Проект

по статистическому практикуму

Зайченко Николай и Мартюшова Кира, 332 гр.

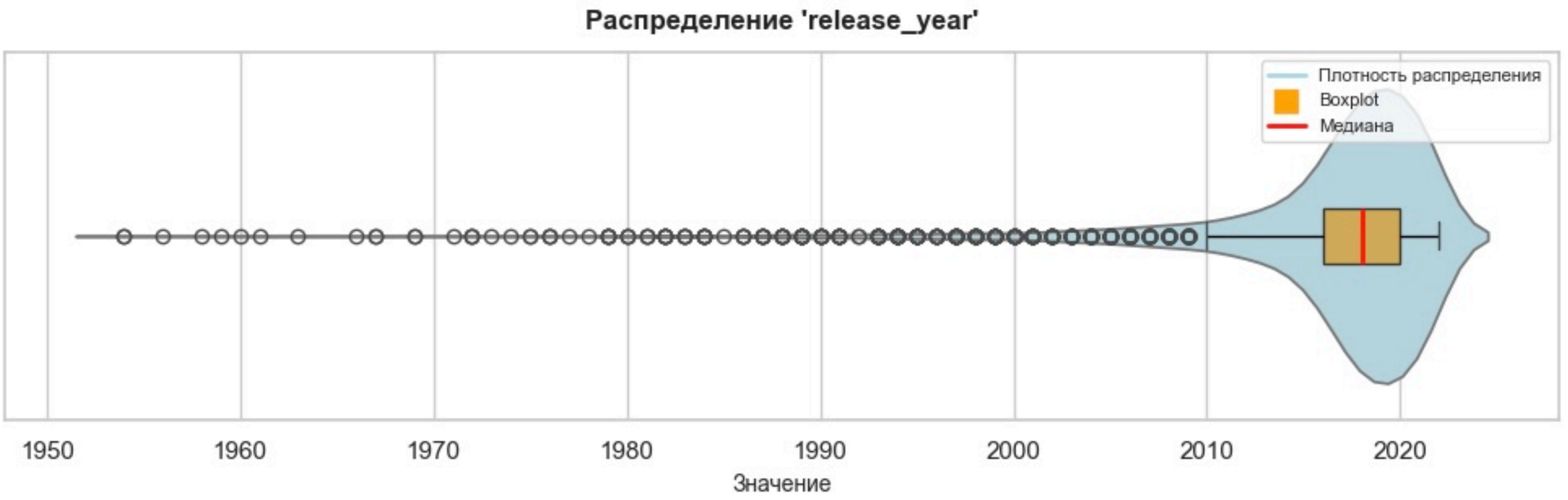
Мы объединили два набора данных:

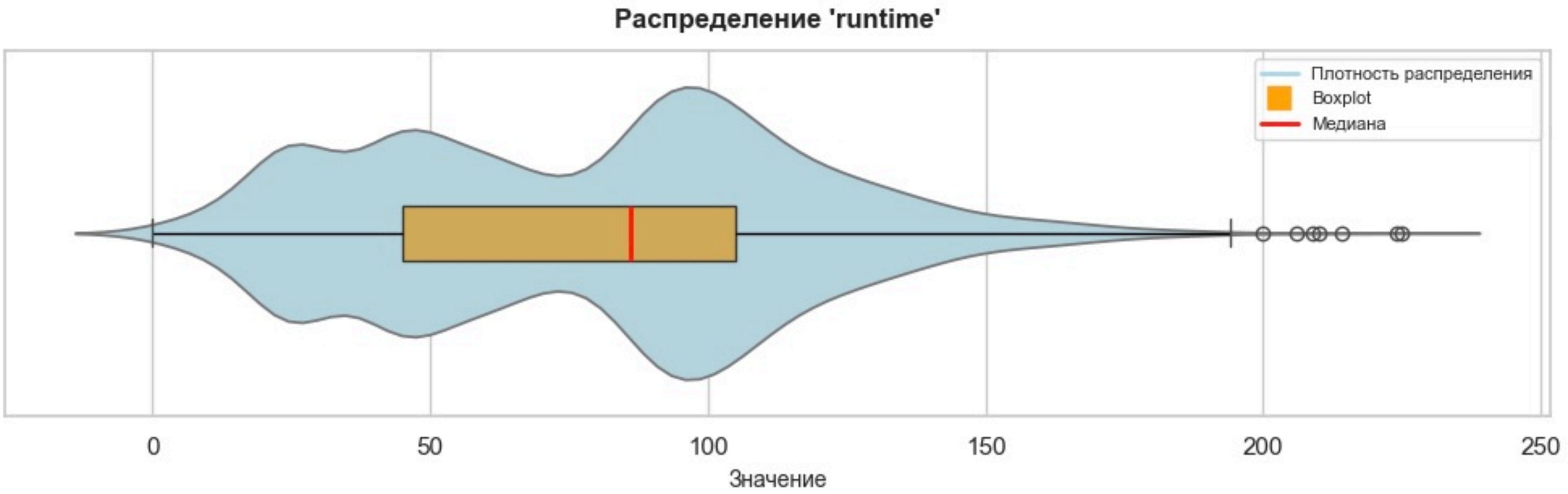
titles.csv — содержит основную информацию о фильмах.
credits.csv — содержит информацию об актёрах и их участии в фильмах.

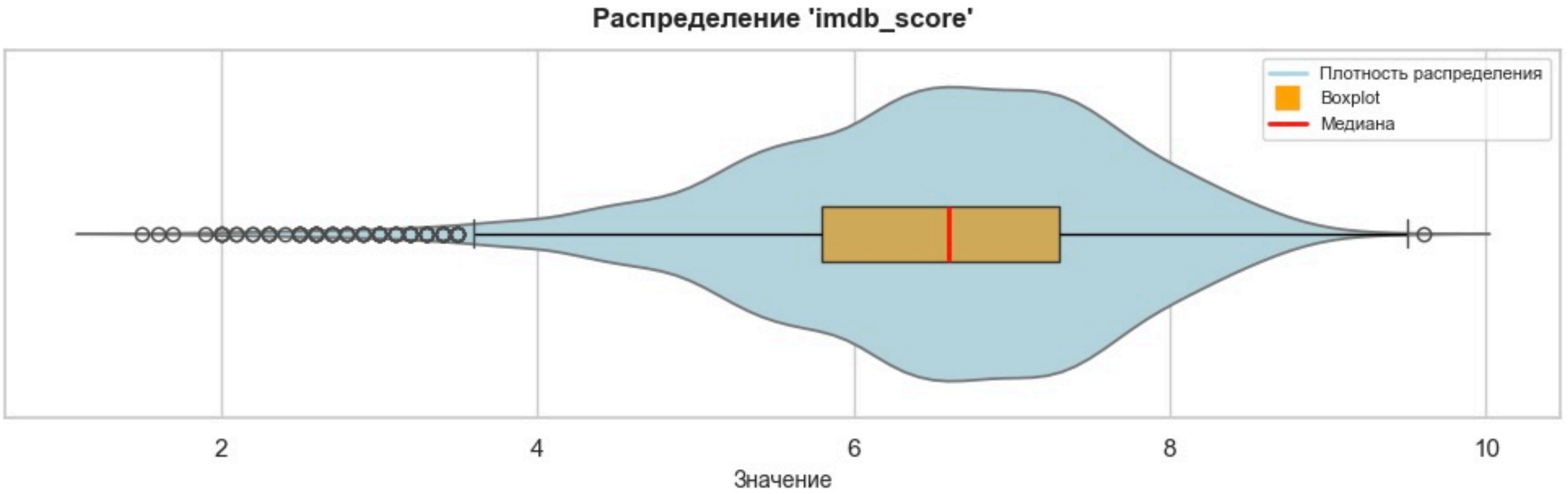
После объединения получаем следующие признаки:

