



Анализ топовых фильмов и предсказание хитовости фильмов

Проект посвящён изучению топовых фильмов и прогнозированию их успеха. Основной источник данных – TMDB Dataset, включающий тысячи фильмов с различными метриками.



Общий Обзор Топ Фильмов (1902-2024)

Общее количество фильмов

8560 фильмов для анализа

Критерии отбора

- Популярность (popularity)
- Средний рейтинг (vote_average)
- Дата выпуска (release_date)

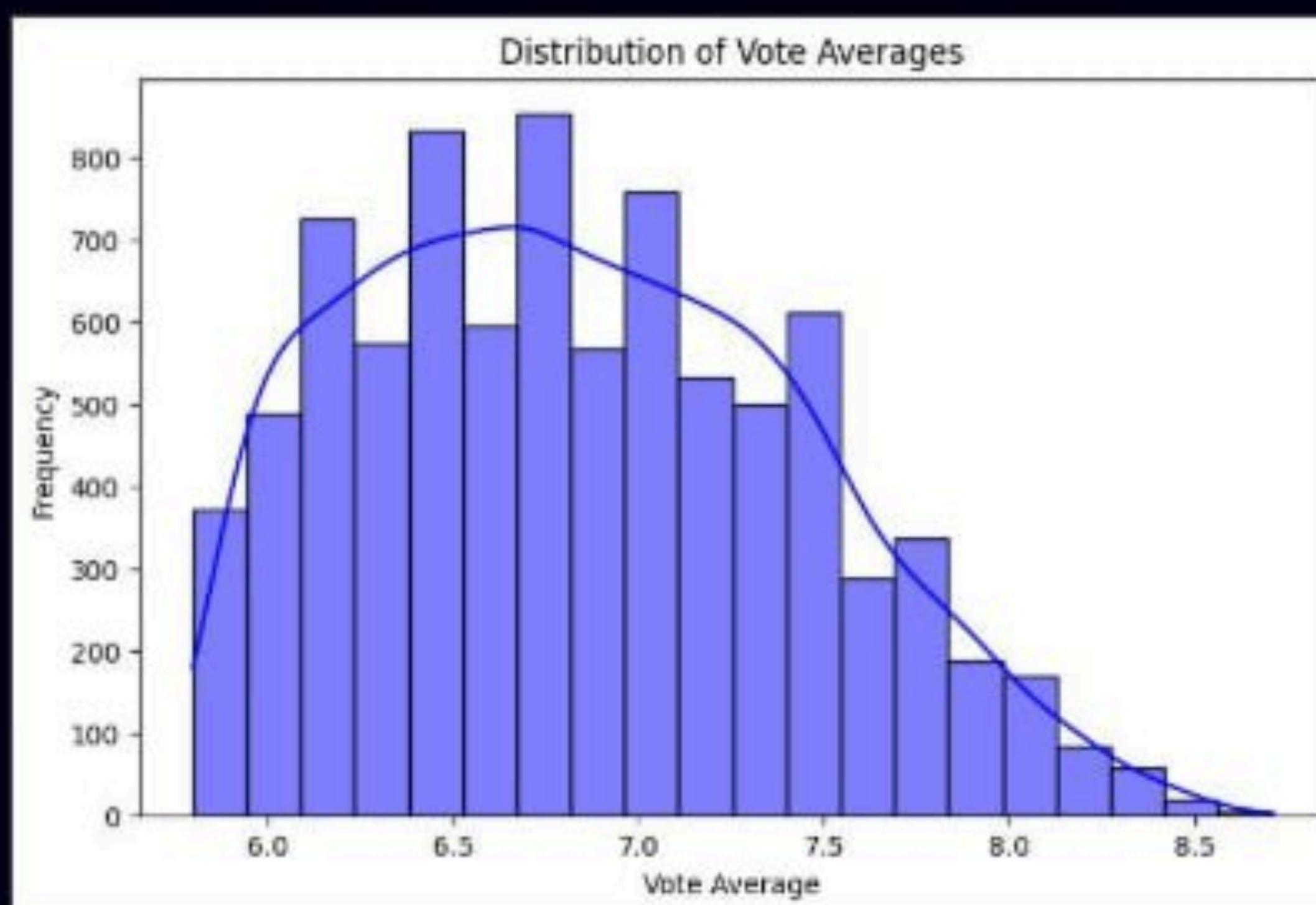
Анализируемые поля

- title
- overview
- release_date
- popularity
- vote_average
- vote_count

Графики распределений рейтинга и популярности

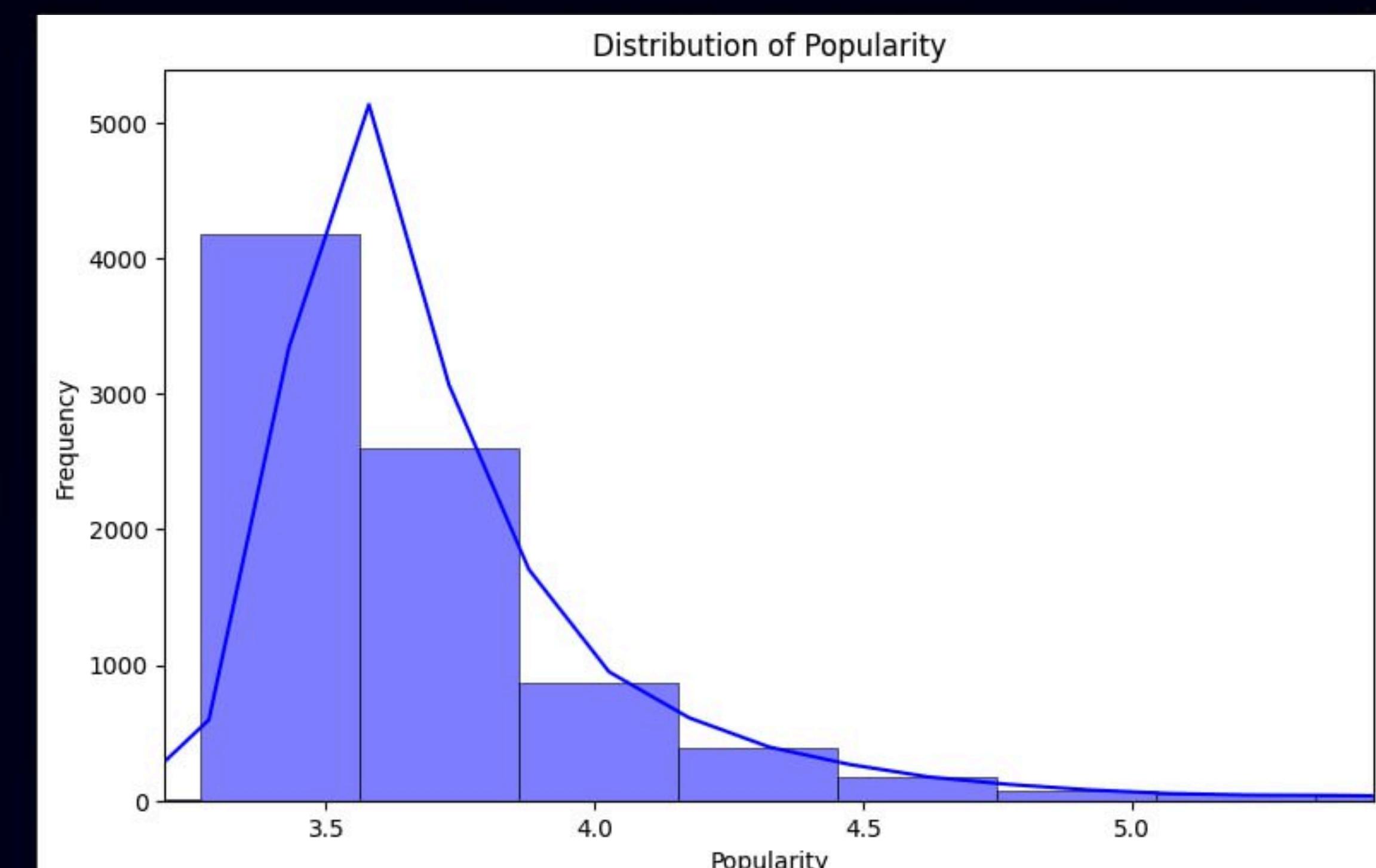
Распределение рейтинга

Как варьируются средние оценки фильмов в выборке.



Распределение популярности

Развёртка по показателям популярности фильмов среди зрителей.

