Университет ИТМО

Практическая работа №4

по дисциплине «Визуализация и моделирование»

Автор: Севумян Наталья Ренатовна
Поток: ВИМ 1.1
Группа: К3221
Факультет: ИКТ
Преподаватель: Чернышева А.В.

Санкт-Петербург, 2021 г.

Описательная статистика по нормализованным данным

Гипотезы из предыдущей практической работы:

1. Зависимость рейтинга от количества проголосовавших - экспоненциальная.

2. Фильмы в жанре анимация более популярны, поэтому у сериалов и меньше общий рейтинг, потому как анимационных сериалов меньше.

3. Фильмы, добавленные на платформу гораздо позже их создания, имеют меньший рейтинг.

4. Контент с одним из жанров Family имеет высокие оценки.

5. Чем больше фильмов или сериалов снял определенный режиссер, тем популярнее контент.

Проверка гипотез

Гипотеза 1

Для проверки этой гипотезы была использована библиотека Plotly. Построен график зависимости рейтинга согласно IMDb от количества проголосовавших (рисунок 1).

Рис. 1: Зависимость рейтинга от количества проголосовавших

Видно, что от рейтинга 6 и выше зависимость похожа на экспоненциальную, но после 8 резко падает. Можно сказать, что гипотеза подтвердилась частично.

Гипотеза 2

Для проверки второй гипотезы сначала визулизировано отношение общего количества фильмов и сериалов к количеству фильмов и сериалов с жанром "Анимация"(рисунки 2 и 3.

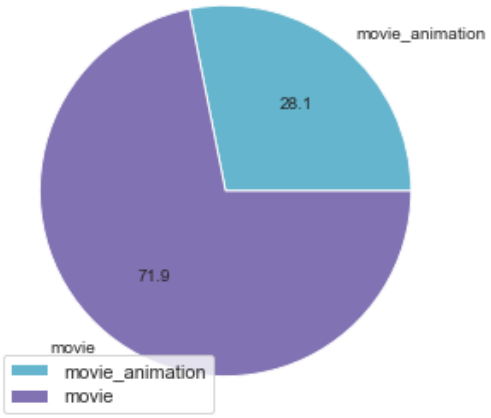

Рис. 2: Фильмы

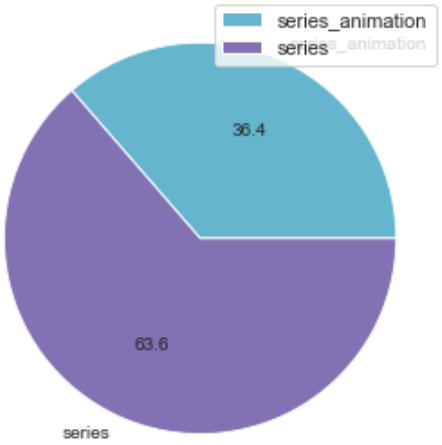

Рис. 3: Сериалы

Теперь,используя библиотеку matplotlib, визуализирую зависимость рейтинга от количества проголосовавших для фильмов и сериалов с жанром анимация и без него.