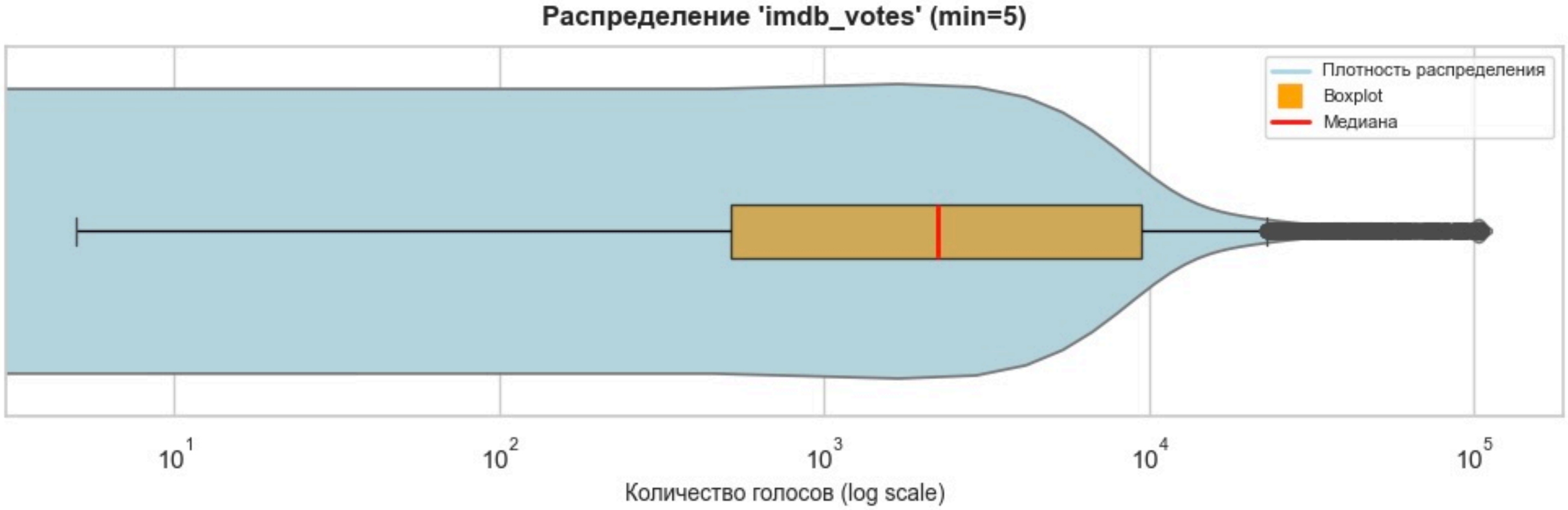
1. title — название фильма
2. type — тип (например, фильм, сериал)
3. release_year — год выпуска
4. age_certification — возрастной рейтинг
5. runtime — продолжительность фильма в минутах
6. genres — жанры (один или несколько)
7. production_countries — страна(-ы) производства
8. imdb_id — идентификатор IMDb
9. imdb_score — оценка IMDb (целевой признак)
10. imdb_votes — количество голосов на IMDb
11. actor_name — имя актёра

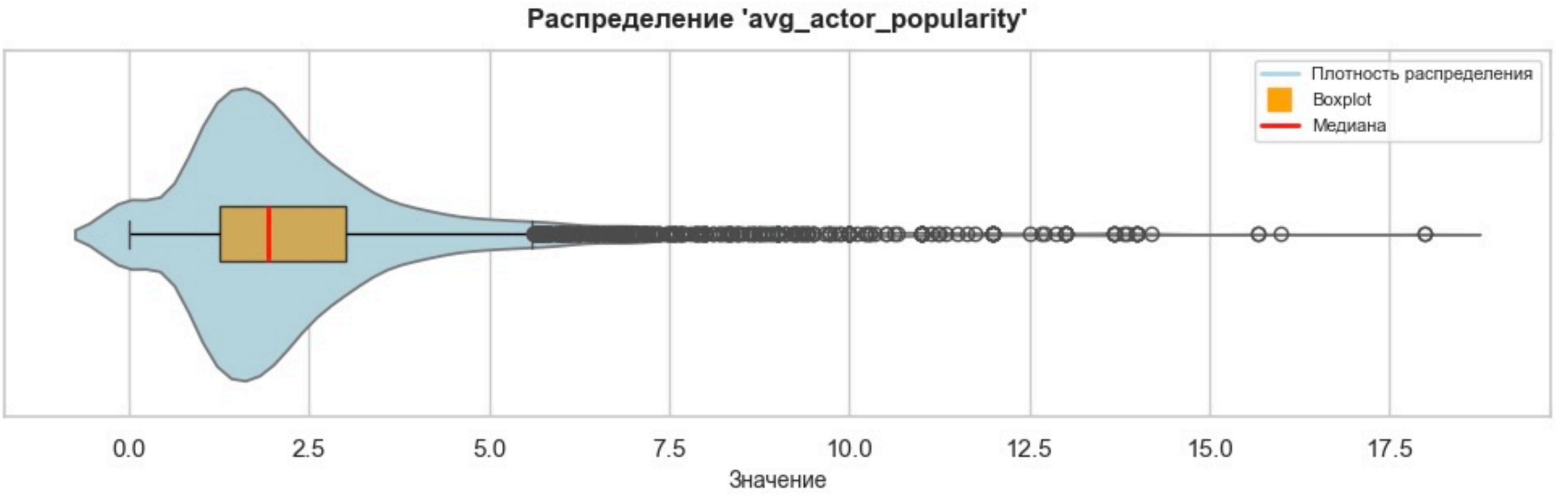
Анализ распределения числовых признаков: Violinplot + Boxplot

