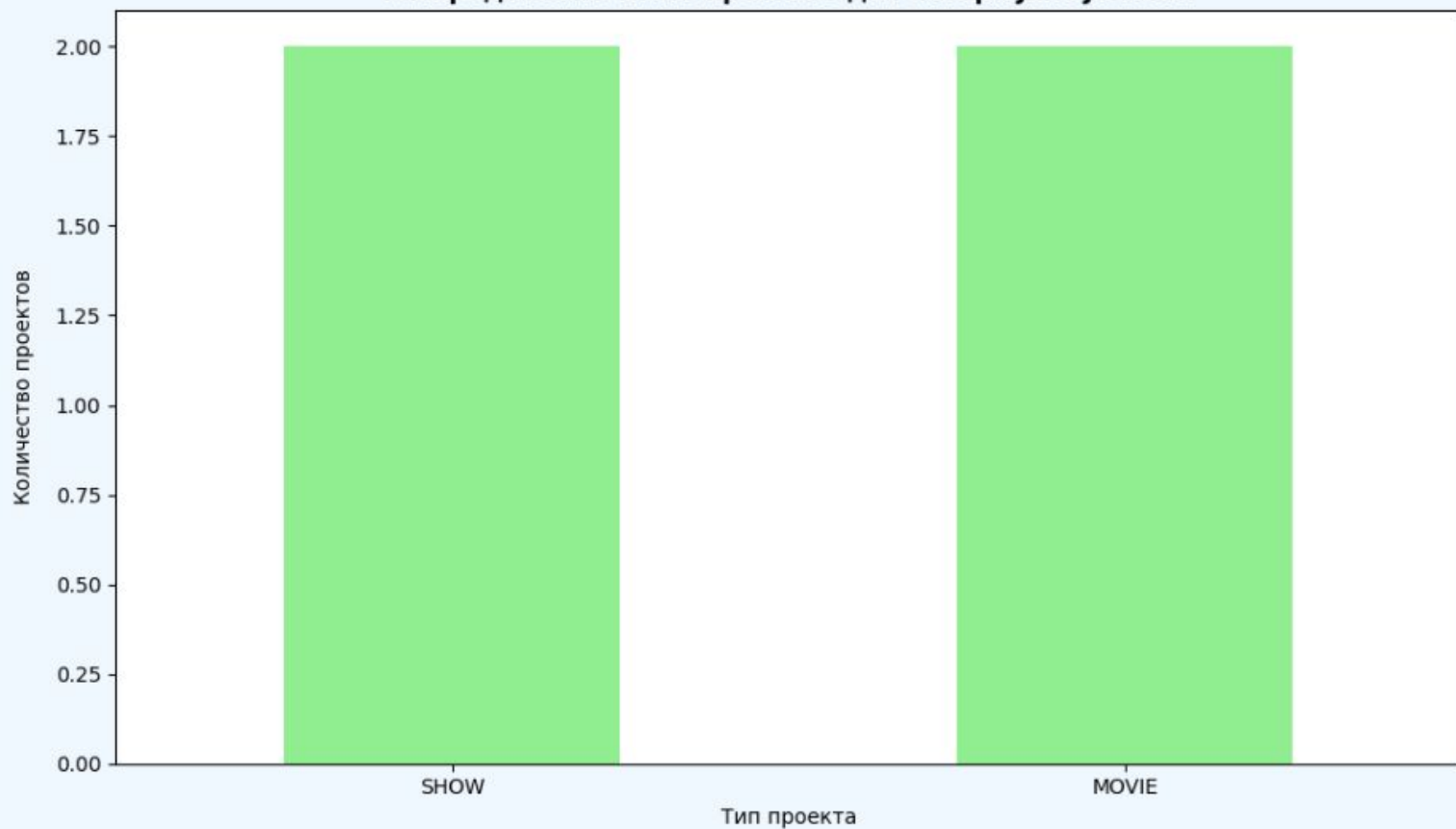
Распределение rating_description для Laraine Newman