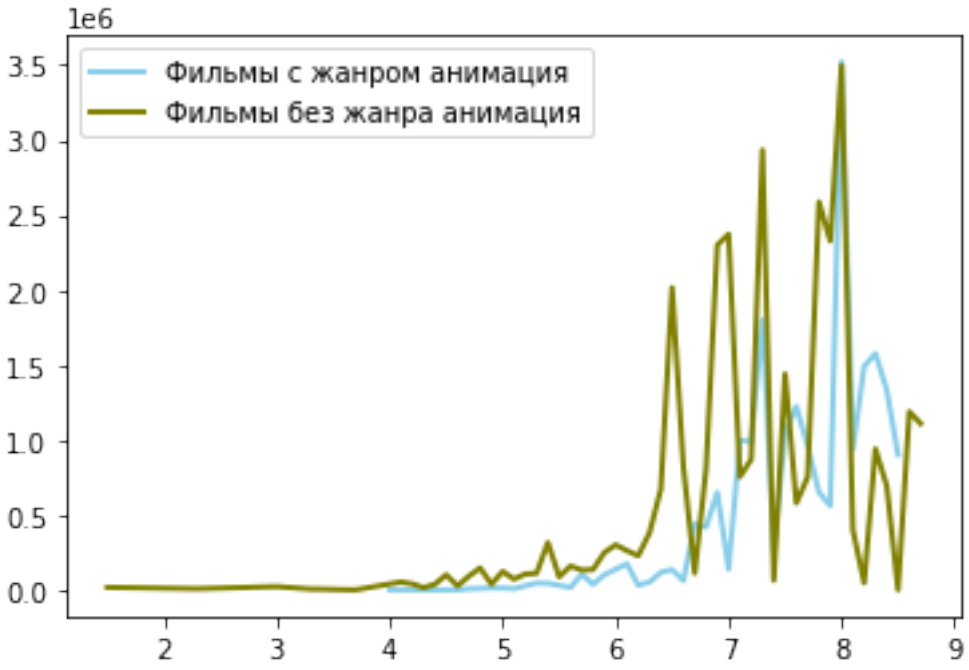

Рис. 4: Фильмы

Рис. 5: Сериалы

Из графиков видно, что сериалы с жанром анимация популярнее, чем без него. Однако для фильмов такая зависимость не работает. К тому же, соотношение сериалов с жанром анимация в процентном соотношении к общему количеству больше, чем фильмов. Можно сделать вывод, что гипотеза не подтвердилась, и популярность не зависит от присутствия жанра "Анимация".

Гипотеза 3

Для проверки третьей гипотезы в датафрейм добавлен еще один столбец "dif в котором записана разница между датой релиза и датой загрузки на платформу в днях. Затем с помощью модуля seaborn построена диаграмма рассеяния для зависимость рейтинга от разницы в днях (рисунок 6).

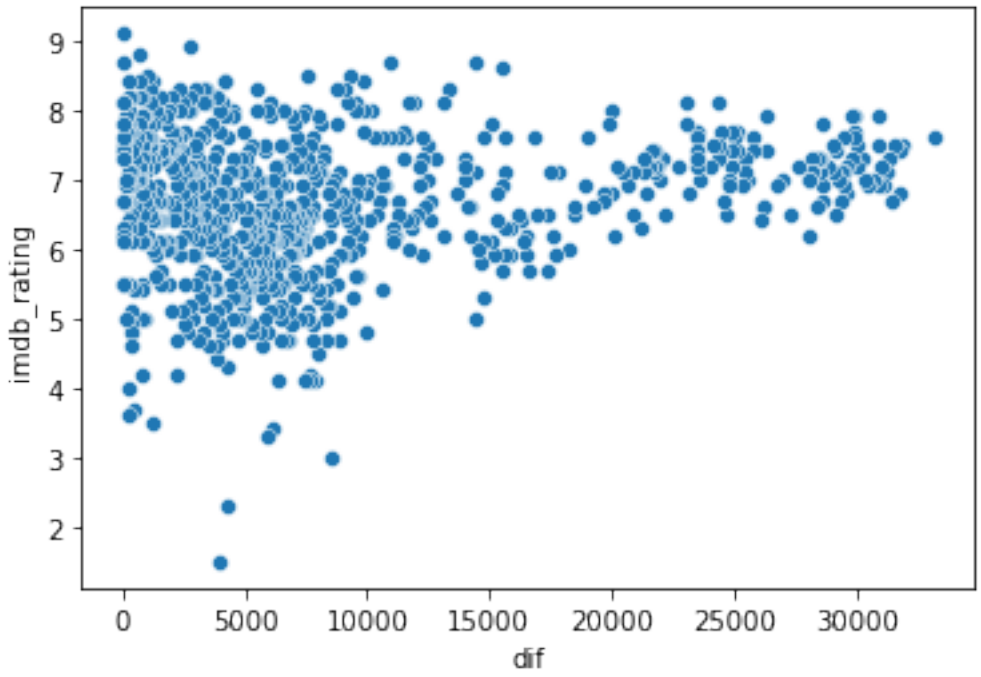

Рис. 6: Зависимость рейтинга от разницы в количестве дней

По графику видно, что гипотеза не подтвердилась. Да, самый высокий рейтинг у фильмов с минимальной разницей, но и фильмы с большой разницей в днях имеют довольно высокие оценки (на среднем уровне).