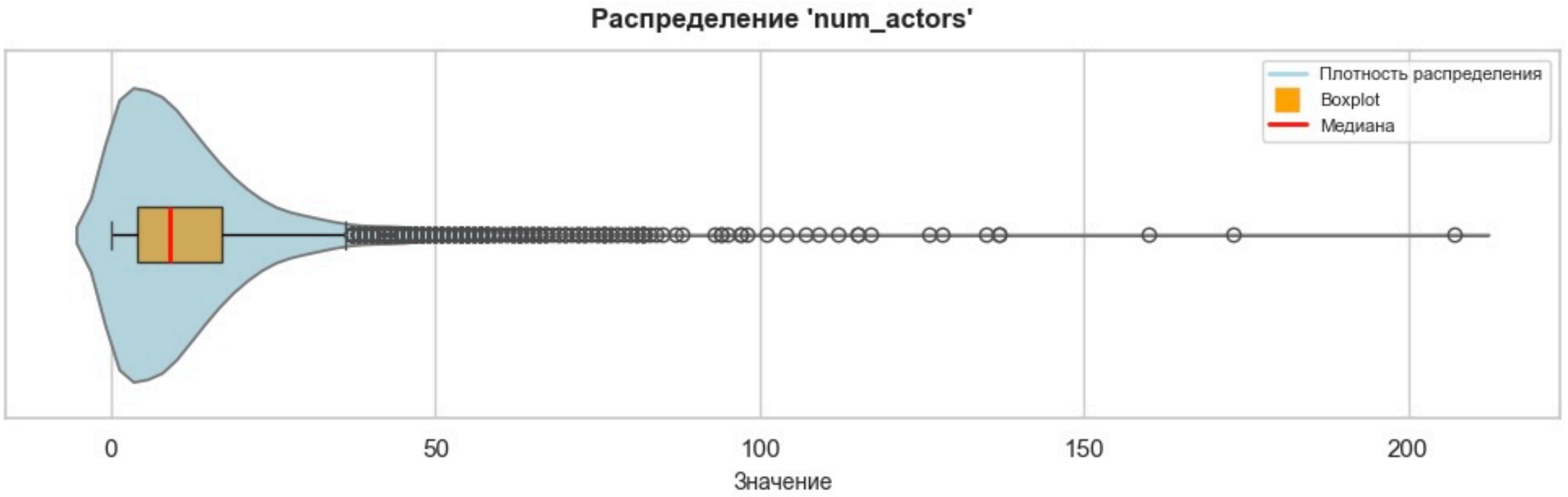


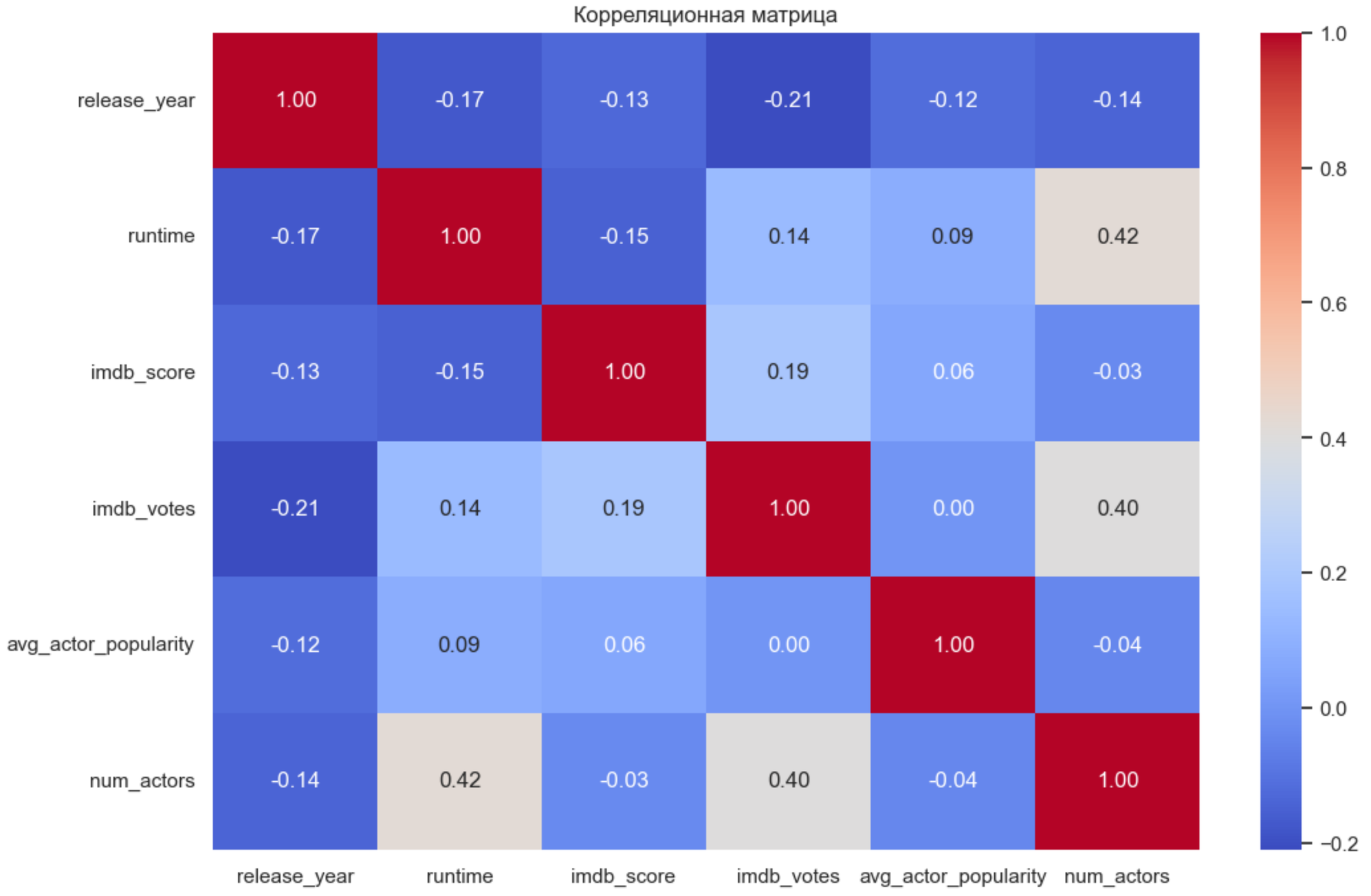

Распределение 'runtime'
Плотность распределения
Boxplot
Медиана
0
50
100
150
200
250
Значение

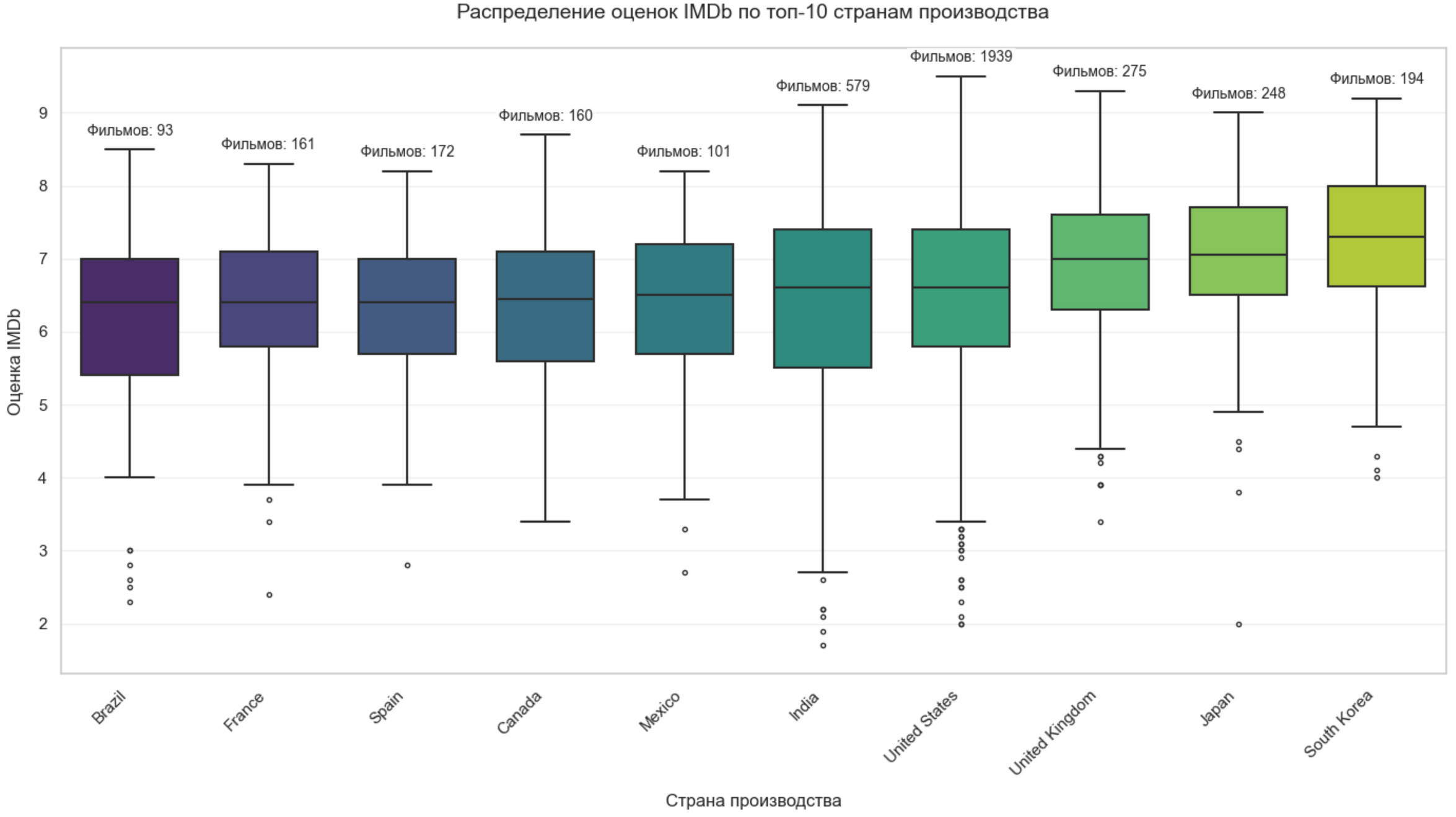

Распределение 'imdb_score'
Плотность распределения
Boxplot
Медиана
2
4
6
8
10
Значение