Rating Descriptions

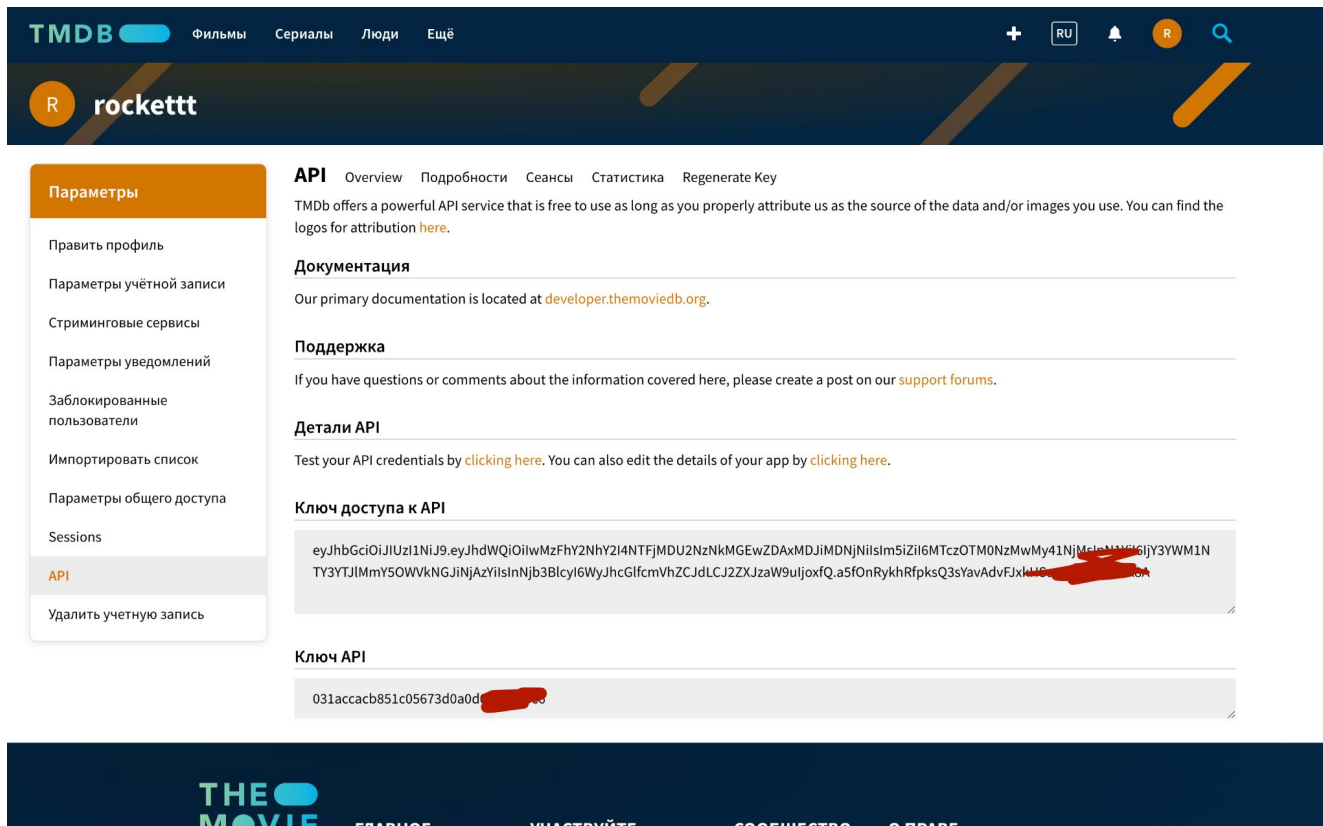
- Parental guidance suggested. May not be suitable for children.
- Suitable for children ages 7 and older. Content may be mild.

Распределение типов проектов для актера Jake Johnson



Применения средств feature engineering

Загружаем данные с помощью TMDB API



Получаем budget и revenue

Считаем roi

	imdb_score	imdb_votes	tmdb_popularity	tmdb_score	budget	\
12	4.5	83519.0	43.924	4.2	40000000.0	
49	6.6	7304.0	38.145	7.5	13938145.0	
78	5.9	9458.0	8.106	6.5	9900000.0	
191	7.5	255099.0	19.789	7.4	53000000.0	
202	6.6	100705.0	37.237	6.8	135000000.0	

	revenue	roi
12	0.0	-1.000000
49	30169000.0	1.164492
78	23600000.0	1.383838
191	83300000.0	0.571698
202	368871007.0	1.732378

```
df['roi'] = (df['revenue'] - df['budget']) / df['budget']
```