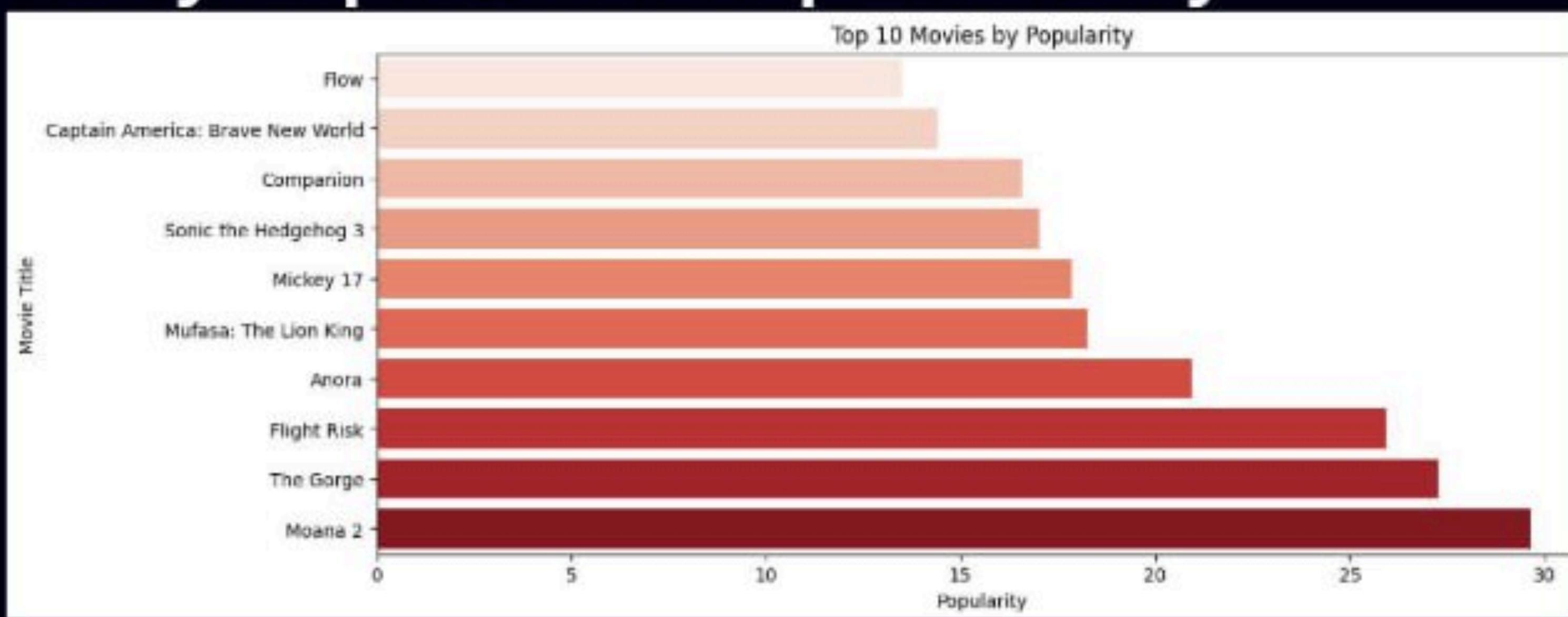


Топ-10 фильмов по популярности и рейтингу



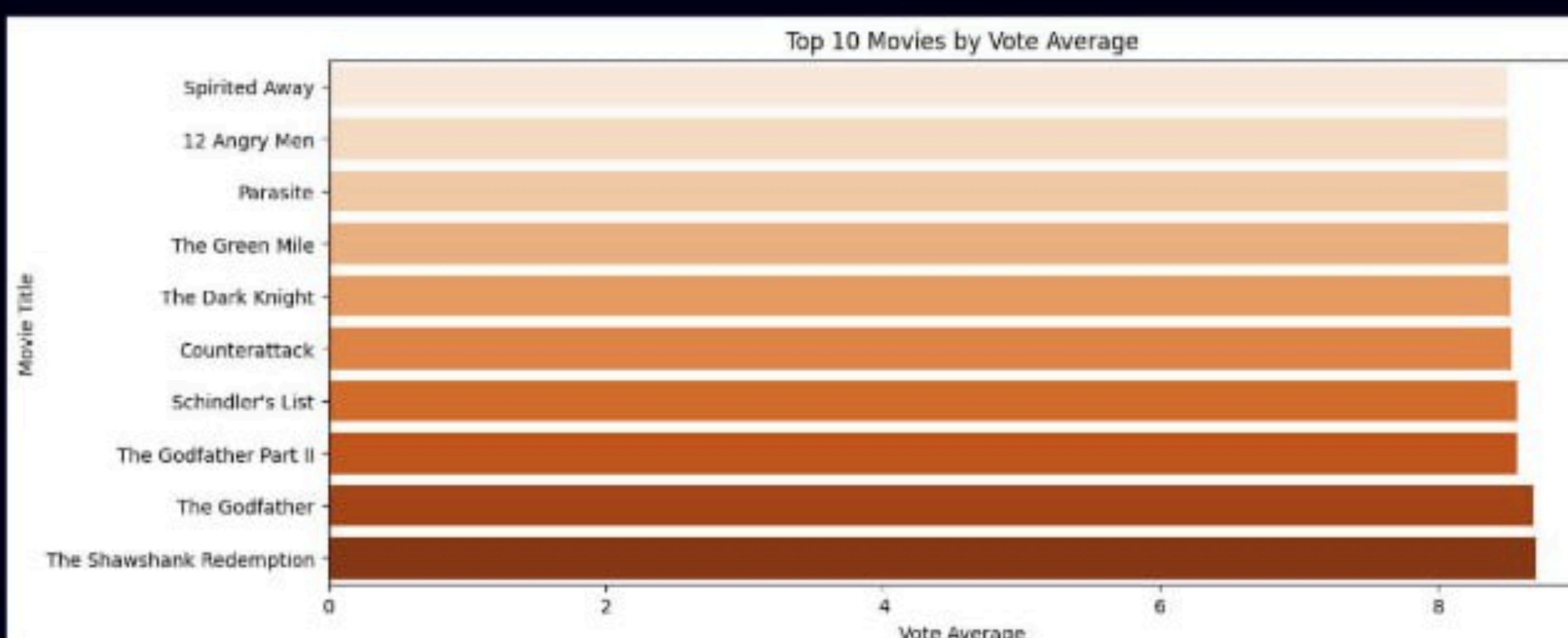
Популярность

Фильмы с наибольшим числом просмотров и обсуждений.