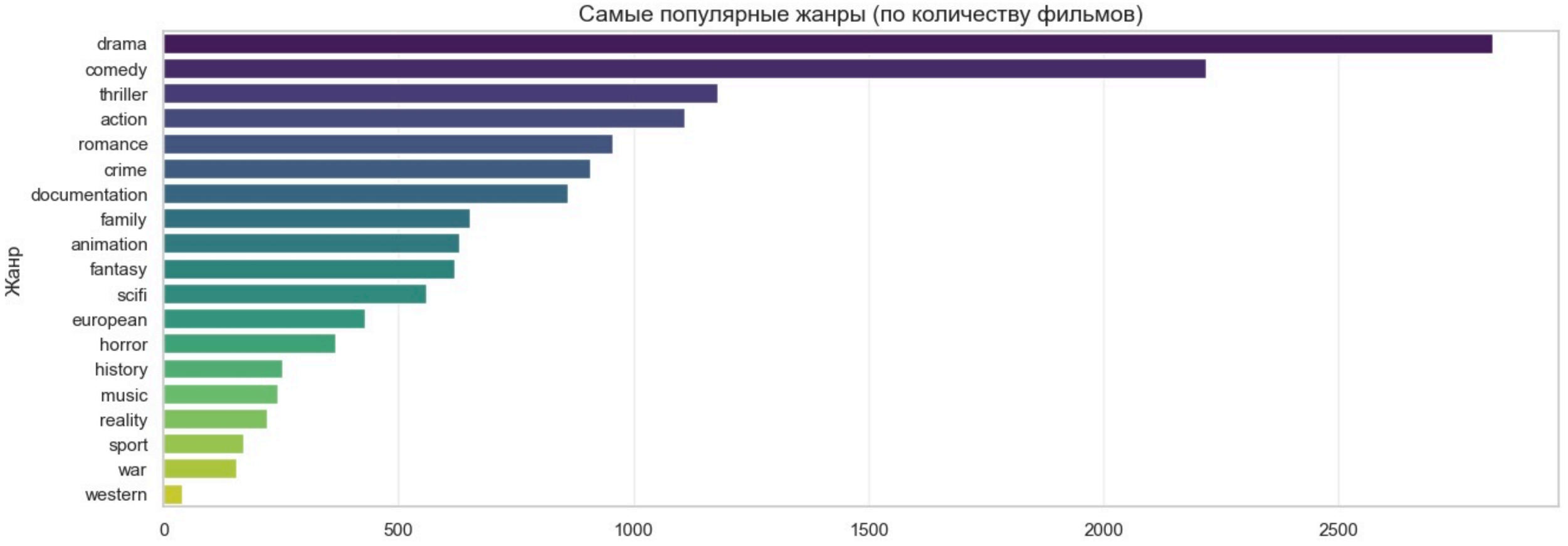

Распределение 'imdb_votes' (min=5)
Плотность распределения
Boxplot
Медиана
10^1
10^2
10^3
10^4
10^5
Количество голосов (log scale)

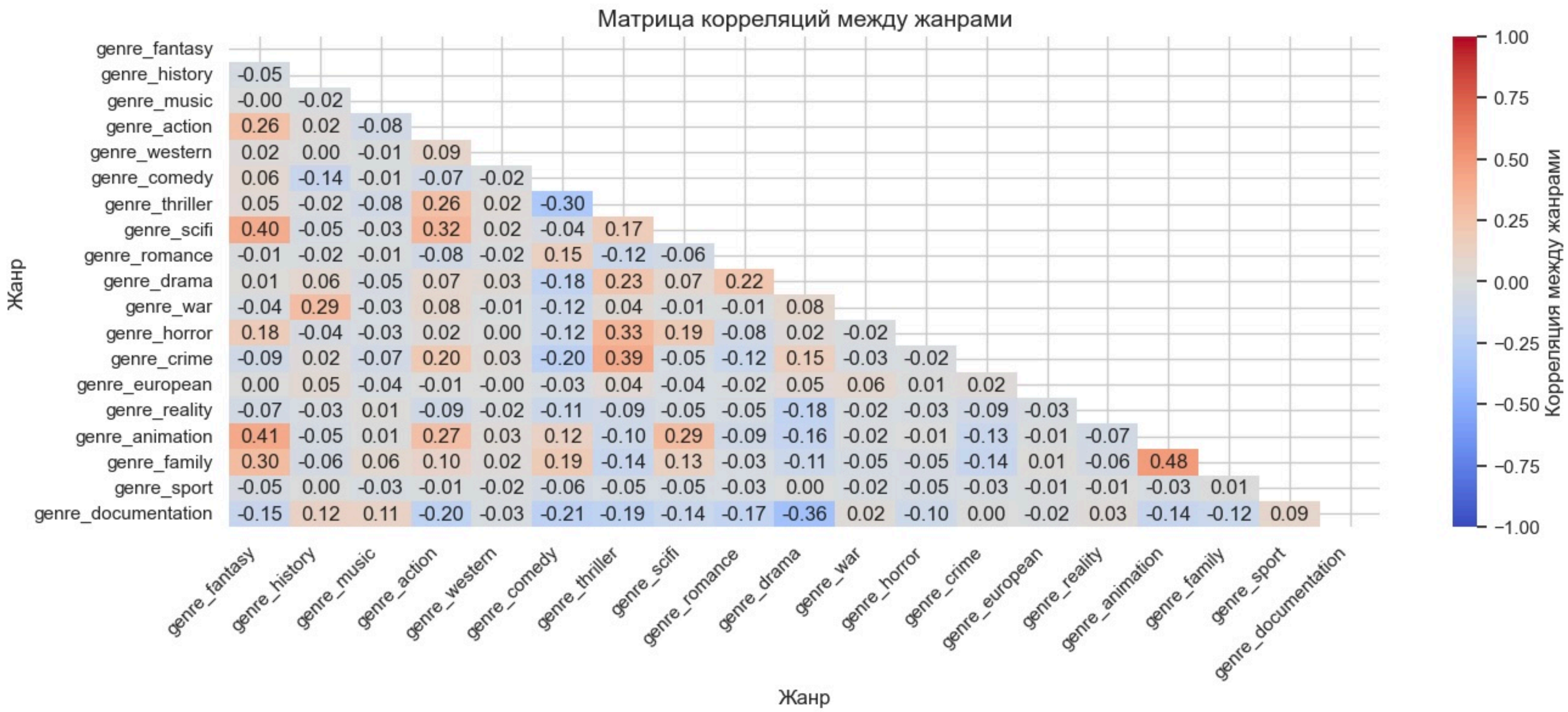

Распределение 'avg_actor_popularity'
Плотность распределения
Boxplot
Медиана
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Значение

