NETFLIX
1000 Shows

Разведочный Анализ Данных

Курсовой проект

Андреев Данила Исламов Айрат Логвинюк Михаил

Масимова Нигяр Рубанов Владислав

ФКН ВШЭ МНОД, 2025

Краткое описание данных

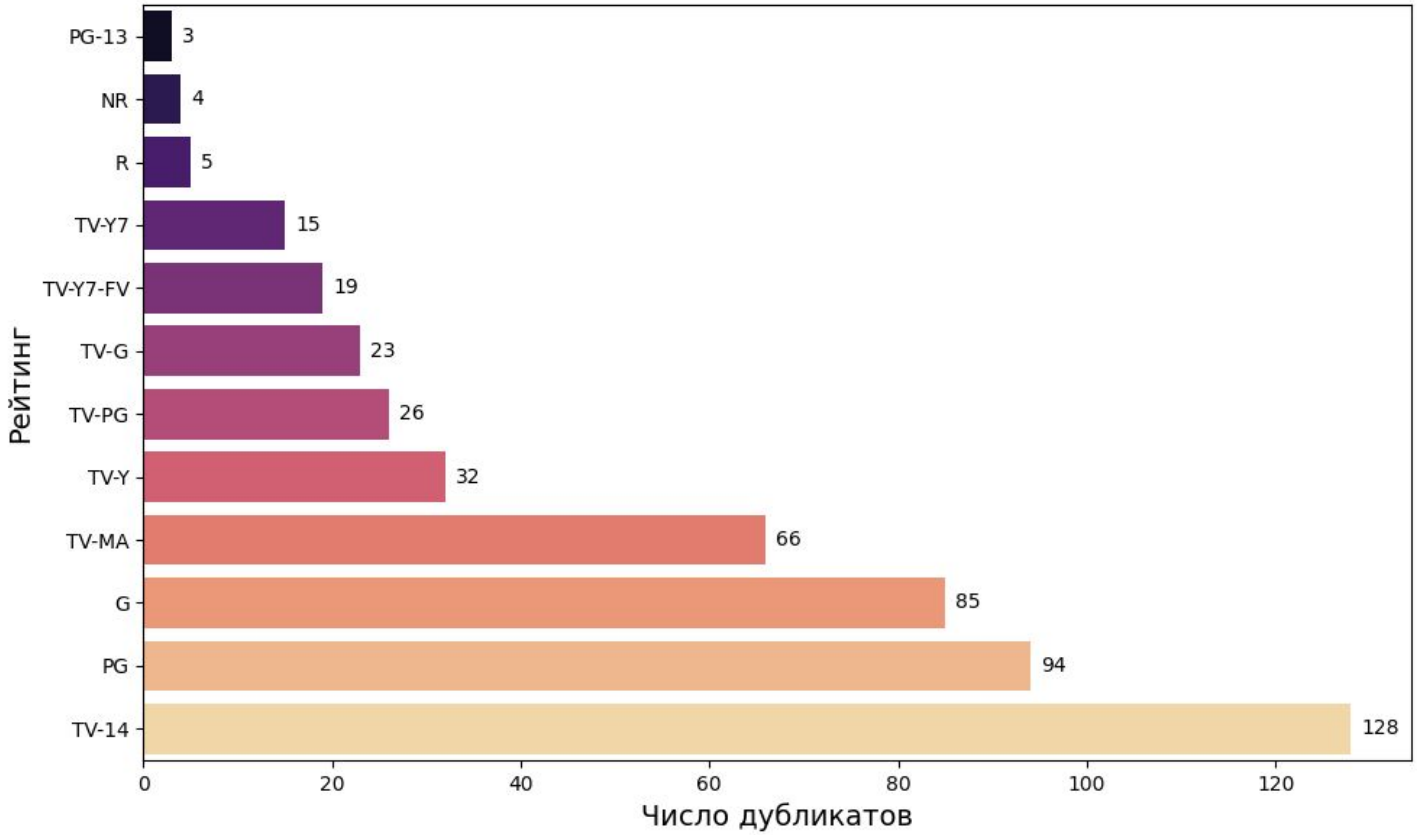
Набор данных **«1000 Netflix Shows»** (собранный по состоянию на 11.06.2017) :

Признак	Описание
title	Название шоу.
rating	Рейтинговая группа, например, **G** (подходит для всех возрастов), **PG** (просмотр с родителями), **TV-14** (для подростков от 14 лет и старше), **TV-MA** (только для взрослых).
rating_level	Описание рейтинговой группы и особенностей показа.
release_year	Год выхода шоу.
user_rating_score	Оценка пользователя (0-100).