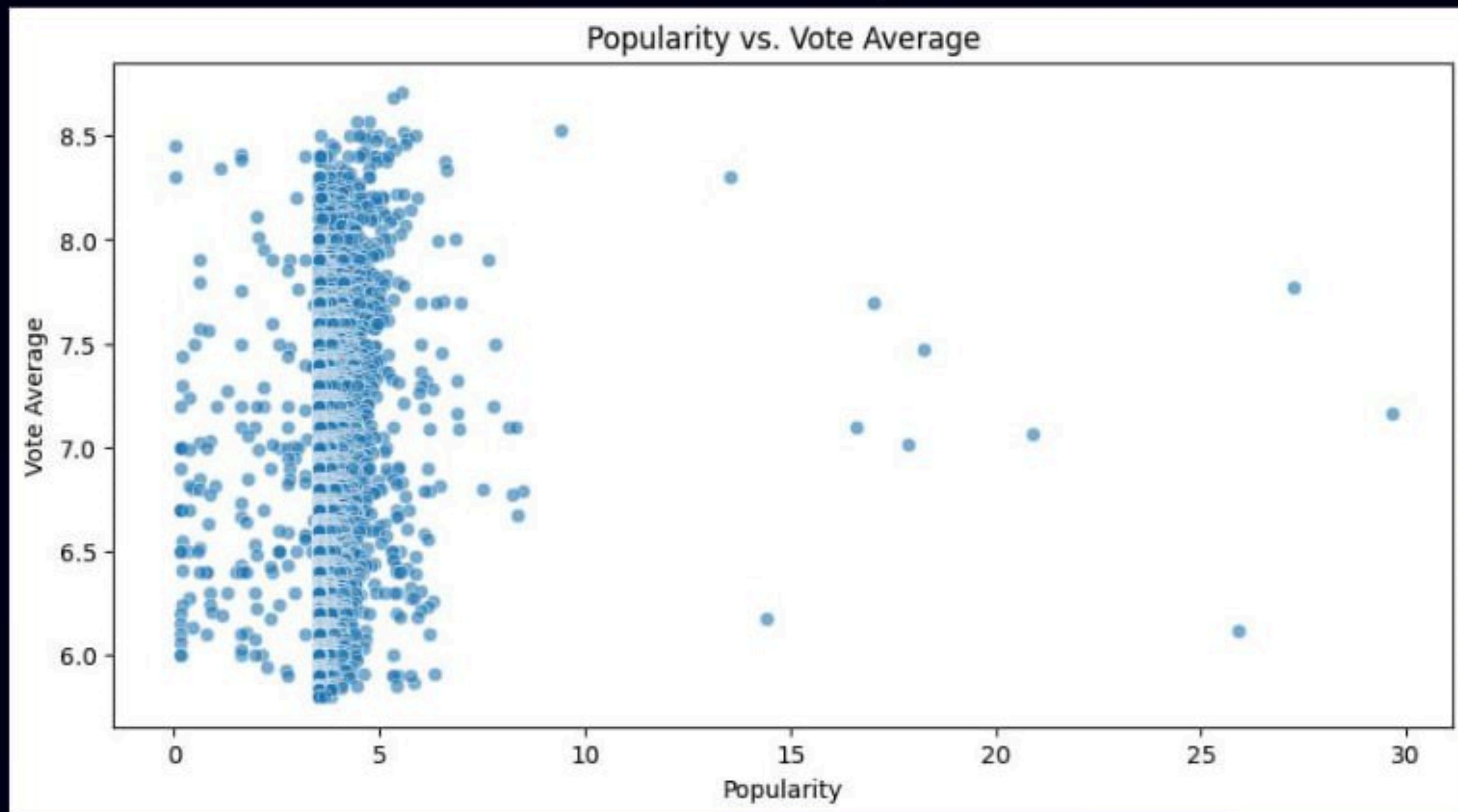
Рейтинг

Лучшие оценки по мнению зрителей и критиков.



Корреляция между популярностью и рейтингом

Исследуем взаимосвязь между оценками зрителей и популярностью фильмов. Корреляция поможет выявить закономерности хитов.





