**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра САПР**

**отчет**

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

**Тема: Кластеризация**

|  |  |  |
| --- | --- | --- |
| Студент гр. 2323 |  | Овчинников Л.Н. |
| Преподаватель |  | Татчина Я.А |

Санкт-Петербург

2024

**Цель работы**

Познакомиться с методом кластеризации K-mean с помощью пакета sklearn

1. **Краткое описание набора данных**

Данные о фильмах в онлайн кинотеатре Amazon Prime. Датасет был взят с сайта www.kaggle.com Этот набор данных содержит фильмы, доступные на сайте Amazon.com.

В датасете имеются числовые данные. В датасете представлены следующие атрибуты:

* title (название фильма) тип данных строковый
* Movie Rating (рейтинг фильма) тип данных числовой
* No\_of\_Ratings (количесвто оценок) тип данных числовой
* ReleaseYear (год релиза фильма) тип данных числовой
* MPAA\_Rating (Возрастной рейтинг фильма) тип данных строковый
* Directed\_By (Режиссёр фильма) тип данных строковый
* Starring (Каст фильма) тип данных строковый
* Price (цена фильма) тип данных числовой

1. **Добавить новый атрибут**

В ходе работы был добавлен новый атрибут подсчета суммарных продаж на основе количества выставленных оценок, умноженных на цену за фильм.

1. **Обработка данных датасета**

В ходе работы была произведена обработка данных.

Дубликатов в наборе данных не оказалось.

Были обнаружены пропущенные значения был в атрибутах MPAA\_Rating, Starring, Directed\_By, Price, ReleaseYear и соответственно обработаны. В атрибутах Starring и Directed\_By значение NaN было заменено на NoName. В Price на значение 0. В MPAA\_Raiting на значение NoRate. И в ReleaseYear на значение 2000.

В наборе данных были обнаружены выбросы. График выбросов представлен на рисунке 1.

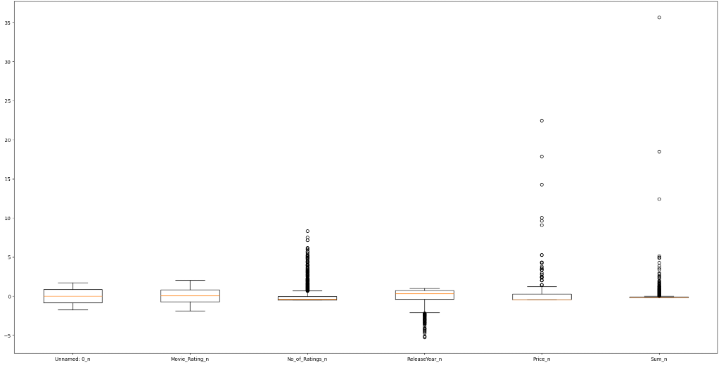


Рисунок 1 - Выбросы в наборе данных

Также были найдены нулевые значения. Аномальные и нулевые значения сначала были опустошены, потом с помощью k-метода ближайших соседей. График выбросов представлен на рисунке 2.

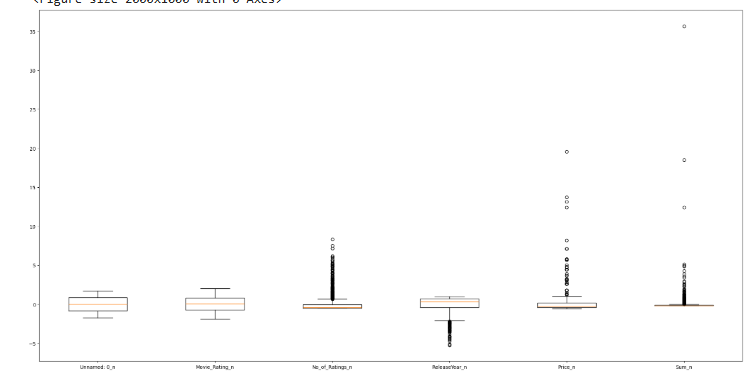


Рисунок 2 - Выбросы в наборе данных

1. **Построить графики зависимости одной переменной от другой**

В ходе работы были рассмотрены 4 зависимости: средних оценок от общего количества выставленных,

Были рассмотрены зависимости четырех атрибутов: Рейтинг фильма, количество оценок, год выхода и цена. График представлен на рисунке 3.

