Университет ИТМО

Практическая работа №2

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 найденного датасета

Таблица с описанием хранящихся в дасатете данных с указанием шкалы и исправленным столбцом "Тип данных".

Таблица 1: Описание колонок

Название	**Краткое описание**	**Тип данных**	**Тип шкалы**
imdb_id	Идентификатор каждого фильма согласно IMDb	str	Номинальная
title	Название фильма или сериала	str	Номинальная
plot	Сюжет фильма или сериала	str	Номинальная
type	Тип контента (фильм или сериал)	str	Номинальная
rated	Рейтинг шоу согласно системе рейтингов Американской киноассоциации	str	Порядковая
year	Годы, в которые выпускались эти кинокартины	str	Номинальная
released_at	Дата объявления о новом проекте	str	Номинальная
added_at	Дата публикации названия картины на сервисе Disney+	str	Номинальная
runtime	Длительность фильма или сериала	int	Относительная
genre	Жанр картины	str	Номинальная
director	Режиссер шоу	str	Номинальная
writer	Сценарист кинокартины	str	Номинальная
actors	Актеры, играющие главные роли	str	Номинальная
language	Язык, на котором можно посмотреть данное видео на этом сервисе	str	Номинальная
country	Страны, в которых возможен доступ к данному шоу	str	Номинальная
awards	Награды, которые получила кинокартина	str	Номинальная
metascore	Рейтинг согласно сайту Metacritic	float	Интервальная
imdb_rating	Оценка согласно IMDb	float	Интервальная
imdb_votes	Количество проголосовавших за данную картину на IMDb	int	Относительная

Визуализация данных

Изначально в датасете всего 2 столбца с числовыми данными - metascore и imdb_rating, но metascore содержал слишком много пустых ячеек, поэтому работать с этими данными было бессмысленно. Поэтому для визуализации я отформатировала несколько столбцов (year, imdb_votes).

Первое, что нужно было узнать - соотношение фильмов и сериалов на платформе, чтобы понимать, с каким контентом будем иметь дело чаще. Результат представлен на рисунке 1.

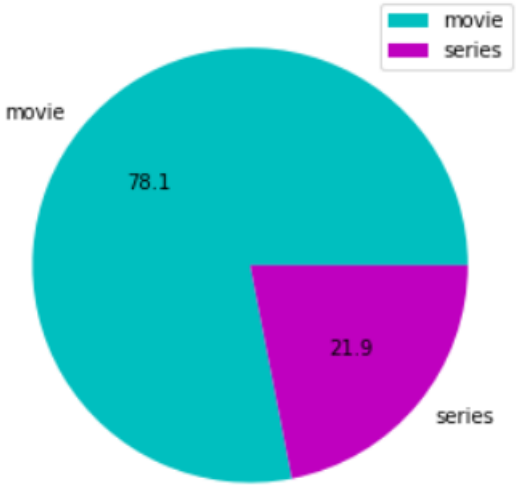

Рис. 1: Соотношение фильмов и сериалов

Видно, что большую часть каталога составляют фильмы, поэтому большую часть буду работать именно с этим типом контента.

Теперь посмотрим, как колеблется число фильмов (рисунок 2) и сериалов (рисунок 3) по разным IMDb рейтингам в группе интерквантильного размаха. Предполагаю, что фильмов с высоким рейтингом больше, чем сериалов.

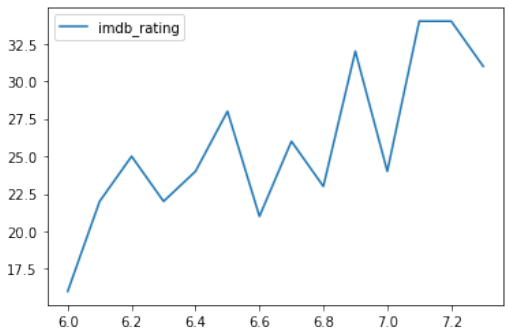

Рис. 2: Количество фильмов в зависимости от рейтинга

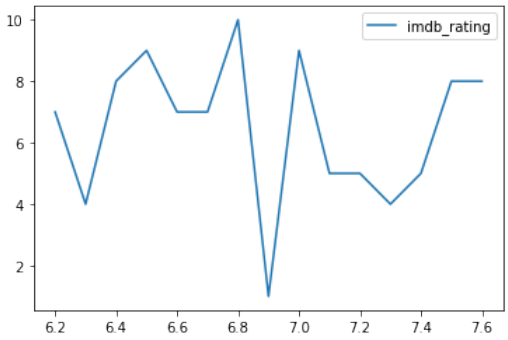

Рис. 3: Количество сериалов в зависимости от рейтинга.