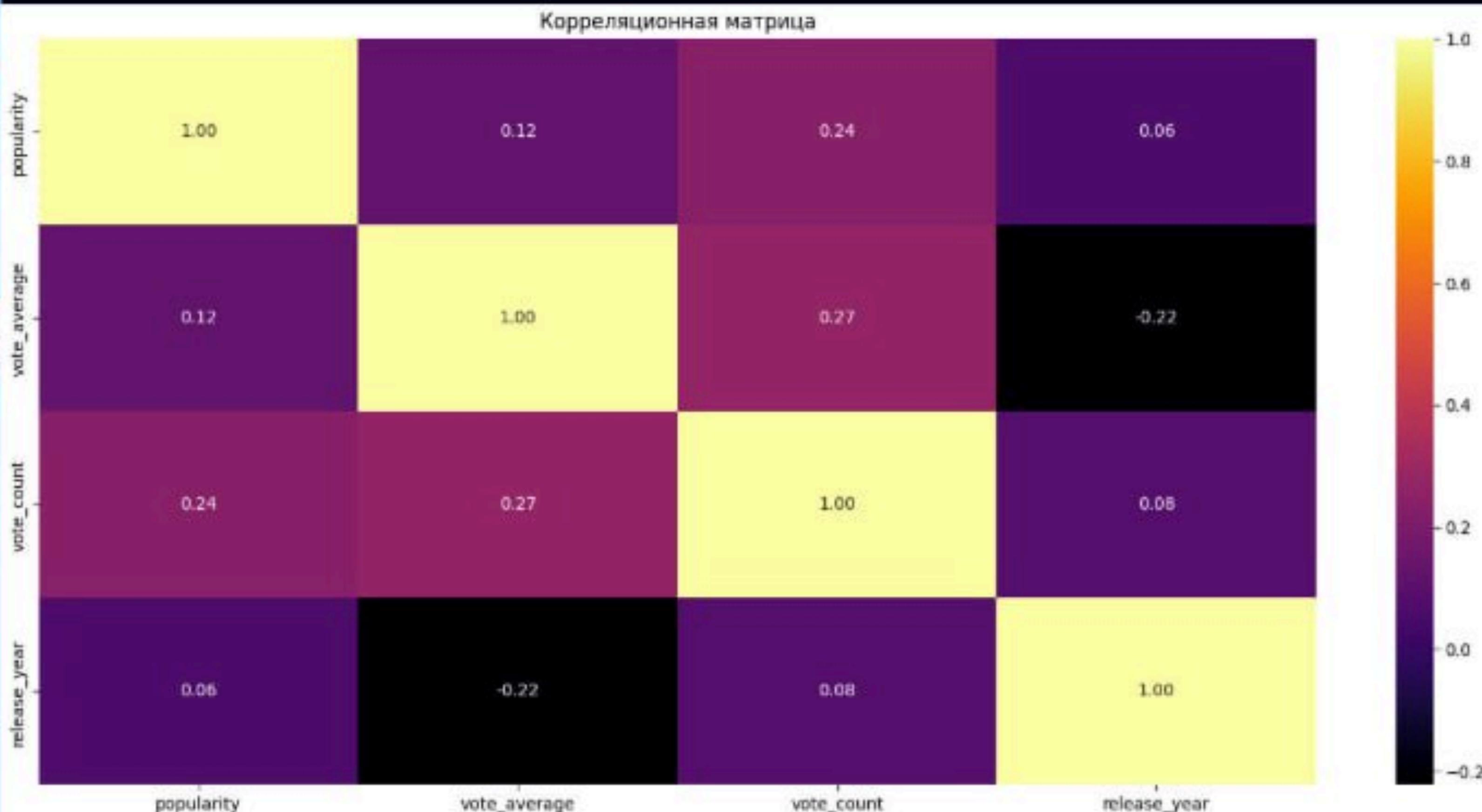
Количество выпущенных фильмов по годам

Анализ тенденций выпуска фильмов за более чем столетний период, выявление пиков и спадов производства.



Корреляционная матрица для датасета

Матрица показывает взаимосвязи между популярностью, рейтингом, датой выпуска и другими переменными.