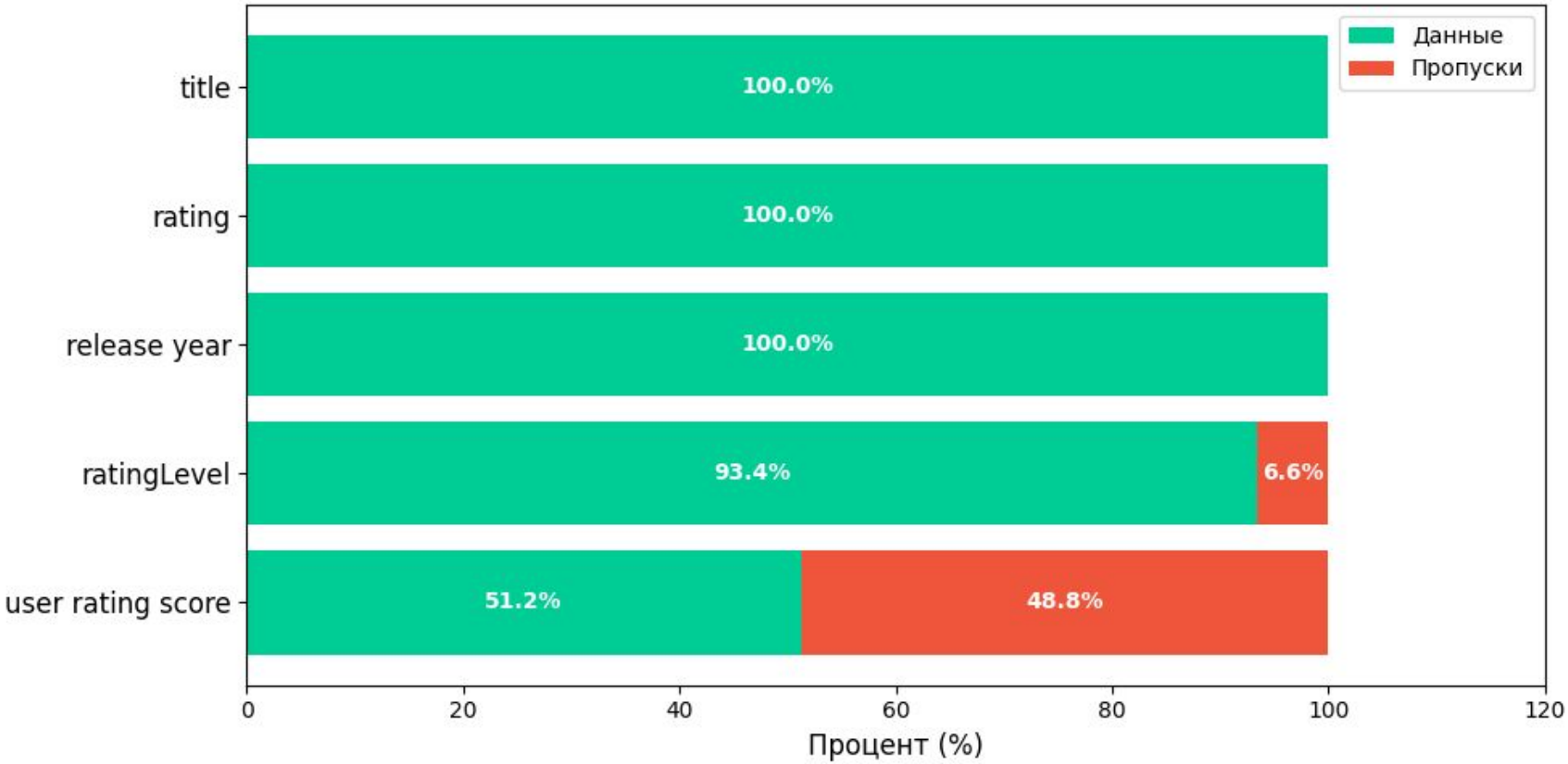
Качество данных

Всего: 1,000 строк

Дубликаты по рейтинговым группам



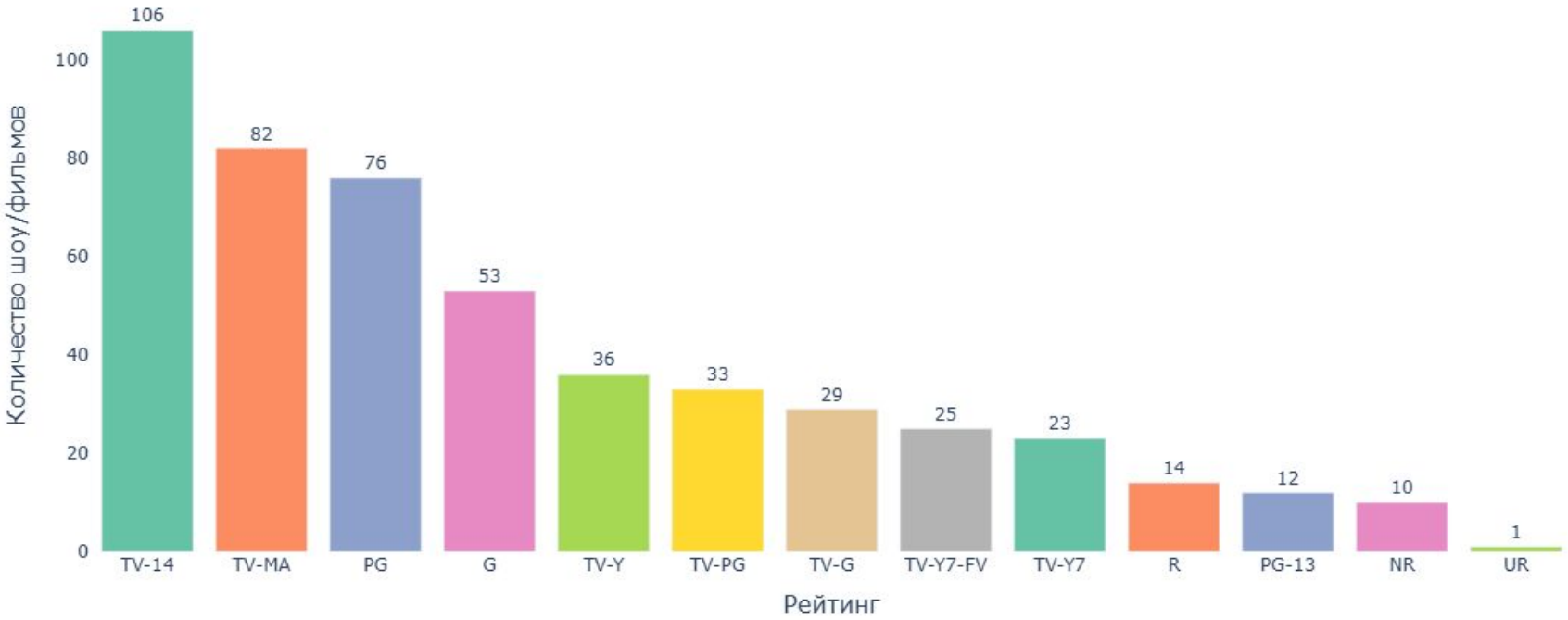
Пропуски по столбцам



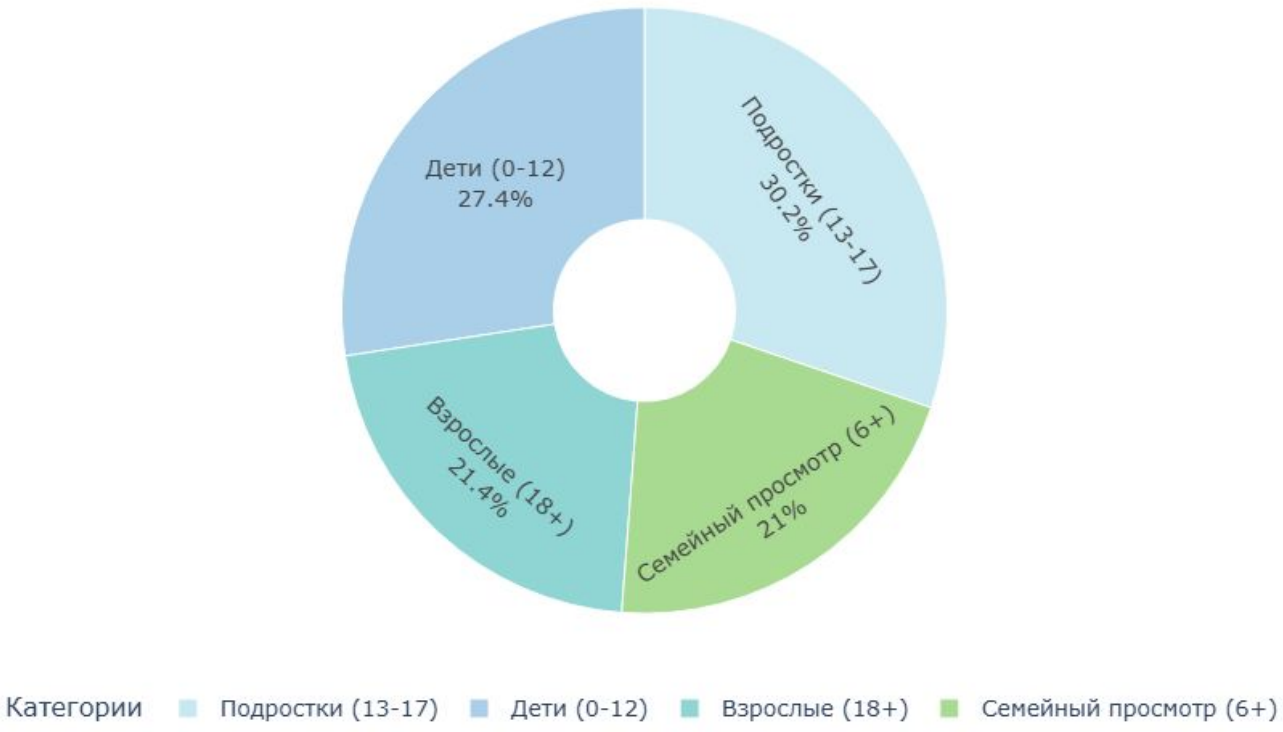
Описание рейтинговых групп

Категория	Рейтинг	Описание
Телевидение	**TV-Y**	Для детей всех возрастов. Контент не содержит ничего страшного.
	TV-Y7	Для детей от 7 лет и старше. Может содержать легкие пугающие сцены.
	TV-Y7-FV	Для детей от 7 лет и старше, с элементами насилия в мультфильмах.
	TV-G	Общедоступный рейтинг. Контент подходит для всей семьи, без ограничений.
	TV-PG	Рекомендуется родительский контроль. Может содержать ненормативную лексику, некоторые интимные ситуации или умеренное насилие.
	TV-14	Не рекомендуется детям младше 14 лет. Может содержать более сильные сцены насилия, сексуального характера или ненормативную лексику.
	TV-MA	Только для взрослых. Контент предназначен для лиц 17+ лет, может содержать откровенный секс, насилие или грубую лексику.
Кинотеатры (MPAA)	**G**	Общедоступный рейтинг. Подходит для всех возрастов. Нет ни насилия, ни неприемлемого контента.
	PG	Рекомендуется родительский контроль. Возможны легкие сцены насилия, мягкие ругательства.
	PG-13	Не рекомендуется детям младше 13 лет. Может содержать сцены насилия, интимные ситуации или ненормативную лексику.
	R	Ограниченный просмотр. Лица младше 17 лет допускаются только с родителями. Может содержать сильный насильственный или сексуальный контент.
Без рейтинга	**NR**	Фильм или программа не прошли официальную рейтинговую систему.
	UR	Фильм не был представлен на рассмотрение рейтинговой комиссии для получения официального рейтинга или это режиссерская версия фильма.