Смотрим корреляцию и делаем вывод что:

Корреляционная матрица:

	roi	user_rating_score	imdb_score	tmdb_score
roi	1.000000	-0.311258	0.260962	0.173652
user_rating_score	-0.311258	1.000000	-0.103557	-0.041828
imdb_score	0.260962	-0.103557	1.000000	0.919951
tmdb_score	0.173652	-0.041828	0.919951	1.000000

Попробуем проверить гипотезу,
известный актер = более высокая
оценка?

The Oscar Award, 1927 - 2024

161

New Notebook

Download

Data Card

Code (14)

Discussion (4)

Suggestions (0)

Please, If you enjoyed this dataset, don't forget to upvote it.

Tags

Movies and TV Shows

Context

The Academy Awards, also officially and popularly known as the Oscars, are awards for artistic and technical merit in the film industry. Given annually by the Academy of Motion Picture Arts and Sciences (AMPAS), the awards are an international recognition of excellence in cinematic achievements as assessed by the Academy's voting membership. The various category winners are awarded a copy of a golden statuette, officially called the "Academy Award of Merit", although more commonly referred to by its nickname "Oscar". The statuette depicts a knight rendered in Art Deco style.

Content

This file contains a scrape of The Academy Awards Database, recorded of past Academy Award winners and nominees between 1927 and

View more

the_oscar_award.csv (970.36 kB)

Download Icon Full Screen Icon Share Icon

Detail

Compact

Column

7 of 7 columns

About this file

Suggest Edits

This file contains all of nomination on the Academy Awards, Oscar.

# year_film	# year_ceremony	# ceremony	Δ category	Δ name	Δ film
Filter screened at	Ceremony happened at	Number of ceremony	Nomination	Name of nomination	File's title

Data Explorer

Version 11 (970.36 kB)