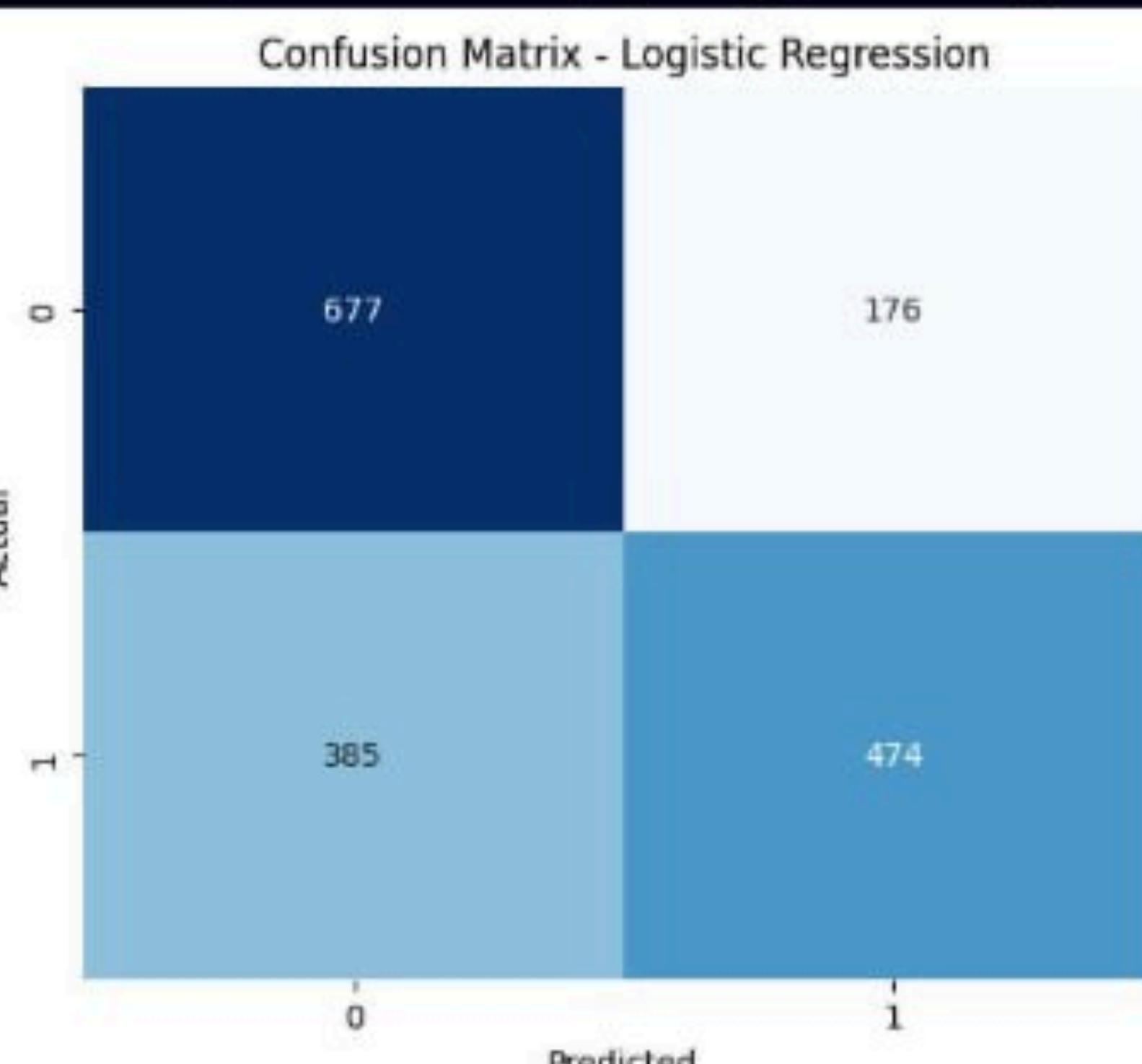


Прогнозное моделирование

1

Логистическая регрессия