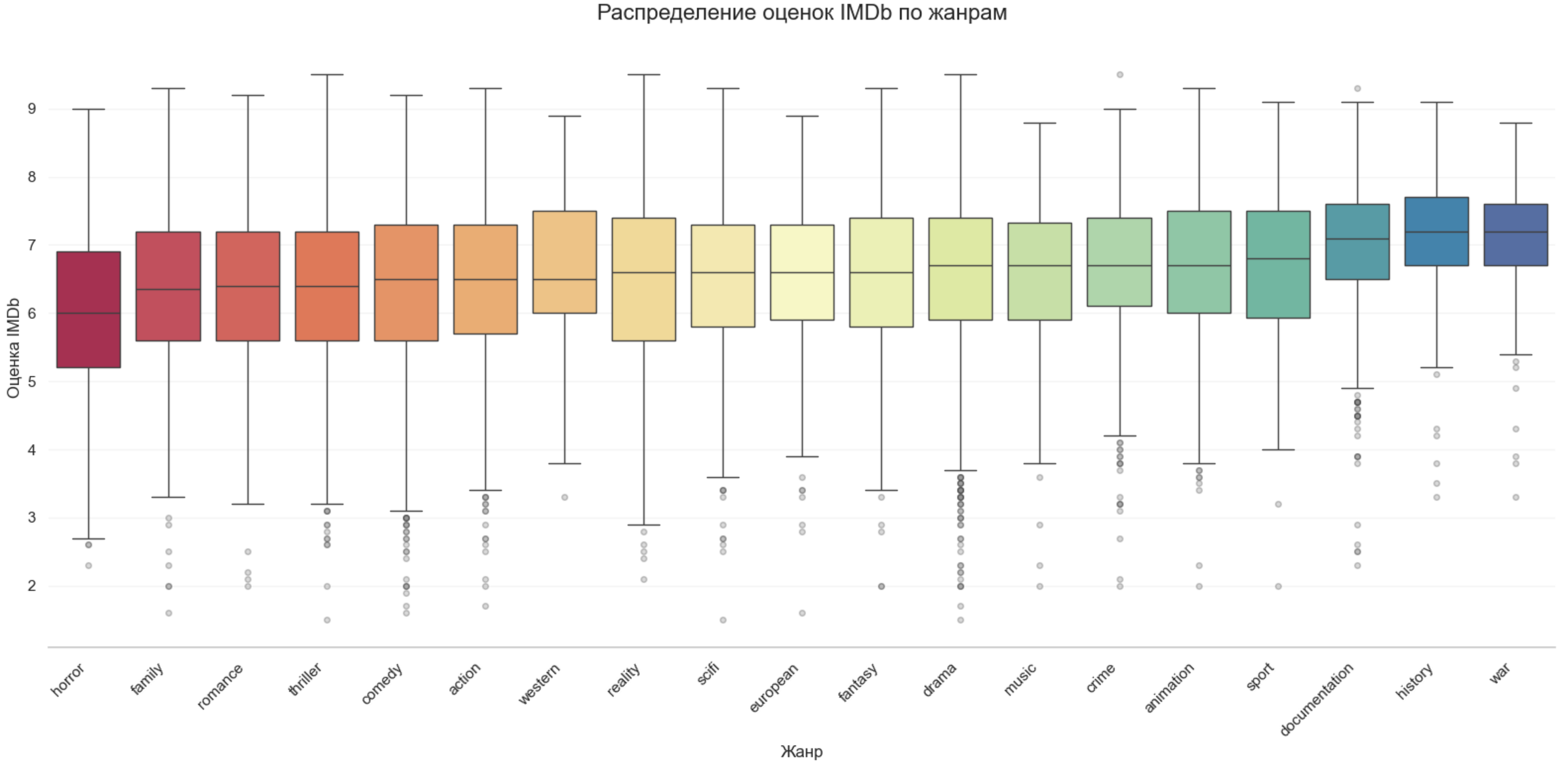

Распределение 'num_actors'
Плотность распределения
Boxplot
Медиана
0
50
100
150
200
Значение


Корреляционная матрица
release_year
runtime
imdb_score
imdb_votes
avg_actor_popularity
num_actors
1.00 -0.17 -0.13 -0.21 -0.12 -0.14
-0.17 1.00 -0.15 0.14 0.09 0.42
-0.13 -0.15 1.00 0.19 0.06 -0.03
-0.21 0.14 0.19 1.00 0.00 0.40
-0.12 0.09 0.06 0.00 1.00 -0.04
-0.14 0.42 -0.03 0.40 -0.04 1.00
release_year
runtime
imdb_score
imdb_votes
avg_actor_popularity
num_actors
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


Распределение оценок IMDb по топ-10 странам производства
Оценка IMDb
Страна производства
Фильмов: 93
Фильмов: 161
Фильмов: 172
Фильмов: 160
Фильмов: 101
Фильмов: 579
Фильмов: 1939
Фильмов: 275
Фильмов: 248
Фильмов: 194
Brazil
France
Spain
Canada
Mexico
India
United States
United Kingdom
Japan
South Korea

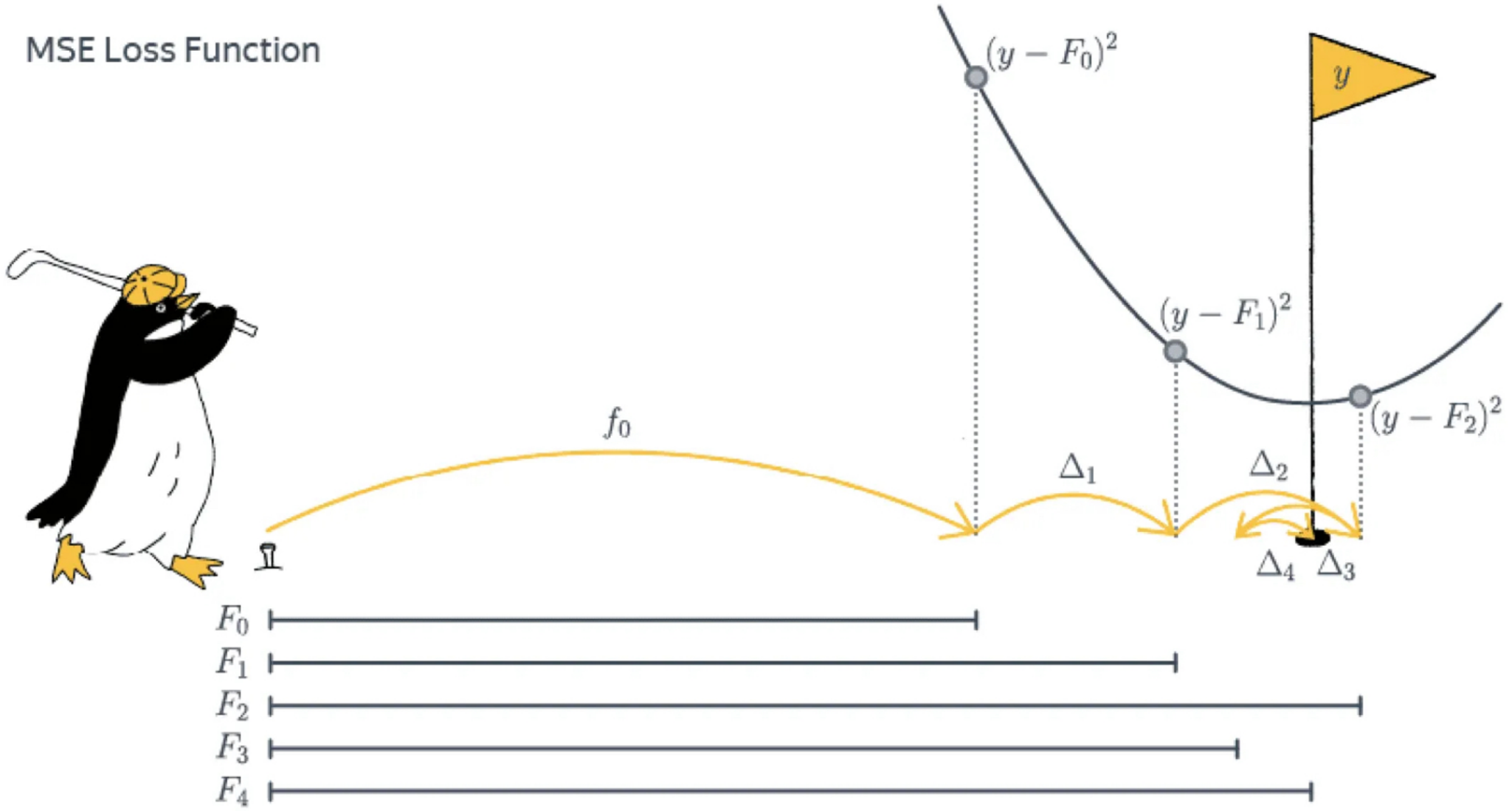

Самые популярные жанры (по количеству фильмов)
Жанр
drama
comedy
thriller
action
romance
crime
documentation
family
animation
fantasy
scifi
european
horror
history
music
reality
sport
war
western
0
500
1000
1500
2000
2500