Из графиков видно, что рейтинг выше среднего имеет большая часть фильмов, а у сериалов самый распространенный рейтинг около 6.8. При этом, сравнивая оба графика, можно увидеть, что почти все фильмы расположены в диапазоне от 6 до 7.2, тогда как сериалы - от 6.2 до 7.6.

Составлю процентное соотношение фильмов и сериалов в разных группах оценок (рисунок 4).

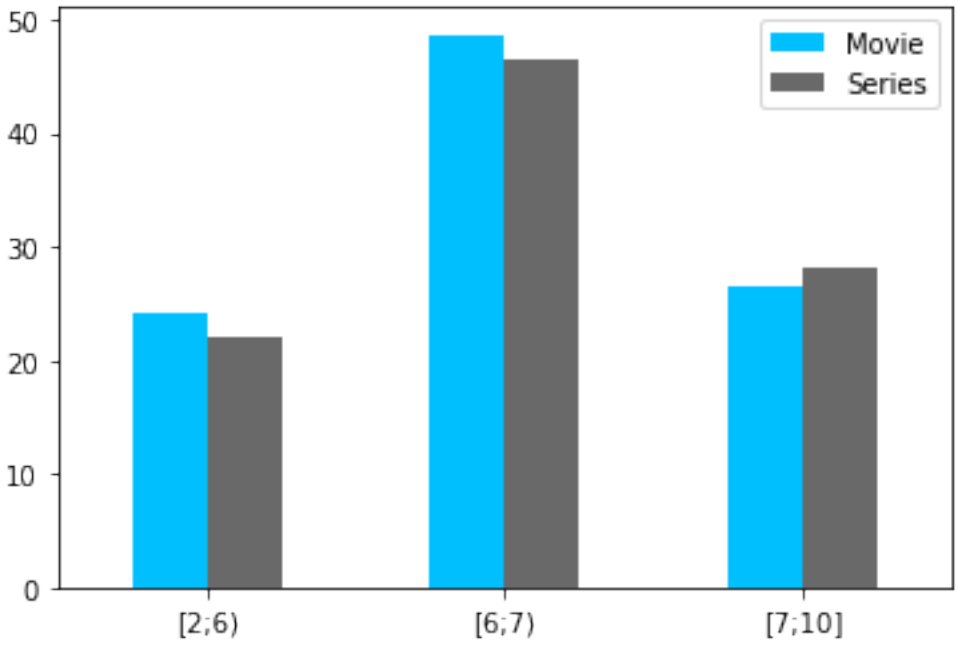

Рис. 4: Соотношение фильмов и сериалов по IMDb рейтингам

Видно, что в процентном соотношении у фильмов и сериалов примерно одинаковые позиции, однако чуть большее количество сериалов получают высокие рейтинги.

Теперь визуализирую распределение контента по рейтингам согласно системе Американской киноассоциации, которая показывает для какой аудитории предназначен фильм (рисунок 5). Составлю два графика - с категорией "UNRATED"(рейтинг не указан) и без нее.