the_oscar_award.csv

Summary

1 file

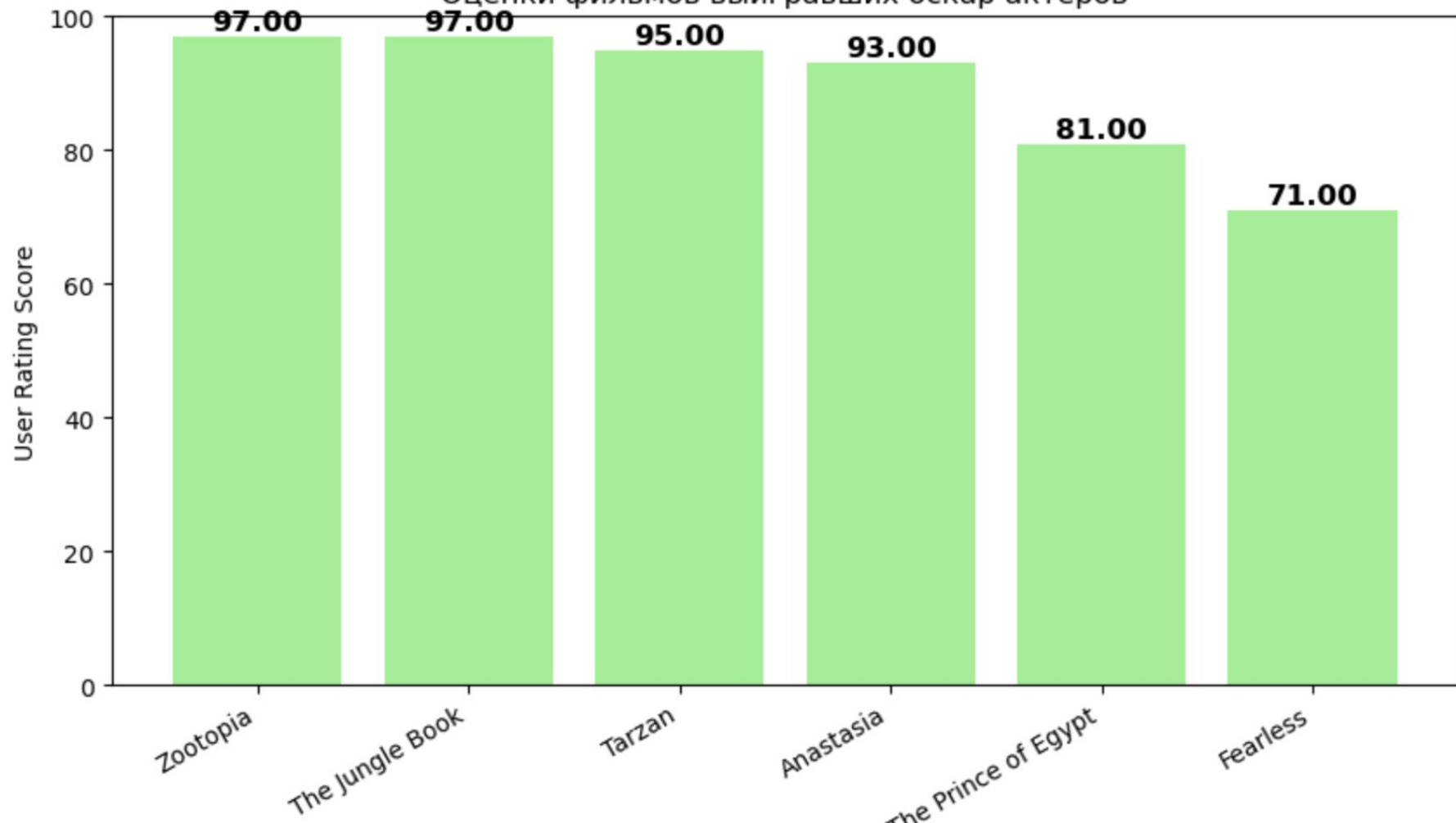
7 columns

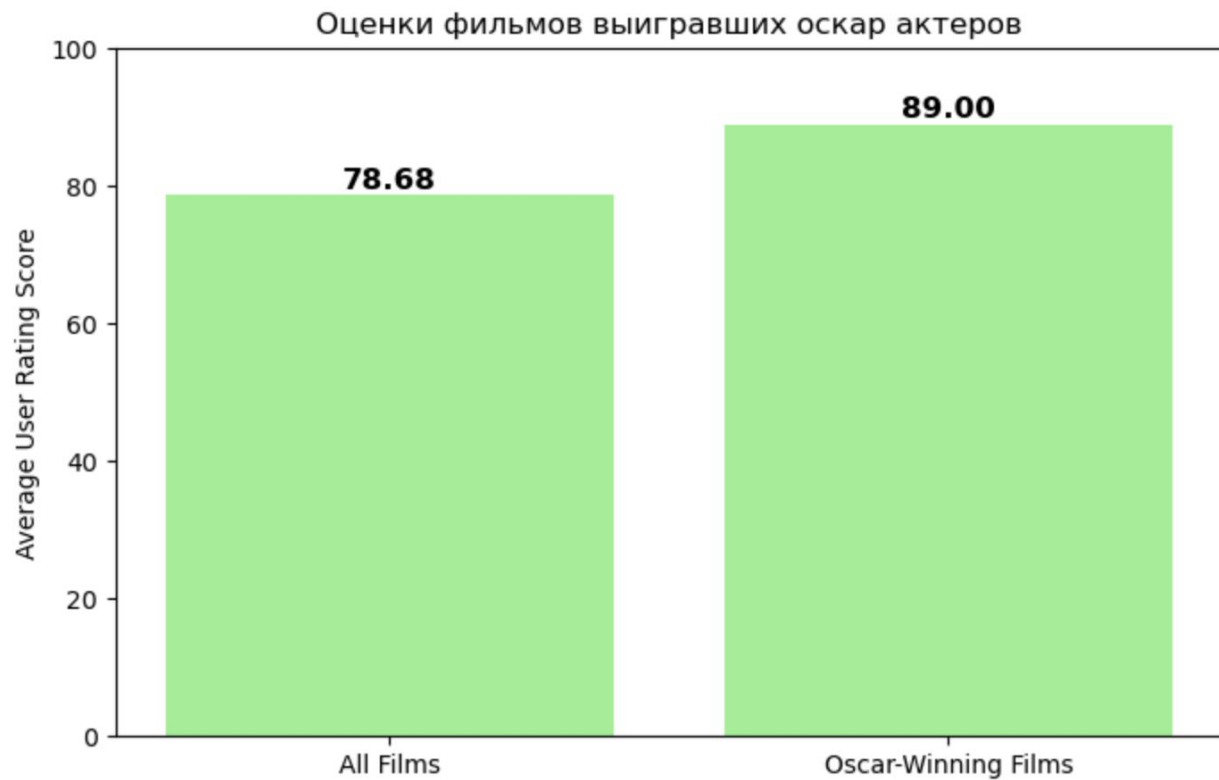
```
df_2 = pd.read_csv('the_oscar_award.csv')
print(df_2.head(5))
```