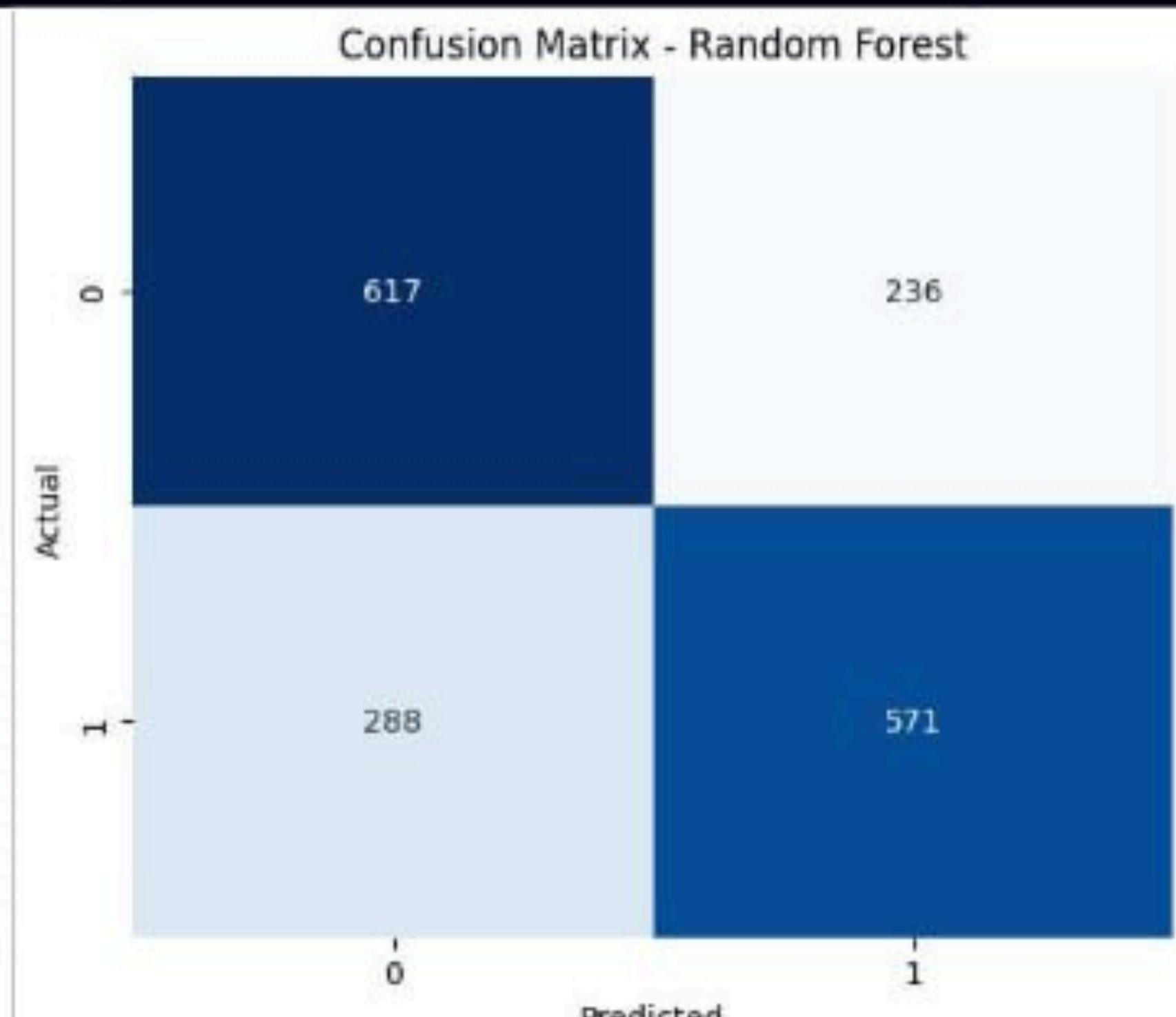
Простая и интерпретируемая модель.