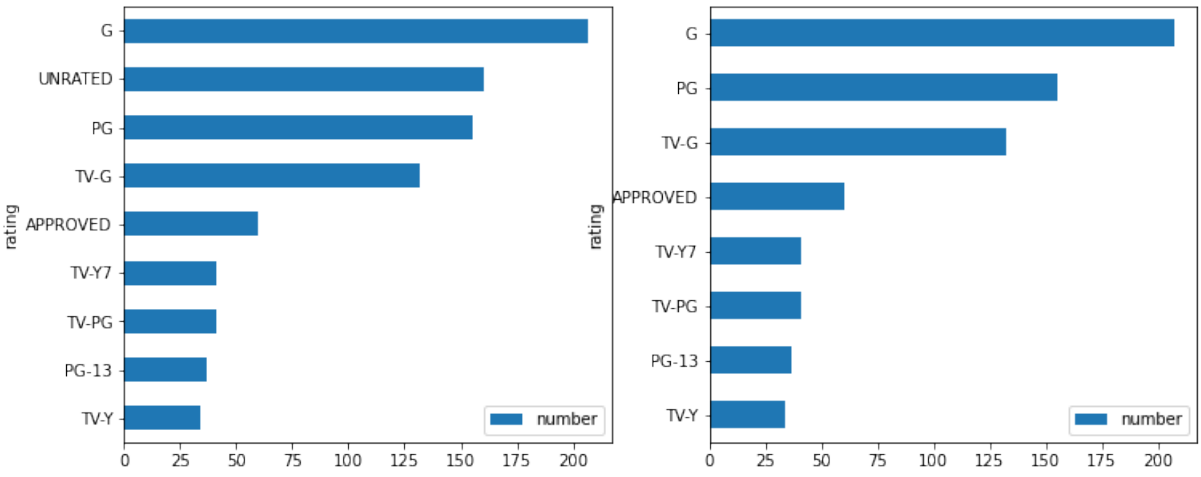

Рис. 5: Количество шоу по разным категориям

Как и ожидалось, большинство шоу будет отнесено к категории "G"и "PG так как Disney - это детский сервис.

Теперь наложу на эту визуализацию распределение по типу контента (рисунок 6).

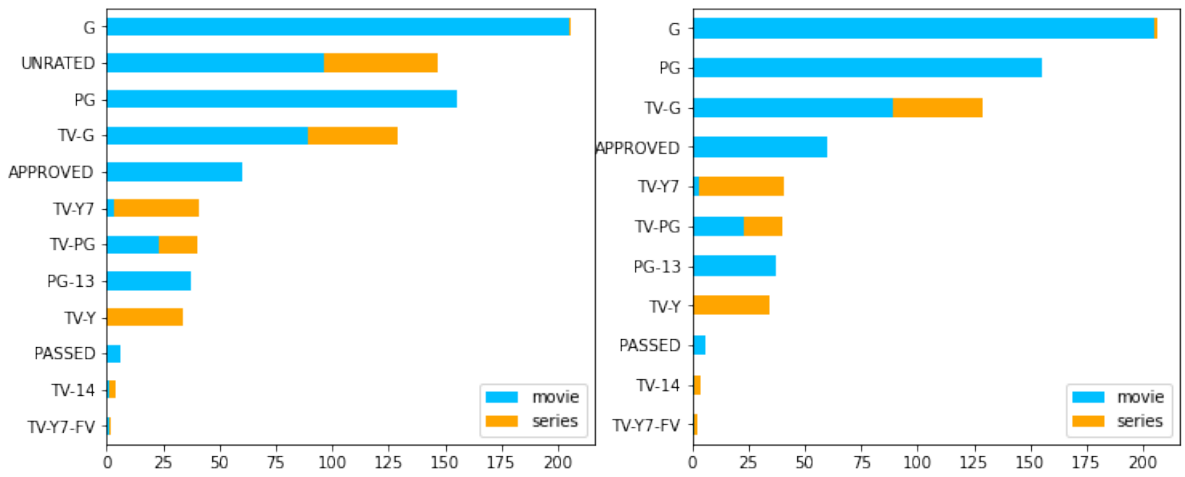

Рис. 6: Количество шоу по разным категориям с наложением его типа

Из такого распределения видно, что есть категории, в которых присутствует только один из типов контента, например, рейтинги PG или TV-Y.

Теперь построю график, демонстрирущий количество фильмов, выпущенных с течением времени, чтобы посмотреть, в какие года Дисней был наиболее активен (рисунок 7).