Рейтинговые группы

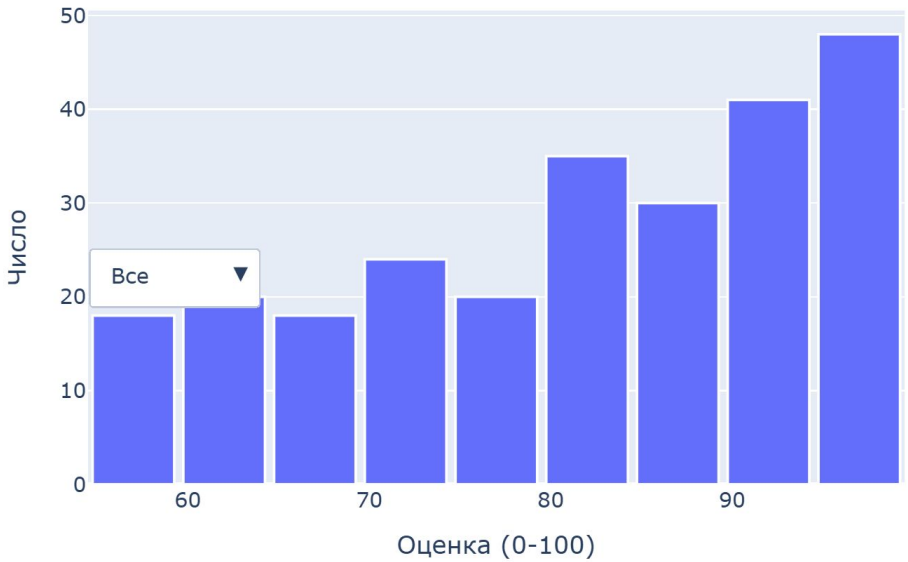


Целевая аудитория

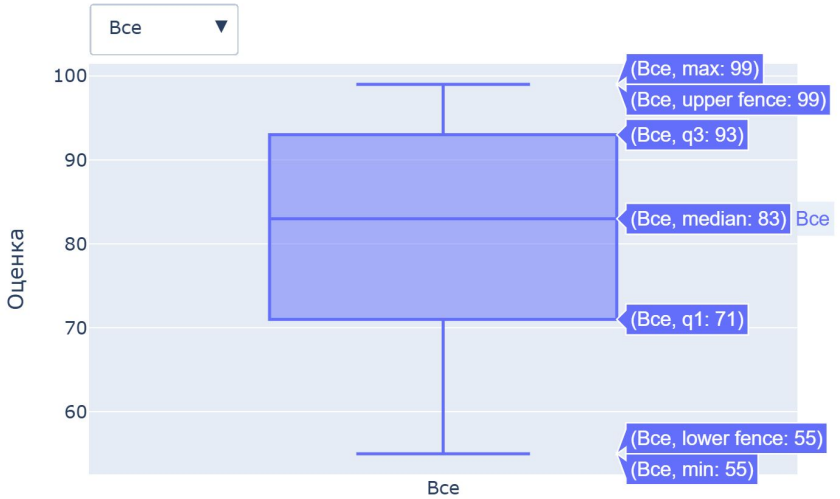


Оценки зрителей

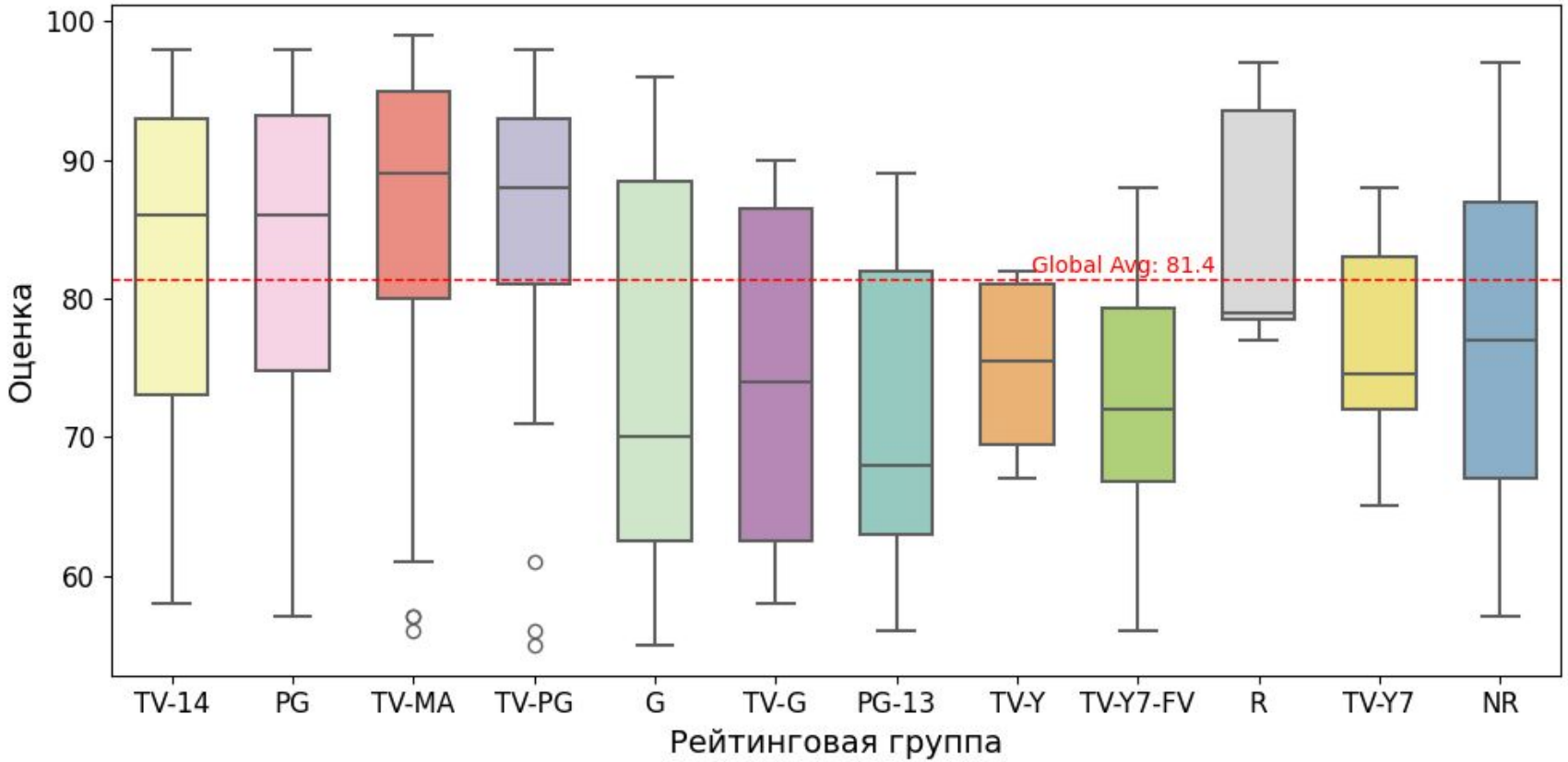
Распределение оценок зрителей



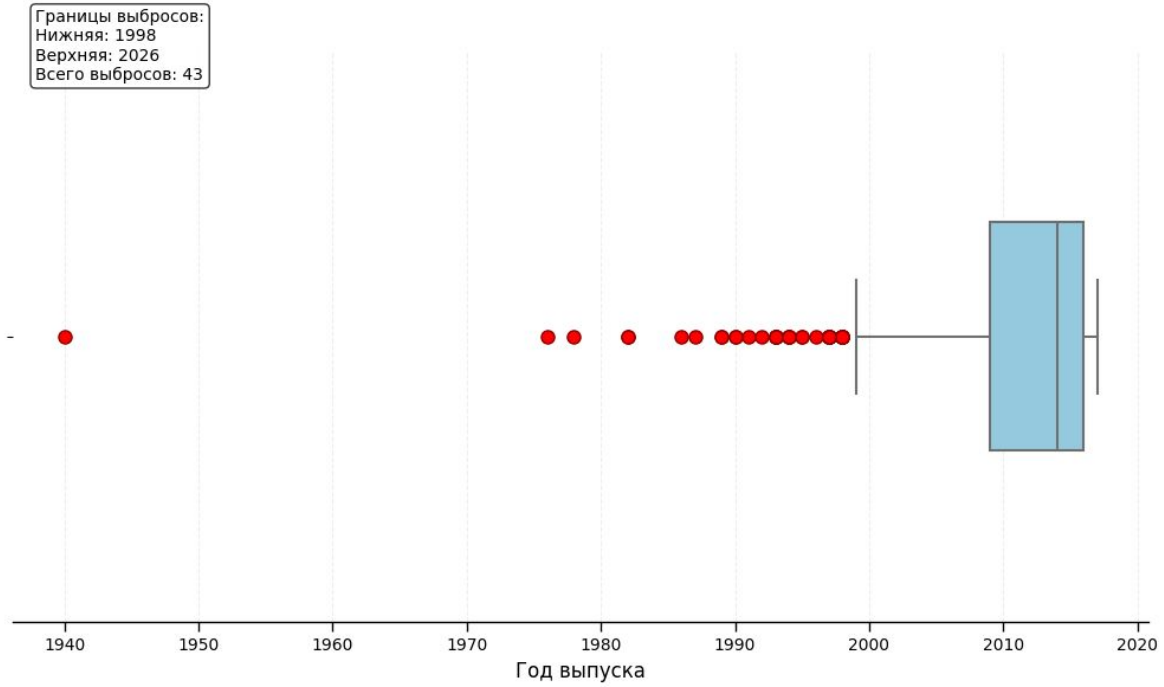
Boxplot для оценок зрителей



Оценки зрителей по рейтинговым группам



Год выхода шоу



Есть фильмы, выпущенные **до запуска Netflix** (1997 г.), большинство из них классифицируются как выбросы

Количество фильмов по году выпуска

- Данные “обрываются” в 2017 г.
- **Смена тренда** может быть связана со сменой стратегии Netflix

Успешность Netflix: 2016 vs 2017?

- 84.3 (2016) vs **88.1 (2017)**
- Важно иметь в виду что **2017 г. не завершился** в этой картине мира
- Для выводов нужно сравнивать другие бизнес метрики:
  - доход