	year_film	year_ceremony	ceremony	category	name \
0	1927	1928	1	ACTOR	Richard Barthelmess
1	1927	1928	1	ACTOR	Emil Jannings
2	1927	1928	1	ACTRESS	Louise Dresser
3	1927	1928	1	ACTRESS	Janet Gaynor
4	1927	1928	1	ACTRESS	Gloria Swanson

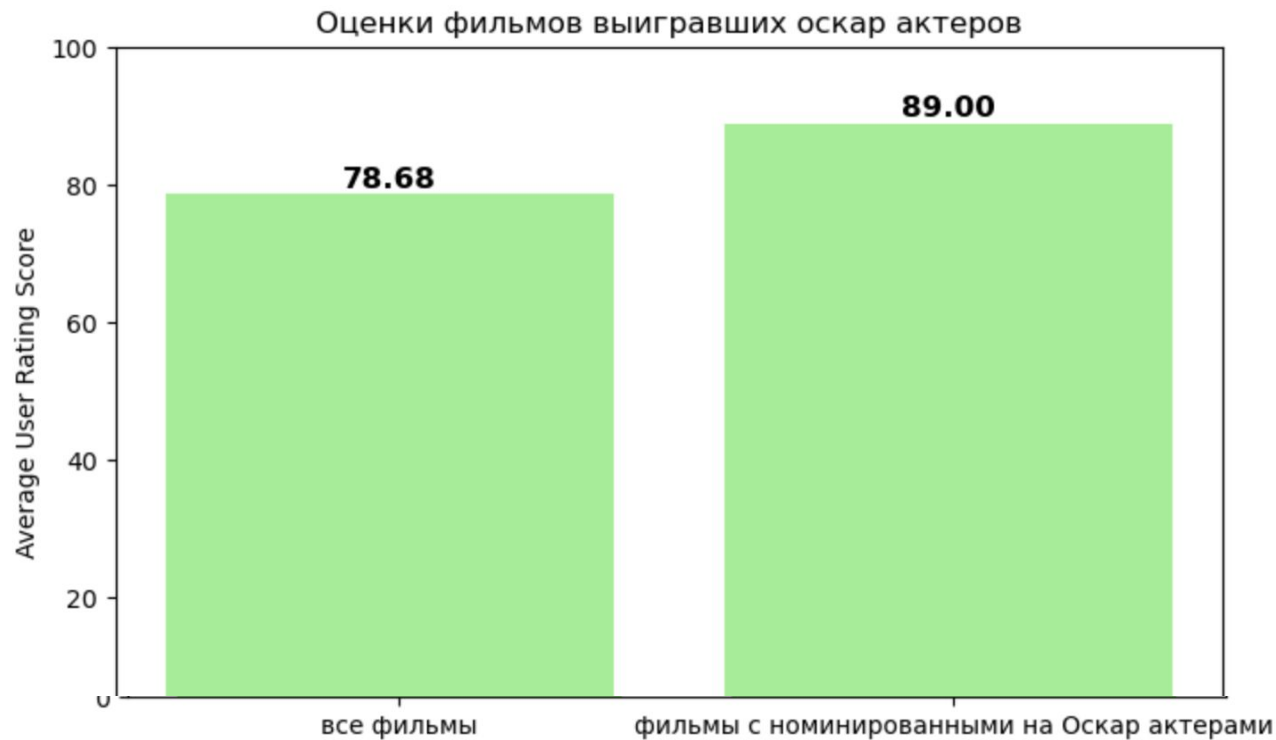
	film	winner
0	The Noose	False
1	The Last Command	True
2	A Ship Comes In	False
3	7th Heaven	True
4	Sadie Thompson	False