Матрица корреляций между жанрами
Жанр
Жанр
Корреляция между жанрами
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
genre_fantasy
genre_history -0.05
genre_music -0.00 -0.02
genre_action 0.26 0.02 -0.08
genre_western 0.02 0.00 -0.01 0.09
genre_comedy 0.06 -0.14 -0.01 -0.07 -0.02
genre_thriller 0.05 -0.02 -0.08 0.26 0.02 -0.30
genre_scifi 0.40 -0.05 -0.03 0.32 0.02 -0.04 0.17
genre_romance -0.01 -0.02 -0.01 -0.08 -0.02 0.15 -0.12 -0.06
genre_drama 0.01 0.06 -0.05 0.07 0.03 -0.18 0.23 0.07 0.22
genre_war -0.04 0.29 -0.03 0.08 -0.01 -0.12 0.04 -0.01 -0.01 0.08
genre_horror 0.18 -0.04 -0.03 0.02 0.00 -0.12 0.33 0.19 -0.08 0.02 -0.02
genre_crime -0.09 0.02 -0.07 0.20 0.03 -0.20 0.39 -0.05 -0.12 0.15 -0.03 -0.02
genre_european 0.00 0.05 -0.04 -0.01 -0.00 -0.03 0.04 -0.04 -0.02 0.05 0.06 0.01 0.02
genre_reality -0.07 -0.03 0.01 -0.09 -0.02 -0.11 -0.09 -0.05 -0.05 -0.18 -0.02 -0.03 -0.09 -0.03
genre_animation 0.41 -0.05 0.01 0.27 0.03 0.12 -0.10 0.29 -0.09 -0.16 -0.02 -0.01 -0.13 -0.01 -0.07
genre_family 0.30 -0.06 0.06 0.10 0.02 0.19 -0.14 0.13 -0.03 -0.11 -0.05 -0.05 -0.14 0.01 -0.06 0.48
genre_sport -0.05 0.00 -0.03 -0.01 -0.02 -0.06 -0.05 -0.05 -0.03 0.00 -0.02 -0.05 -0.03 -0.01 -0.01 -0.03 0.01
genre_documentation -0.15 0.12 0.11 -0.20 -0.03 -0.21 -0.19 -0.14 -0.17 -0.36 0.02 -0.10 0.00 -0.02 0.03 -0.14 -0.12 0.09
genre_fantasy
genre_history
genre_music
genre_action
genre_western
genre_comedy
genre_thriller
genre_scifi
genre_romance
genre_drama
genre_war
genre_horror
genre_crime
genre_european
genre_reality
genre_animation
genre_family
genre_sport
genre_documentation


Распределение оценок IMDb по жанрам
Оценка IMDb
9
8
7
6
5
4
3
2
horror
family
romance
thriller
comedy
action
western
reality
scifi
european
fantasy
drama
music
crime
animation
sport
documentation
history
war
Жанр

Градиентный бустинг

MSE Loss Function



Градиентный бустинг строит взвешенную композицию алгоритмов:

$$a(x) = \sum_{i=1}^{T} \alpha_i b_i(x),$$

где $b_i(x)$ — базовые алгоритмы, а α_i — их веса.

$$a(x) = \sum_{i=1}^{T} \alpha_i b_i(x)$$

$$Q(\alpha, b) = \sum_{i=1}^{l} L(a(x_i), y_i) = \sum_{i=1}^{l} L\left(\underbrace{\underbrace{\sum_{t=1}^{T-1} \alpha_t b_t(x_i)}_{f_{T-1,j}} + \alpha b(x_i)}_{f_{T,j}}, y_i \right) \to \min_{\alpha, b}$$

$$Q(\alpha, b) = \sum_{i=1}^{l} L(a(x_i), y_i) = \sum_{i=1}^{l} L\left(\underbrace{\underbrace{\sum_{t=1}^{T-1} \alpha_t b_t(x_i)}_{f_{T-1,j}} + \alpha b(x_i)}_{f_{T,j}}, y_i \right) \to \min_{\alpha, b}$$

$$f_{T-1} = \left[f_{T-1,1}, f_{T-1,2}, \ldots, f_{T-1,l} \right]^T$$

$$f_T = \left[f_{T,1}, f_{T,2}, \ldots, f_{T,l} \right]^T$$

$$Q(\alpha, b) = \sum_{i=1}^{l} L(a(x_i), y_i) = \sum_{i=1}^{l} L\left(\underbrace{\underbrace{\sum_{t=1}^{T-1} \alpha_t b_t(x_i)}_{f_{T-1,j}} + \alpha b(x_i)}_{f_{T,j}}, y_i \right) \to \min_{\alpha, b}$$

$$f_{T-1} = \left[f_{T-1,1}, f_{T-1,2}, \ldots, f_{T-1,l} \right]^T$$

$$f_T = \left[f_{T,1}, f_{T,2}, \ldots, f_{T,l} \right]^T$$

$$g_i = \frac{\partial L(f_{T-1,i}, y_i)}{\partial f_{T-1,i}} = L'(f_{T-1,i}, y_i), \quad i = 1, \ldots, l$$

$$Q(\alpha, b) = \sum_{i=1}^{l} L(a(x_i), y_i) = \sum_{i=1}^{l} L\left(\underbrace{\underbrace{\sum_{t=1}^{T-1} \alpha_t b_t(x_i)}_{f_{T-1,j}} + \alpha b(x_i)}_{f_{T,j}}, y_i \right) \to \min_{\alpha, b}$$

$$f_{T-1} = \left[f_{T-1,1}, f_{T-1,2}, \ldots, f_{T-1,l} \right]^T \qquad f_{T,i} = f_{T-1,i} - \alpha g_i, \quad i = 1, \ldots, l$$

$$f_T = \left[f_{T,1}, f_{T,2}, \ldots, f_{T,l} \right]^T \qquad f_{T,i} = f_{T-1,i} + \alpha b(x_i), \quad i = 1, \ldots, l$$

$$g_i = \frac{\partial L(f_{T-1,i}, y_i)}{\partial f_{T-1,i}} = L'(f_{T-1,i}, y_i), \quad i = 1, \ldots, l$$

$$b_T = \arg\min_b \sum_{i=1}^{l} \left(b(x_i) + g_i\right)^2$$

$$b_T = \arg\min_b \sum_{i=1}^{l} \left(b(x_i) + g_i\right)^2$$

$$\alpha_T = \arg\min_{\alpha>0} \sum_{i=1}^{l} L(f_{T-1,i} + \alpha b_T(x_i), y_i)$$

В качестве входных признаков ($\mathbf{X}$) используются числовые характеристики фильмов:

- Длительность фильма (runtime);
- Год выпуска (release_year);
- Число актеров (num_actors)
- Популярность актера (avg_actor_popularity);
- Есть ли в фильме топ-актер (has_top_actor);
- Жанр фильма (genres);
- Страна производства (country).

В качестве целевой переменной ($\mathbf{Y}$) берётся рейтинг IMDb (imdb_score), который необходимо предсказать.

Для модели градиентного бустинга наиболее популярной функцией потерь является среднеквадратичная ошибка (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Для модели градиентного бустинга наиболее популярной функцией потерь является среднеквадратичная ошибка (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Средняя абсолютная ошибка (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|;$$

Коэффициент детерминации (R^2):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Результаты

Модель	MSE	MAE	R^2
Градиентный бустинг	0,9885	0,7575	0,2914

Кластеризация: OPTICS и спектральный подход

OPTICS (Ordering Points To Identify the Clustering Structure)
- Похоже на DBSCAN, но не требует фиксированного радиуса
- Строит упорядоченный список точек по плотности
- Обнаруживает кластеры разной плотности и шум

Параметры OPTICS

- min_samples: определяет минимальный размер кластера
- cluster_method: метод кластеризации (в нашем случае это метод “xi”)
- xi: чувствительность к изменению плотности

Ключевые понятия OPTICS

Core Distance — радиус, включающий min_samples

Reachability Distance(q,p) = max(core_dist(p), dist(p,q))