  - число часов на платформе
  - число подписчиков
  - число просмотров

Обогащение данных – датасеты

Netflix popular movies dataset (kaggle) – 9957 стр. / 9 кол.

- **movies title**
- **cast of the movie**
- desc of movies
- duration
- **rating on IMDB**
- voted by people
- **year**
- **genre**
- certificate

Предобработка:

- Удаление дублей
- Приведение колонок к нужному формату
- Парсинг json полей

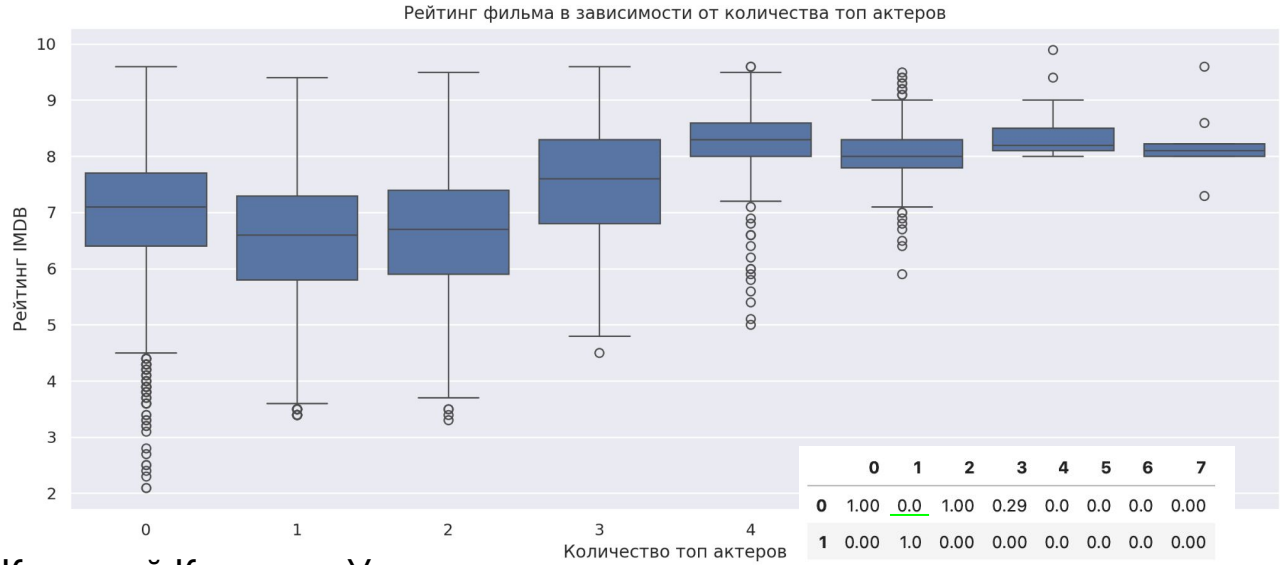
Movie Dataset: Budgets, Genres, Insights (kaggle) – 4803 стр. / 24 кол.

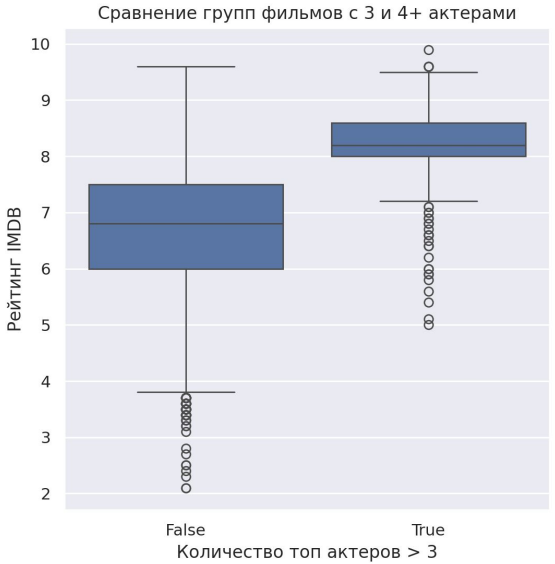
- **budget**
- genres
- homepage
- original_language
- original_title
- popularity
- production_companies
- **production_countries**
- **release_date**
- **revenue**
- runtime
- spoken_languages
- **title**
- cast
- crew
- Director

И другие...