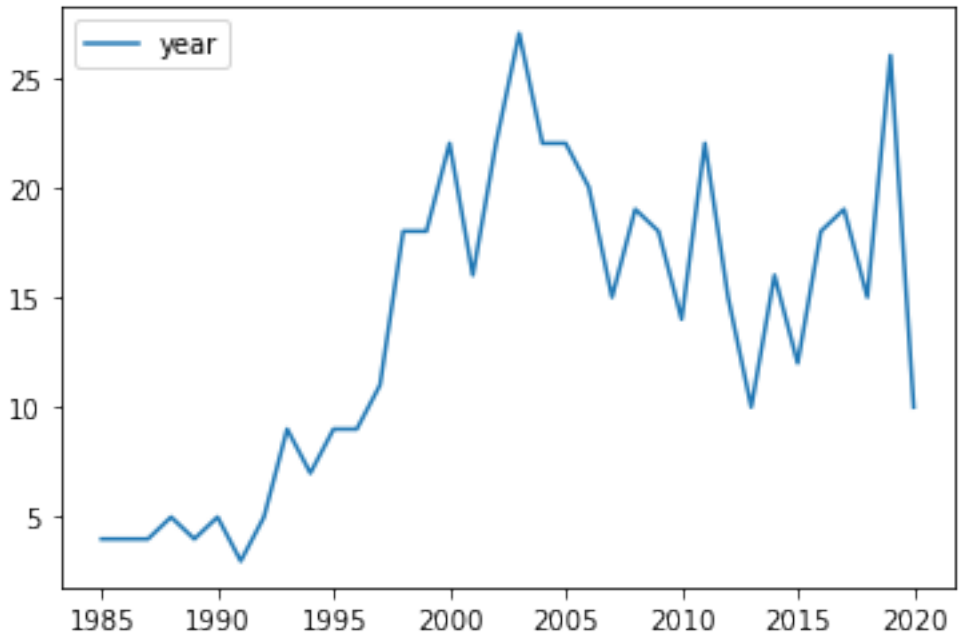

Рис. 7: Количество фильмов по годам

Ожидаемым результатом было, что в самом начале своей деятельности "Дисней"выпускал мало фильмов. Пики активности пришлись приблизительно на 2005 и 2018 год.

Затем визулизирую зависимость средней продолжительности фильма от года выпуска (рисунок 8).

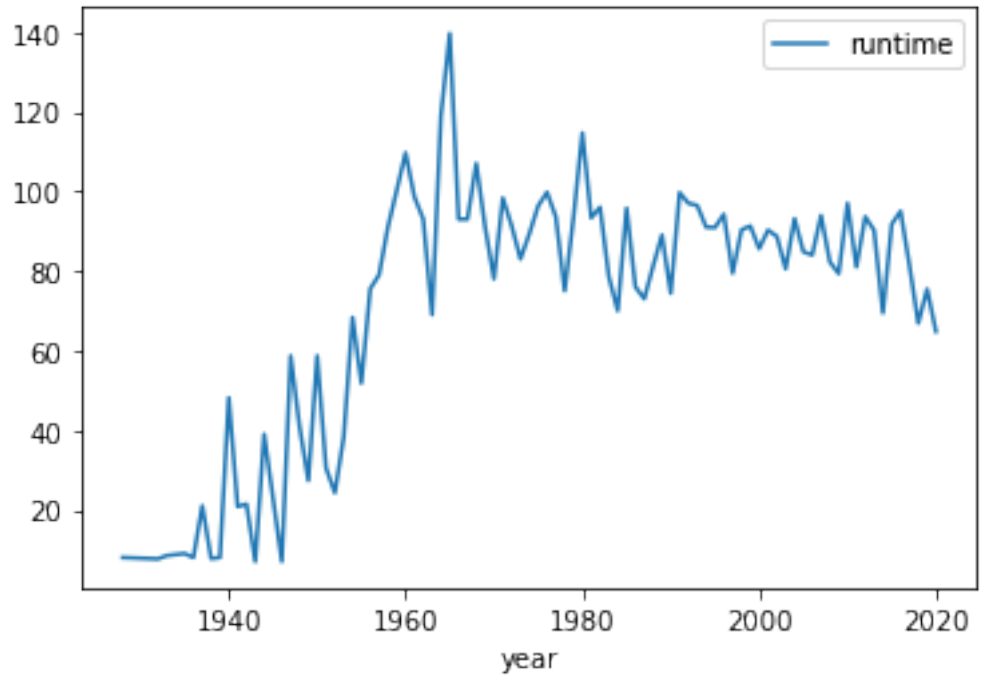

Рис. 8: Длительность фильмов по годам

График показывает, что изначально (из-за отсутствия техники) фильмы были короткометражными, в определенный период (1950-1970) стали увеличивать хронометраж, а затем пришли к оптимальной длине фильмов - около 80 минут.

Попробую оценить зависимость IMDb рейтинга фильма от его длительности (рисунок 9).