Гипотеза 4

С помощью модуля seaborn были визуализированы графики, показывающие количество контента по рейтингу с жанром "Family"и без него (рисунки 7 и 8).

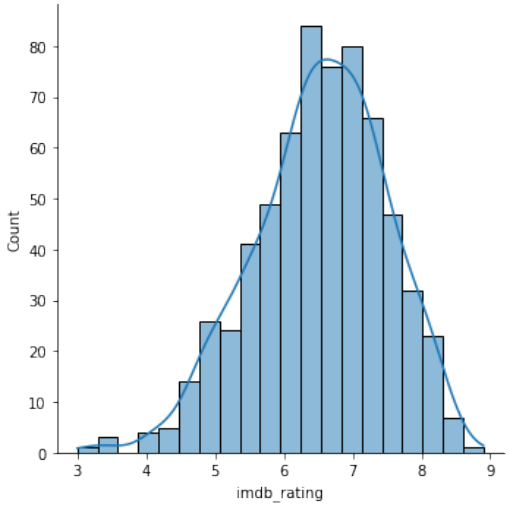

Рис. 7: Фильмы с жанром Family

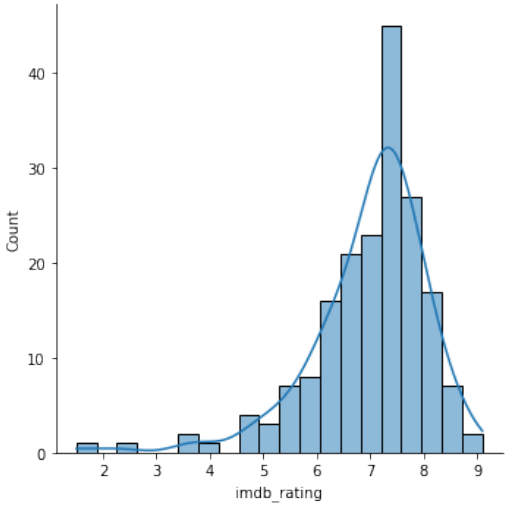

Рис. 8: Фильмы без жанра Family

По этим графикам видно, что рейтинг фильмов с этим жанром выше. Можно сделать общую визуализацию двух зависимостей (рисунок 9).