Влияние количества топ актеров на рейтинг

Из внешнего датасета Netflix было выбрано 1106 (топ 5%) актеров из 22743





Критерий Краскела-Уоллиса:
H-статистика: 1170.26
P-значение: **0.000**

Post-hoc Dunn:

	0	1	2	3	4	5	6	7
0	1.00	0.0	1.00	0.29	0.0	0.0	0.0	0.00
1	0.00	1.0	0.00	0.00	0.0	0.0	0.0	0.00
2	1.00	0.0	1.00	0.21	0.0	0.0	0.0	0.00
3	0.29	0.0	0.21	1.00	0.0	0.0	0.0	0.11
4	0.00	0.0	0.00	0.00	1.0	1.0	1.0	1.00
5	0.00	0.0	0.00	0.00	1.0	1.0	1.0	1.00
6	0.00	0.0	0.00	0.00	1.0	1.0	1.0	1.00
7	0.00	0.0	0.00	0.11	1.0	1.0	1.0	1.00

Значимые различия

Нет различий

Фильмы с 4 топ актерами и более получают статистически более высокие оценки

Обогащение данных – процесс объединения

1. Точное соответствие названий
32% / 15%

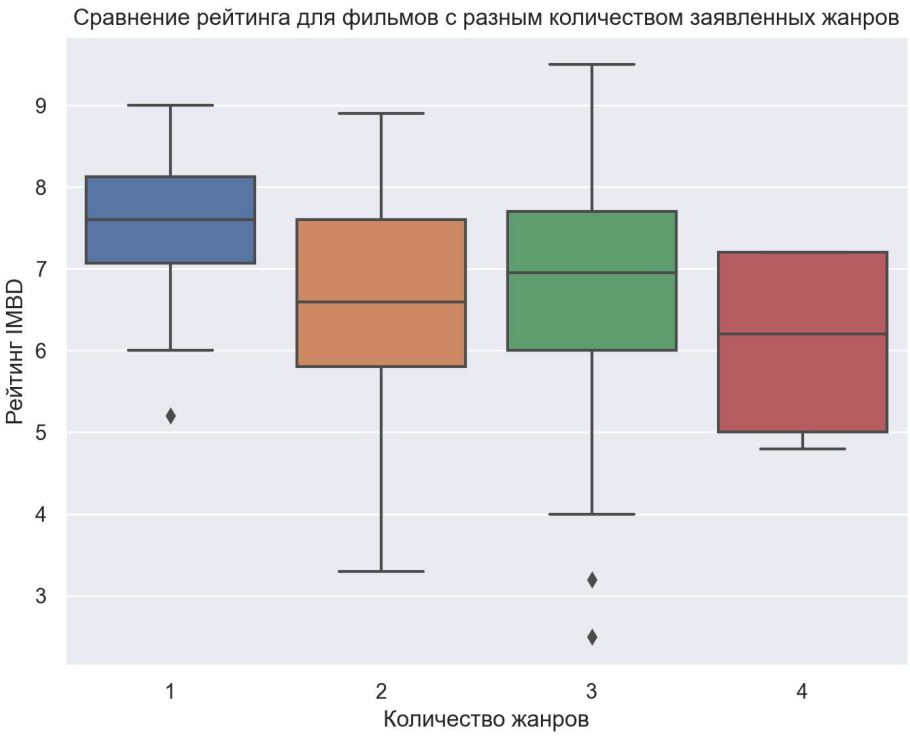
2. Неточные соответствия названий (локальные названия, артефакты, регистр, спецсимволы)
FuzzyWuzzy - fuzz.token_sort_ratio
37% / 16%

3. Поиск недостающей информации с помощью LLM с доступом в Интернет
Gemeni 2.5 Flash
(Выборочный контроль корректности выдачи)

Итоговая полнота используемых в дальнейшем полей, %:

genre	100.000000
rating_y	98.387097
stars	92.338710
budget	27.016129
production_countries	100.000000
revenue	26.209677

Исследование обогащенных данных
IMBD rating vs genre count

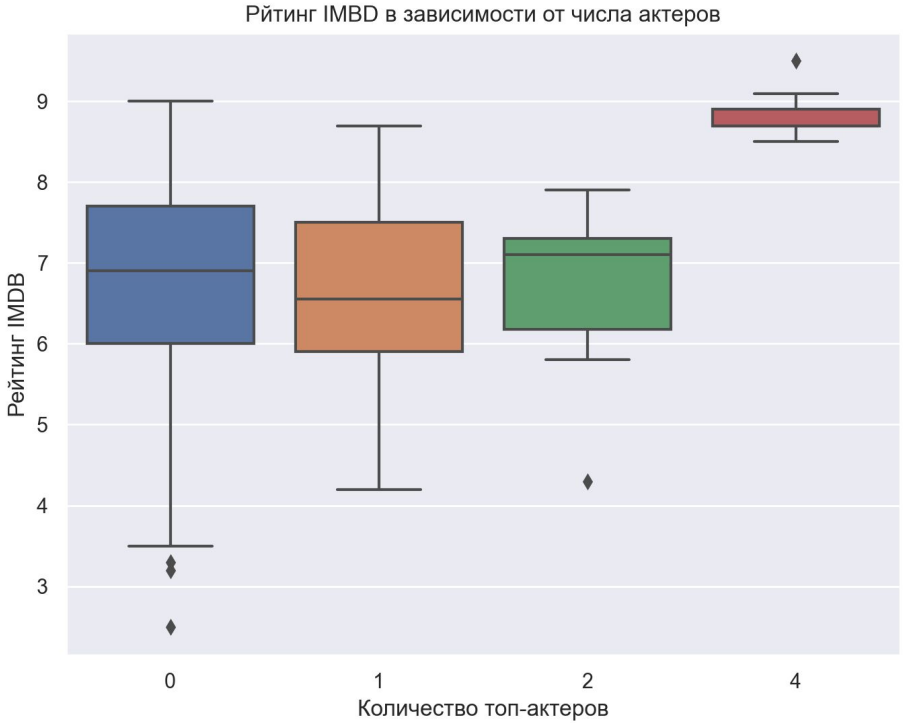


Критерий **Краскела-Уоллиса**:
H-статистика: 27.81
P-значение: **0.000**

Post-hoc Dunn:

	1	2	3	4
1	1.00	0.00	0.00	0.03
2	0.00	1.00	0.58	1.00
3	0.00	0.58	1.00	0.91
4	0.03	1.00	0.91	1.00

Исследование обогащенных данных
IMBD rating vs top stars count



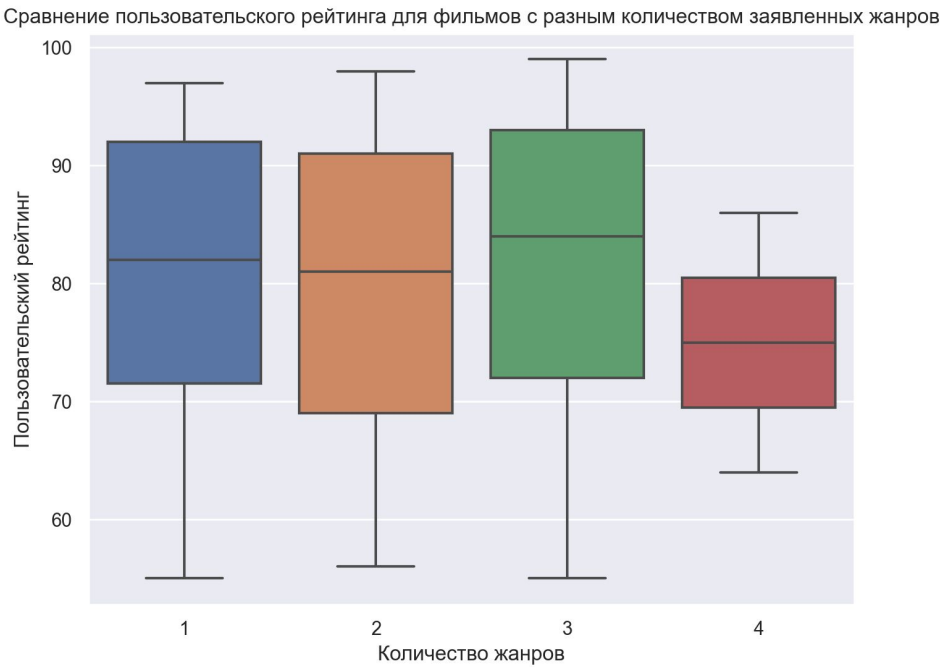
Коэффициент корреляции **Кендалла**: 0.069
p-значение: 0.061
Коэффициент корреляции **Спирмена**: 0.087
p-значение: 0.055