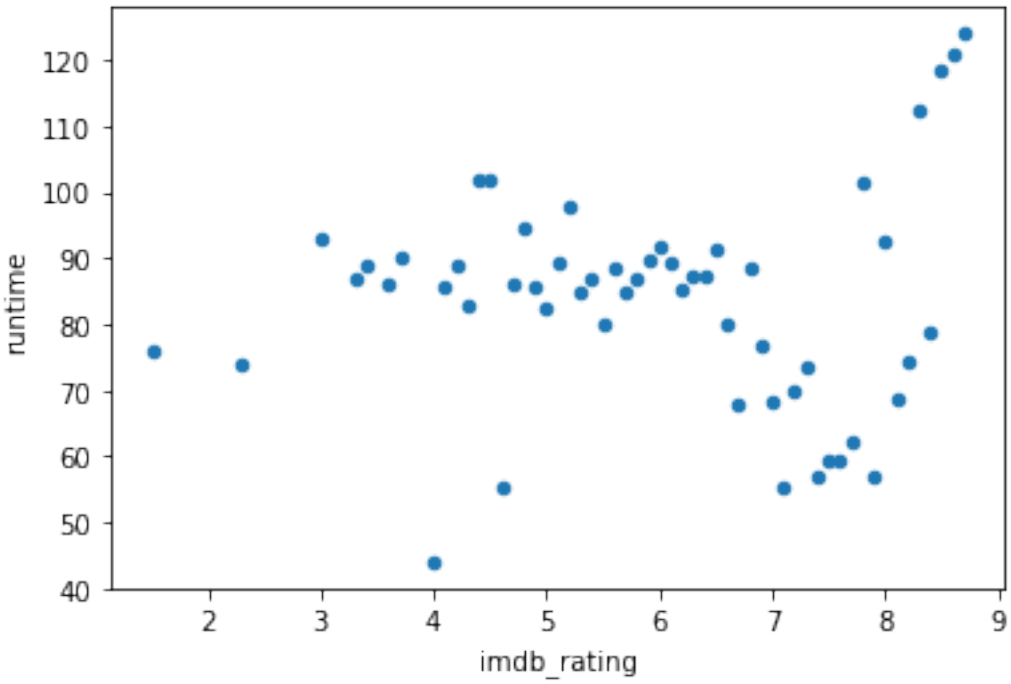

Рис. 9: Рейтинг фильма в зависимости от длительности

Этот график не дает сделать каких-то конкретных выводов, но показывает, что

максимальные оценки у самых продолжительных фильмов.

Попробую изучить зависимость IMDb рейтинга от количества проголосовавшихт(рисунок 10).

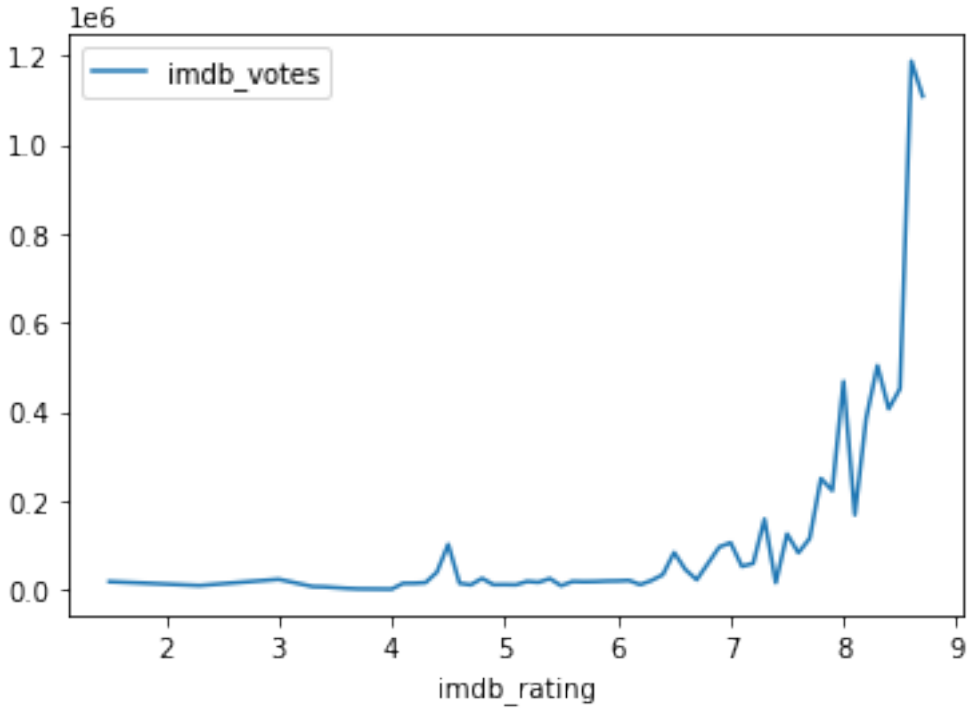

Рис. 10: Рейтинг фильма в зависимости от количества проголосовавших

Из-за большого разброса значений сложно делать какие-то выводы, но по графику видно, что чем больше проголосовавших, тем выше рейтинг. То есть массовая аудитория оставляет в основном только положительные отзывы.