2

Случайный лес

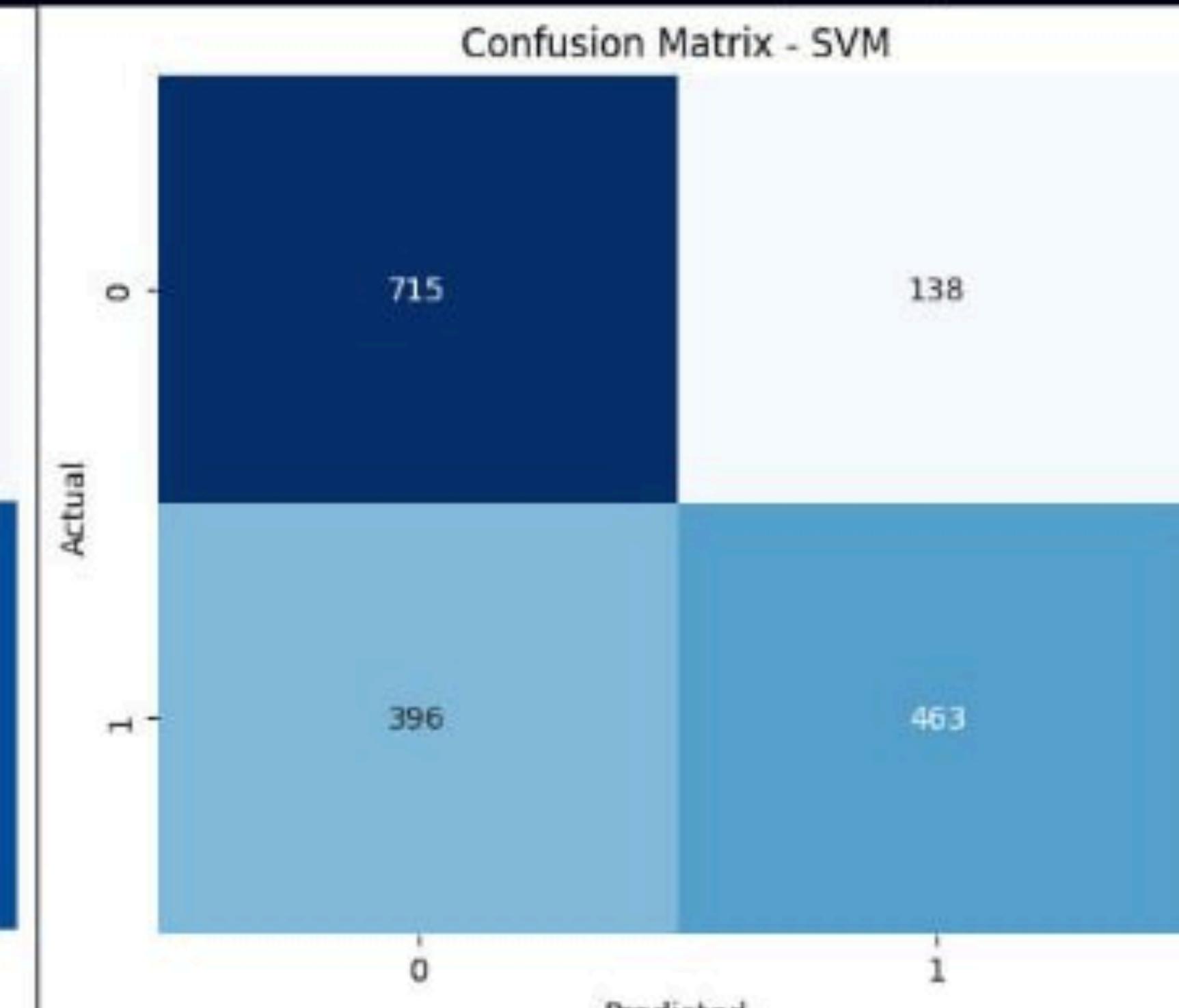
Обработка сложных взаимосвязей в данных.



3

SVM

Высокая точность на сложных границах.



Logistic Regression: Accuracy = 0.67

Random Forest: Accuracy = 0.69

SVM: Accuracy = 0.69

Наглядная классификация фильмов на хит/не хит

Используя два ключевых признака, показан результат разделения на хиты и не хиты с помощью логистической регрессии.