Критерий **Краскела-Уоллиса**:
H-статистика: 49.01
P-значение: **0.000**

Post-hoc Dunn:

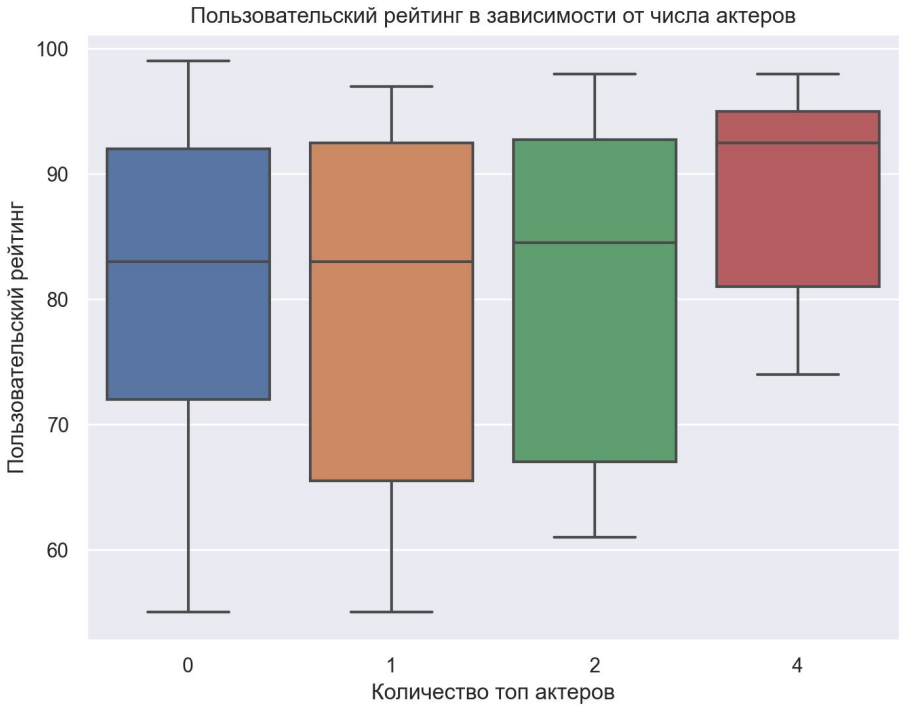
	0	1	2	4
0	1.0	1.0	1.0000	0.0000
1	1.0	1.0	1.0000	0.0000
2	1.0	1.0	1.0000	0.0002
4	0.0	0.0	0.0002	1.0000

Исследование обогащенных данных
user rating score vs genre count



Критерий **Краскела-Уоллиса**:
H-статистика: 1.63
P-значение: 0.65319

Исследование обогащенных данных
user rating score vs top stars count

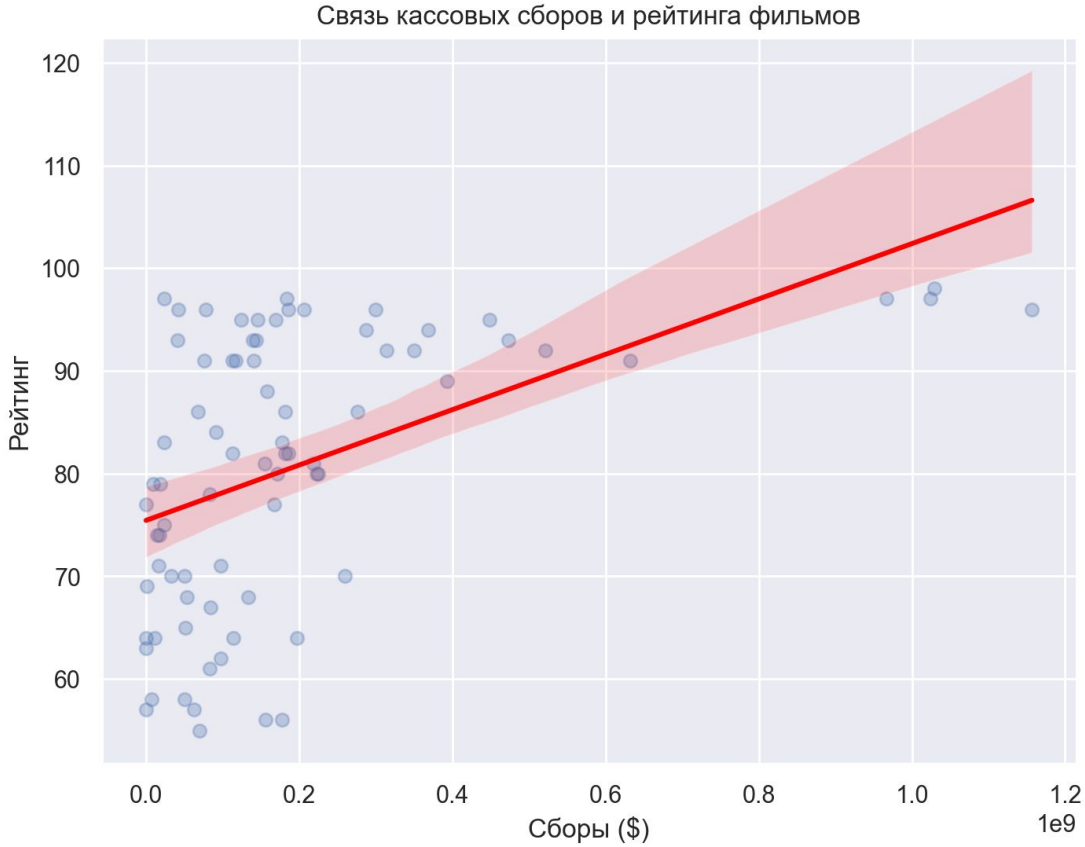


Коэффициент корреляции **Кендалла**: 0.034
p-значение: 0.507
Коэффициент корреляции **Спирмена**: 0.042
p-значение: 0.501