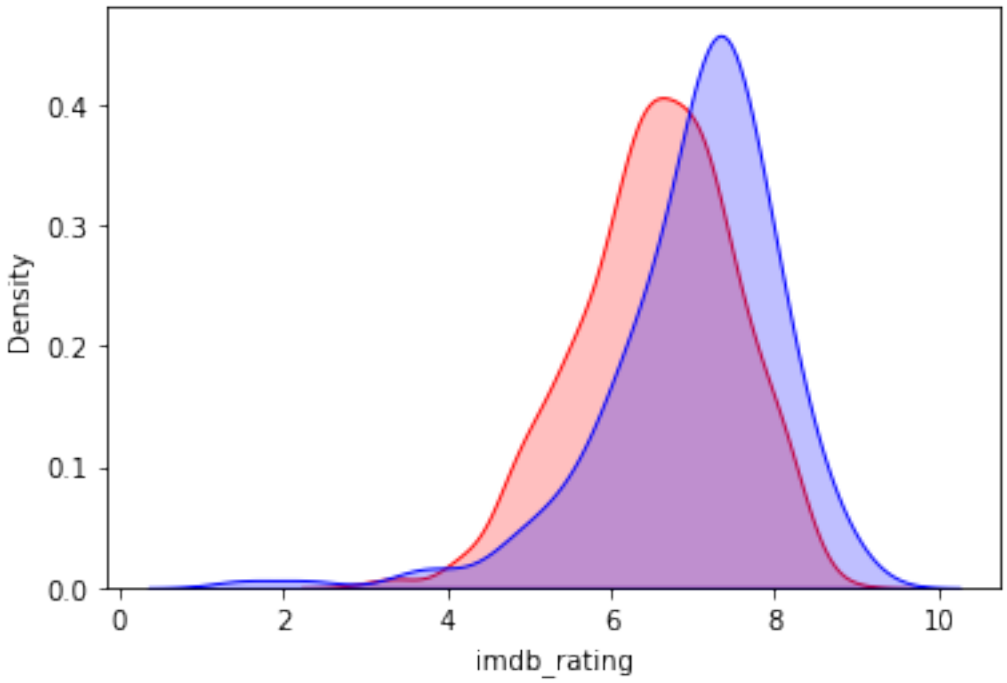

Рис. 9: Контент с жанром Family и без него

Гипотеза подтвердилась.

Гипотеза 5

Для проверки последней гипотезы был сформирован словарь с режиссерами, количество фильмов которых превышает 5. С помощью библиотеки Plotly был визуализирован средний рейтинг по фильмам этих режиссеров (рисунок 10).

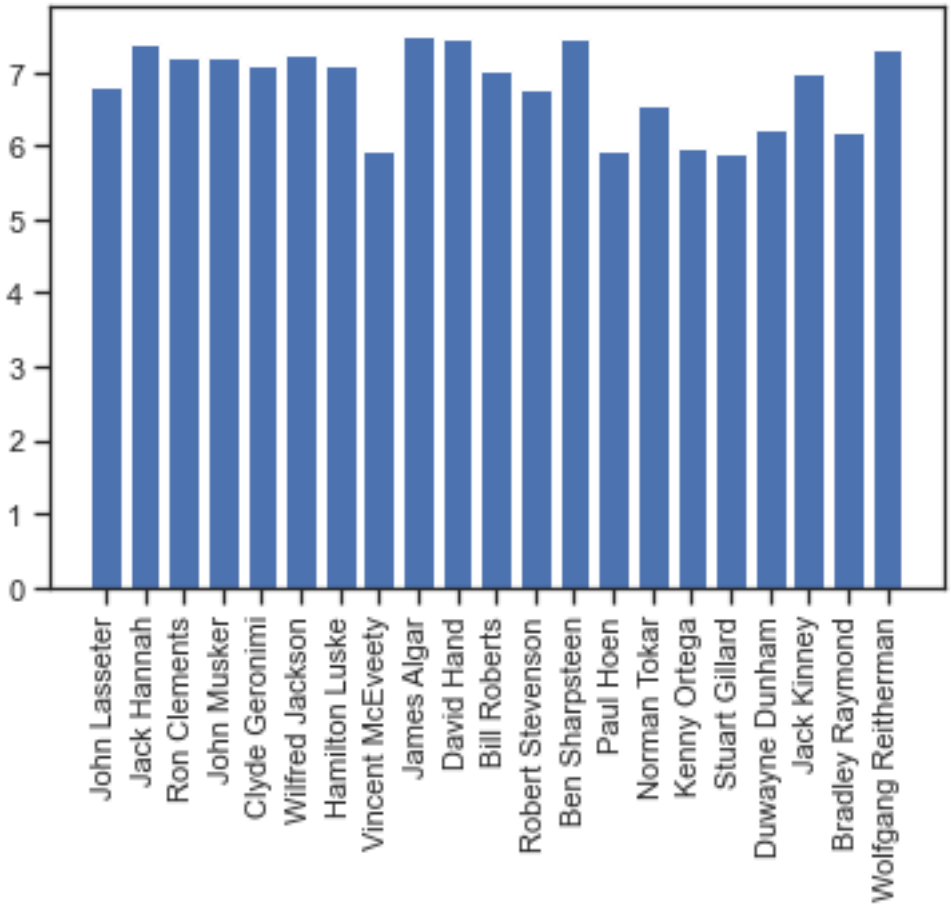

Рис. 10: Средний рейтинг фильмов режиссеров

По графику видно, что средний рейтинг у всех этих режиссеров довольно высок, что делает гипотезу верной.