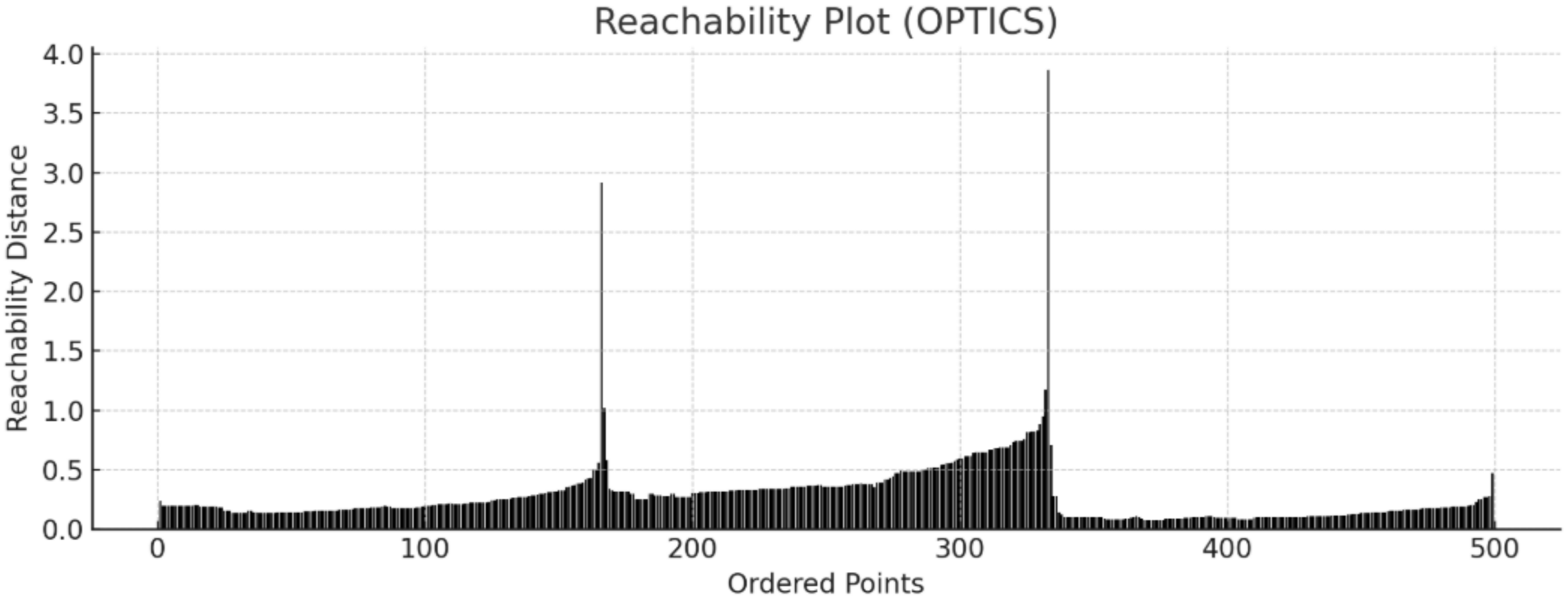
Метод xi в OPTICS

- Кластеры определяются как области **стабильной плотности**, между которыми резкие изменения.
- Используется относительное изменение расстояния достижимости:

$$\frac{\text{reachability\_dist}_{i+1} - \text{reachability\_dist}_i}{\text{reachability\_dist}_i} > \xi$$

- Параметр $\xi \in (0, 1)$ — чувствительность к границам кластеров (обычно 0.03–0.05).

Пример графика достижимости



Спектральная кластеризация

1. Строим граф похожести
2. Вычисляем лапласиан графа
3. Находим собственные вектора
4. Применяем k-means

Матрица смежности и лапласиан

- Матрица A: $A_{ij} = \begin{cases} 1, & \text{если } j \in kNN(i) \\ 0, & \text{иначе} \end{cases}$
- Матрица степеней: $D_{ii} = \sum_j A_{ij}$
- Нормализованный лапласиан: $L = I - D^{-1/2} A D^{-1/2}$

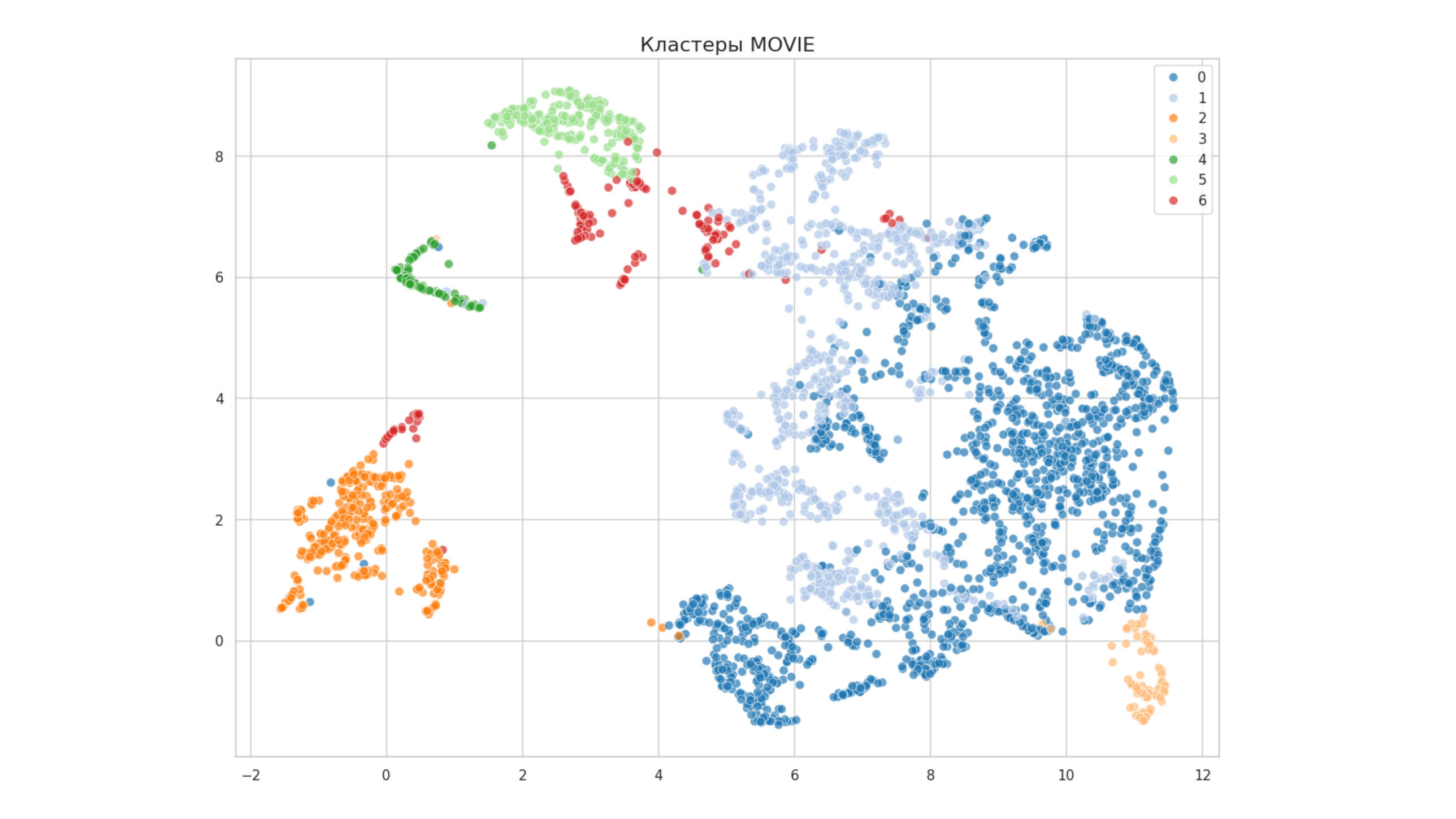
Кластеризация фильмов с помощью комбинации данных методов:

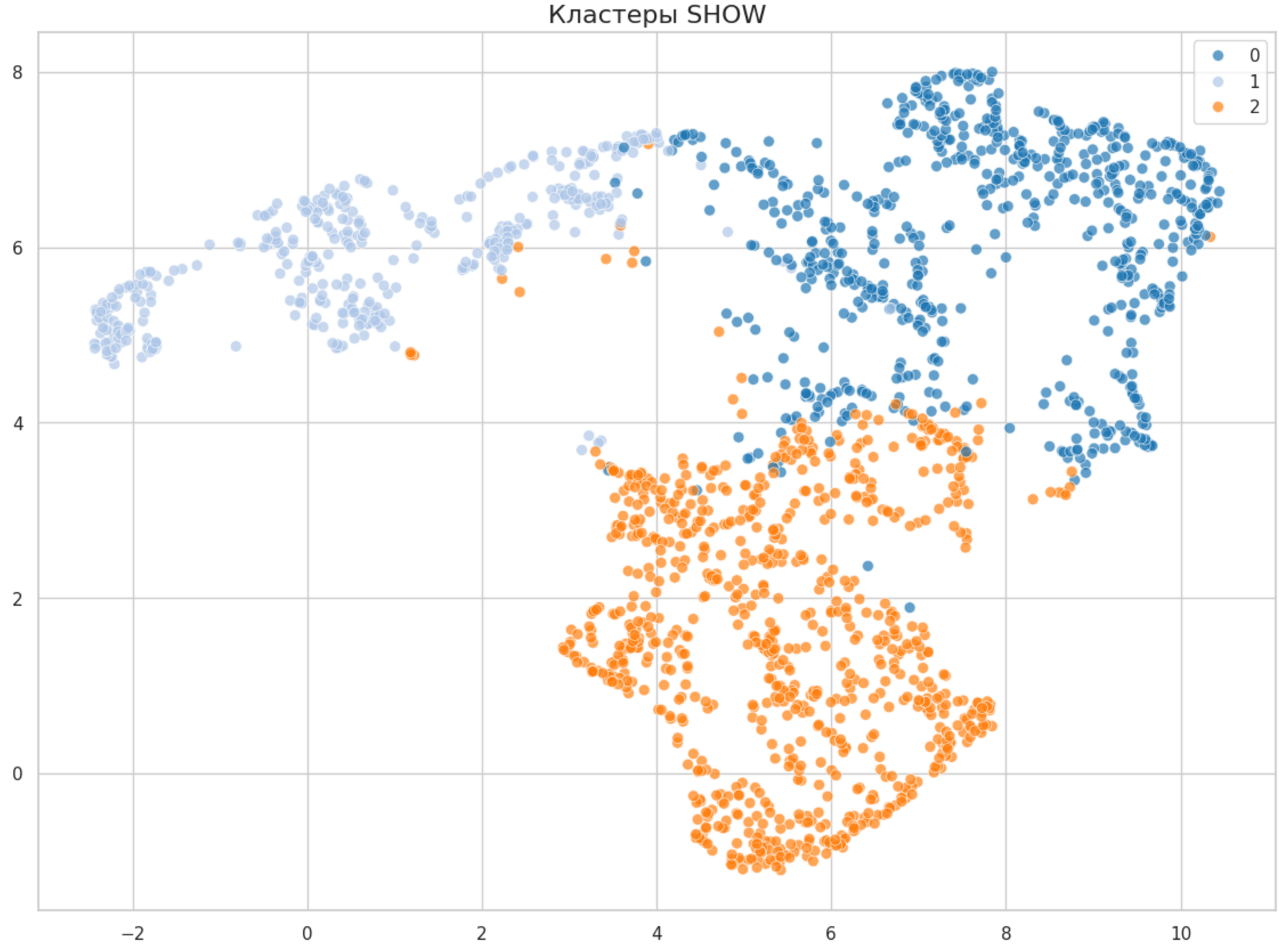
Цель: Мы хотим автоматически разделить фильмы на группы (кластеры), основываясь на количественных характеристиках.

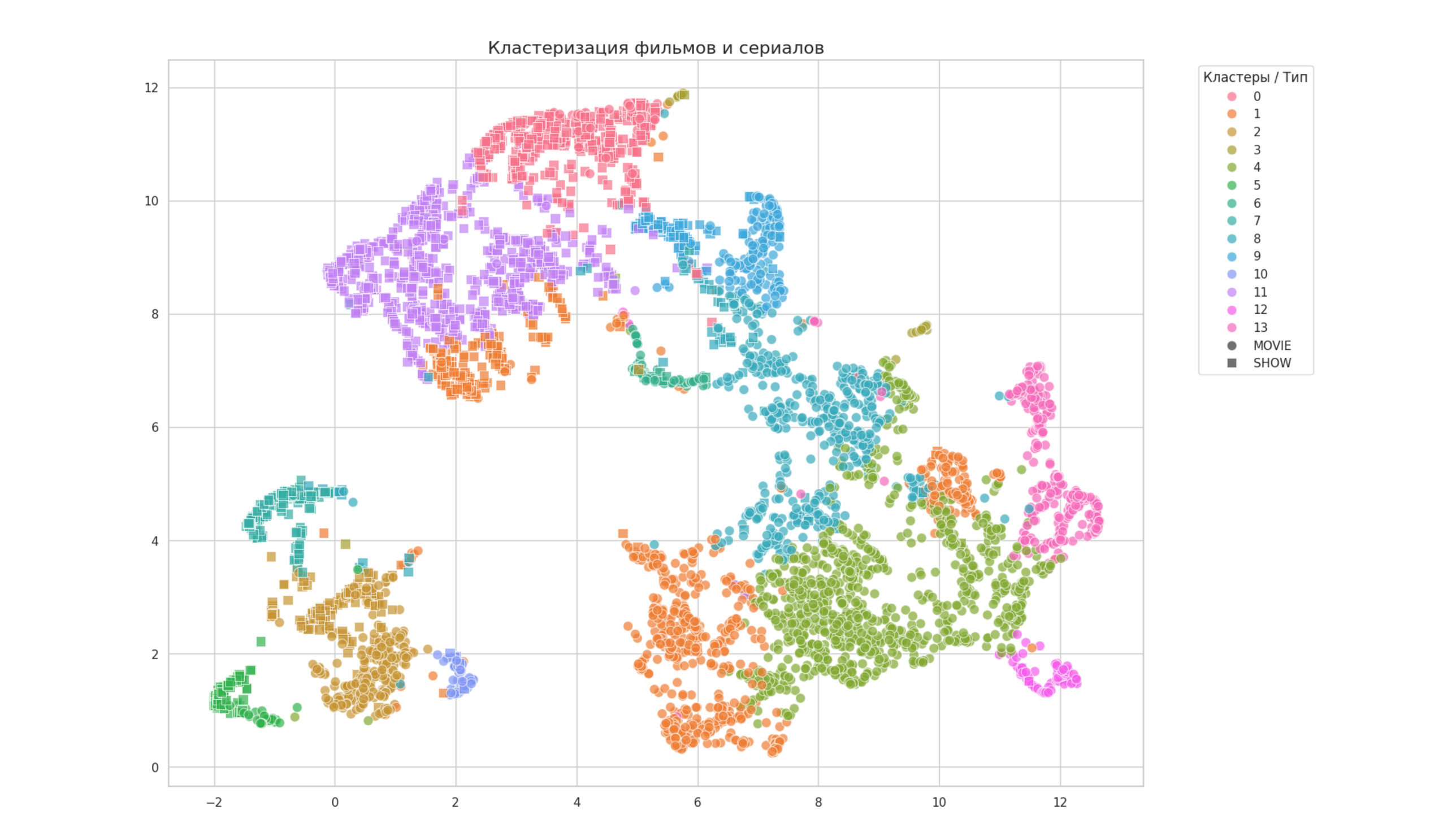
Это позволяет понять, какие типы фильмов существуют в датасете, какие у них общие свойства, и как можно использовать эту информацию, например, для рекомендаций или анализа рынка.

Используемые признаки:

- runtime — длительность фильма;
- release_year — год выхода;
- imdb_score — рейтинг IMDb;
- imdb_votes — количество голосов на IMDb;
- жанры.


Кластеры MOVIE
0
1
2
3
4
5
6
8
6
4
2
0
−2
0
2
4
6
8
10
12


Кластеры SHOW
0
1
2
8
6
4
2
0
−2
0
2
4
6
8
10


Кластеризация фильмов и сериалов
Кластеры / Тип
0
1
2
3
4
5
6
7
8
9
10
11
12
13
MOVIE
SHOW
12
10
8
6
4
2
0
−2
0
2
4
6
8
10
12