Критерий **Краскела-Уоллиса**:
H-статистика: 5.74
P-значение: 0.125

Исследование обогащенных данных

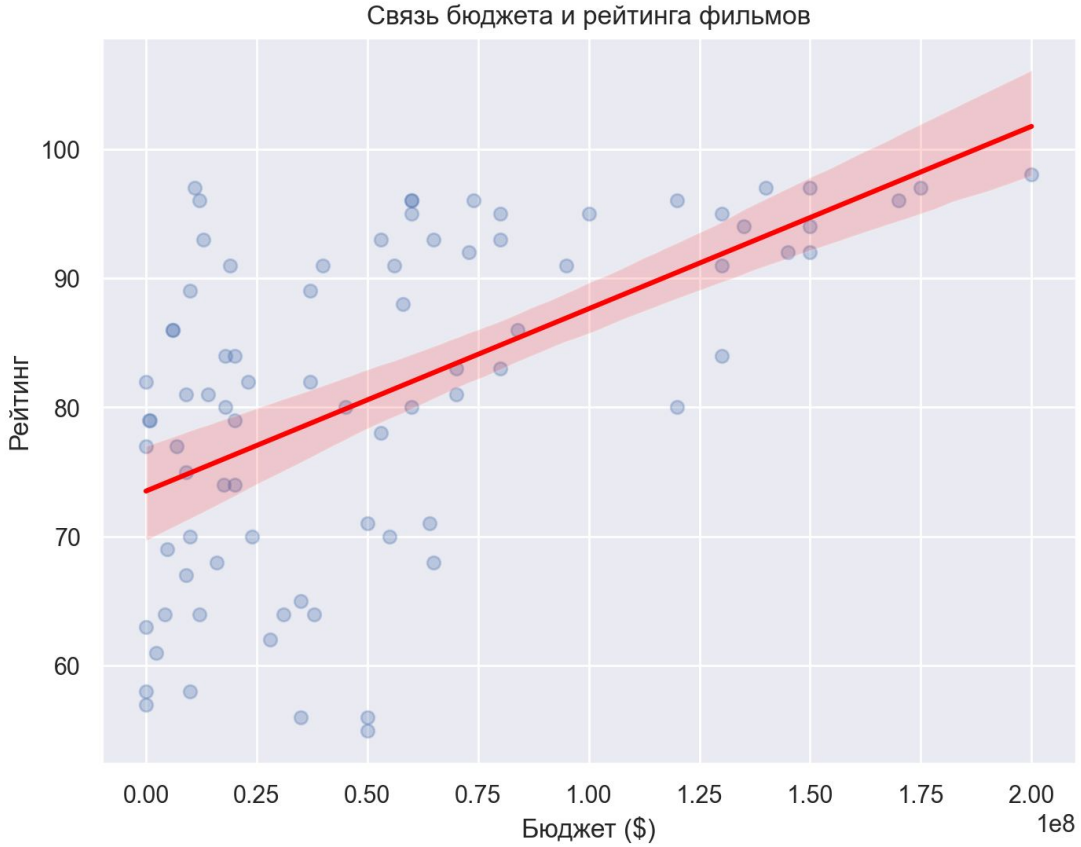
Зависимость пользовательской оценки и сборов



Коэффициент корреляции **Пирсона**: 0.484
p-значение: **0.000**

Исследование обогащенных данных

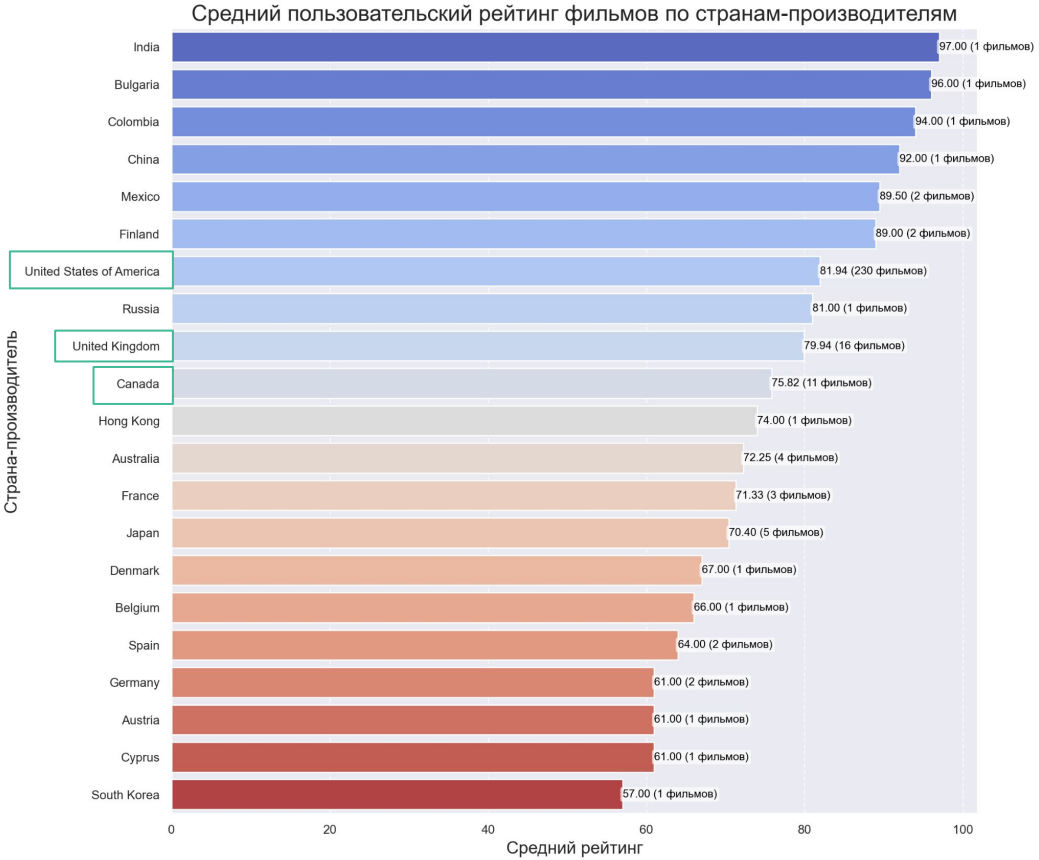
Зависимость пользовательской оценки и бюджета



Коэффициент корреляции **Пирсона**: 0.559
p-значение: **0.000**

Исследование обогащенных данных

Страны производства

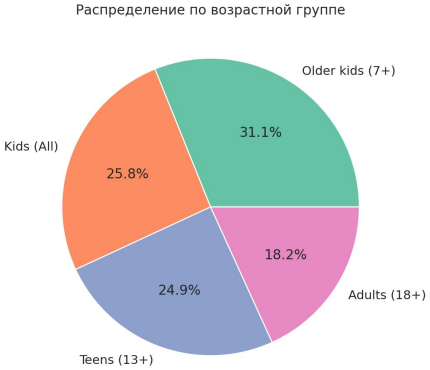


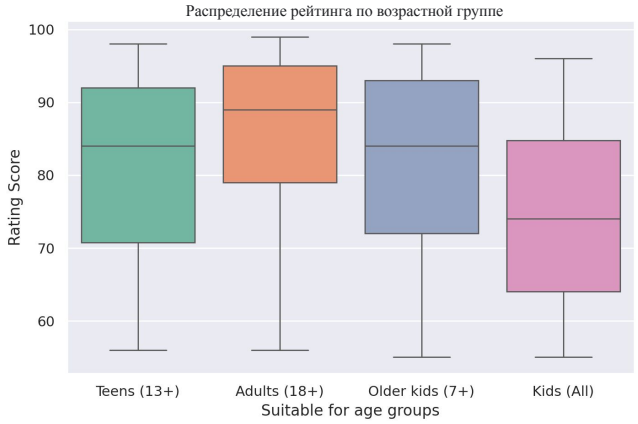
Основные страны-производители:
США, Великобритания, Канада

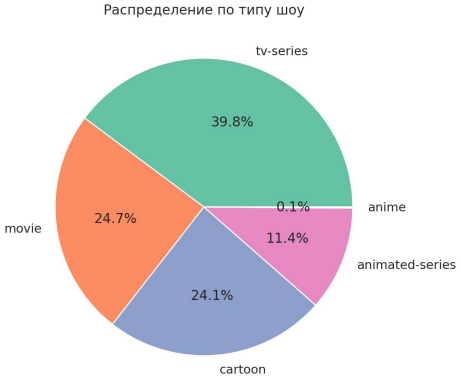
Топ (по рейтингу фильмов) страны имеют малое количество фильмов в прокате Netflix.