Оценки фильмов выигравших оскар актеров

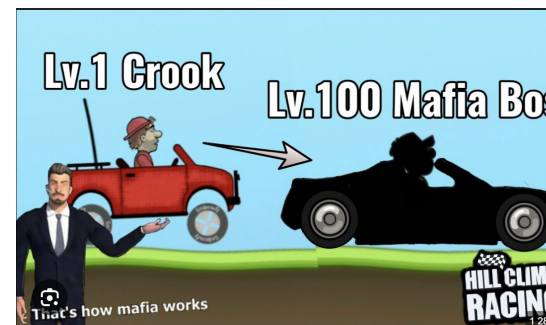


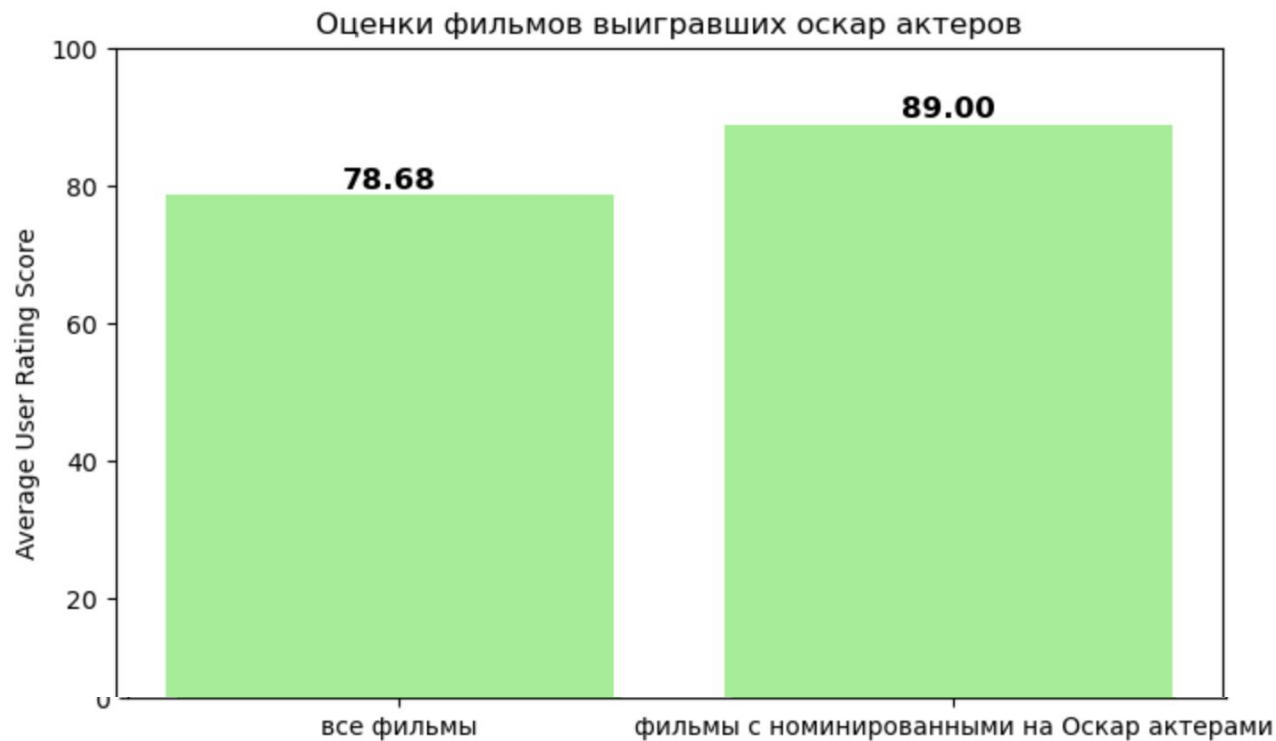


ВЫВОД?



Вывод?





Вывод?

p-value = 0.243

Посмотрим фильмы с людьми
которых хотя бы просто
номинировали





ВЫВОД?



Вывод?

p-value = 0.037

Значит появляется новая фича

есть ли номинированный актер в фильме: True False