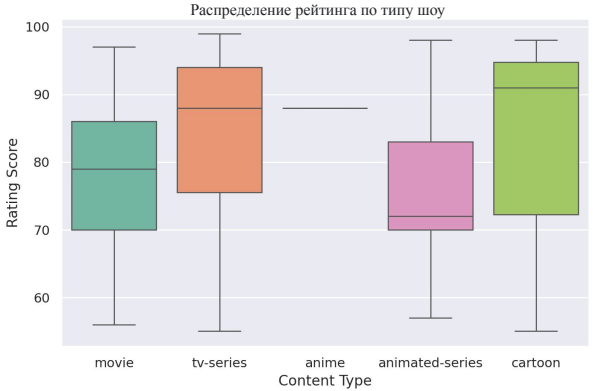
Внешний датасет kinopoisk

Были подгружены данные через API kinopoisk.dev для первичного датасета

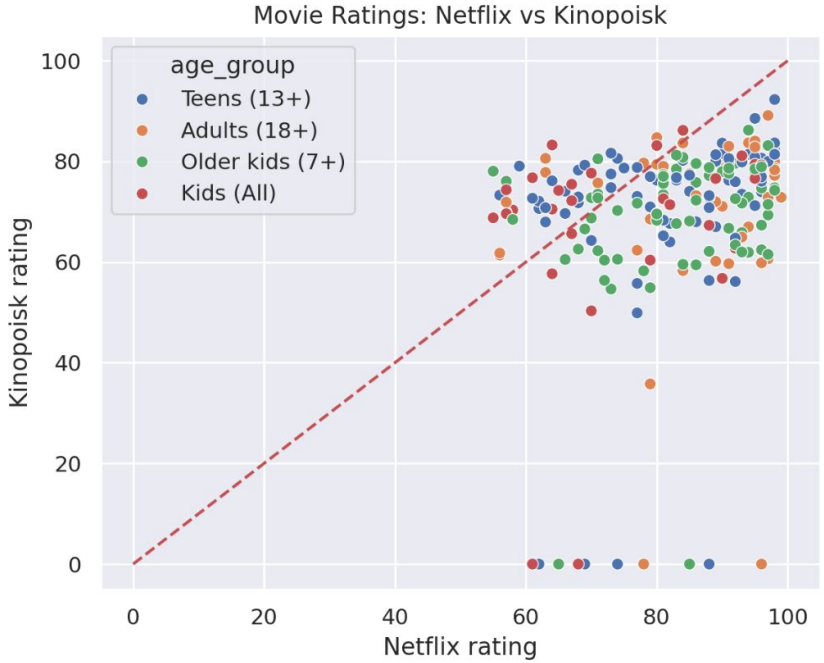


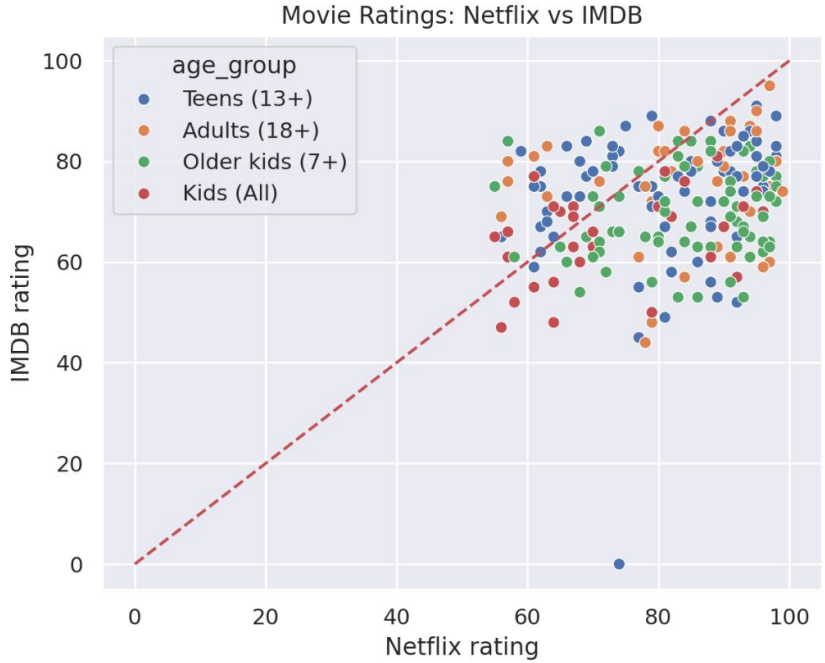






Внешний датасет kinopoisk





Основные инсайты

- **IMBD рейтинг** выше у фильмов отмеченных **одним жанром.**
- **IMBD рейтинг** выше у фильмов с 4 (и более) **топ-актерами.**
- Между **бюждетом фильма и пользовательской оценкой** есть монотонная положительная связь.
- Между **сборами фильма и пользовательской оценкой** есть монотонная положительная связь.
- Основные страны производства: **США, Великобритания, Канада.**
- Однако топовые по рейтингу фильмов страны имеют малое количество фильмов в прокате Netflix.
- Русскоязычные пользователи склонны ставить оценку в меньшем диапазоне, чем на **Netflix**
- Фильмы с возрастной категорией более 18 получают более высокие оценки

	title_x	title_y
0	LEGO: Marvel Super Heroes: Maximum Overload	Lego Marvel Super Heroes: Maximum Overload
1	Barbie Life in the Dreamhouse	Barbie: Life in the Dreamhouse
2	Tayo the Little Bus	Tayo, the Little Bus
3	Haters Back Off	Haters Back Off!
4	I.T	I.T.
5	LEGO Bionicle: The Journey to One	Lego Bionicle: The Journey to One
6	MerlÌ?	Merlí
7	Jane The Virgin	Jane the Virgin
8	Gabriel Iglesias: I‰Ûªm Sorry For What I Said ...	Gabriel Iglesias: I'm Sorry for What I Said Wh...
9	Iron Man & Captain America: Heroes United	Iron Man and Captain America: Heroes United
10	The Mr. Peabody and Sherman Show	The Mr. Peabody & Sherman Show
12	100 Metros	100 Meters
13	Power Rangers Super Megaforce	Power Rangers Megaforce
14	The Great British Baking Show	The Great British Baking Show: Holidays
15	Operaì¤ìµes Especiais	Operações Especiais
16	Transformers: Rescue Bots	Transformers: Rescue Bots Academy
17	Marvel's Hulk: Where Monsters Dwell	Hulk: Where Monsters Dwell
19	American Crime	American Crime Story
29	Angry Birds	Angry Birds Toons
21	Octonauts	The Octonauts
22	Marvel's Agents of S.H.I.E.L.D.	Agents of S.H.I.E.L.D.
23	Shameless (U.S.)	Shameless
24	Scandal	A Scandall

	title_x	title_y
0	Thunder and the House Of Magic	Thunder and the House of Magic
1	Spy Kids 3: Game Over	Spy Kids 3-D: Game Over
2	Inspector Gadget 2	Inspector Gadget
4	D2: The Mighty Ducks	The Mighty Ducks
9	Star Wars: The Clone Wars	Star Wars: Clone Wars: Volume 1
34	Ninja Turtles: The Next Mutation	Teenage Mutant Ninja Turtles III

title
rating_x
ratingLevel
release year
user rating score
year
certificate
duration
genre
rating_y
description
stars
votes
top_stars
genre_count
index
budget

genres
homepage
id
keywords
original_language
original_title
overview
popularity
production_companies
production_countries
release_date
revenue
runtime
spoken_languages
status
tagline
vote_average
vote_count
cast
crew
director

привет! помоги мне обогатить таблицу с фильмами данными из интернета, в первую очередь проверяй википедию.

production_countries: список стран производства фильма через запятую

budget: бюджет фильма в долларах США, только число без дополнительных знаков

revenue: доходы (сборы) фильма в долларах США, только число без дополнительных знаков

Обязательно проверяй соответствие года выпуска

Для тех позиций, по которым ты не сможешь найти информацию, я укажу "Информация не найдена".

пожалуйста, оформи итог в таблицу и ПИШИ НА АНГЛИЙСКОМ

укажи ссылки, откуда ты берешь информацию

спасибо!

вот моя таблица:

title release year

Grey's Anatomy 2016 ….