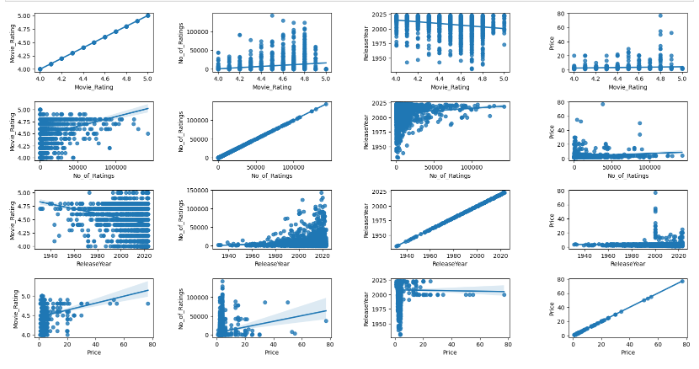


Рисунок 3 – График зависимостей

Представленные данные можно разбить на группы по возрастному рейтингу. Были построены те же графики, но в зависимости от возрастного рейтинга. Графики представлены на рисунках 4–9

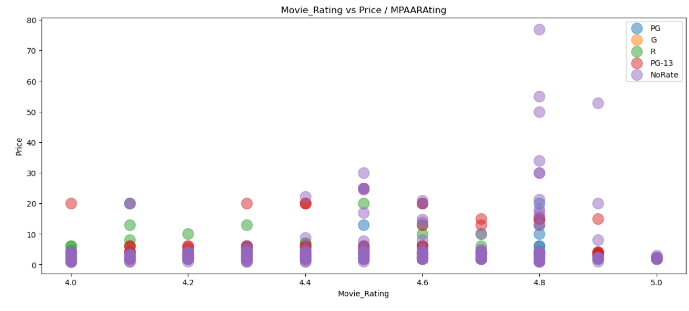


Рисунок 4 – График зависимости оценки фильма от цены в возрастном рейтинге

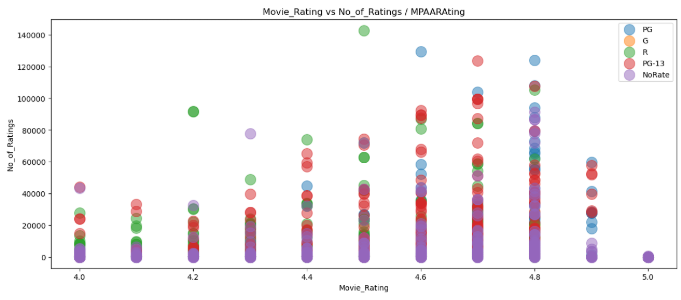


Рисунок 5 - График зависимости оценки фильма от количества оценок в возрастном рейтинге



Рисунок 6 - График зависимости цены от года выхода в возрастном рейтинге

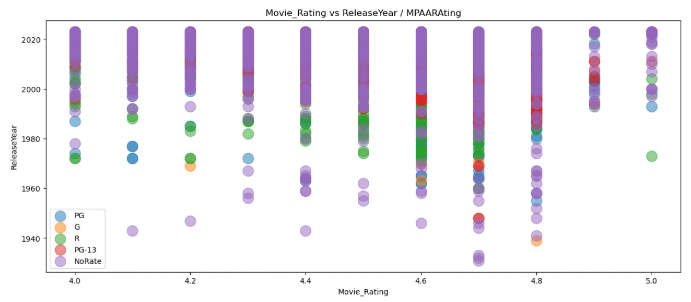


Рисунок 7 - График зависимости оценки фильма от года выхода в возрастном рейтинге

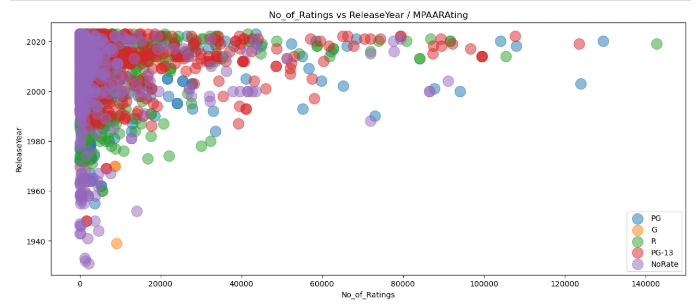


Рисунок 8 - График зависимости количества оценок от года выхода в возрастном рейтинге

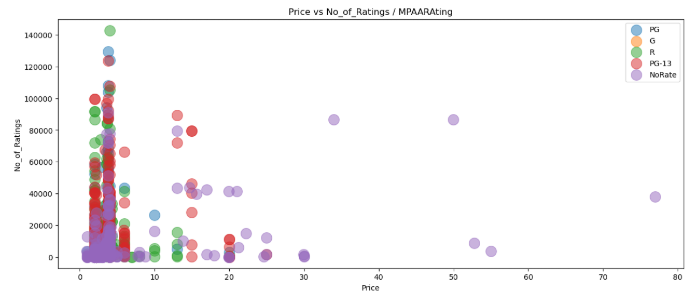


Рисунок 9 - График зависимости цены от количества оценок в возрастном рейтинге

Из указанных графиков следует что фильмы, которые не обладают возрастным рейтингом, имеет большое преимущество перед фильмами с возрастным рейтингом в этом наборе данных.

1. **Метод KMeans**

Графики изображены на рисунках 10 -13

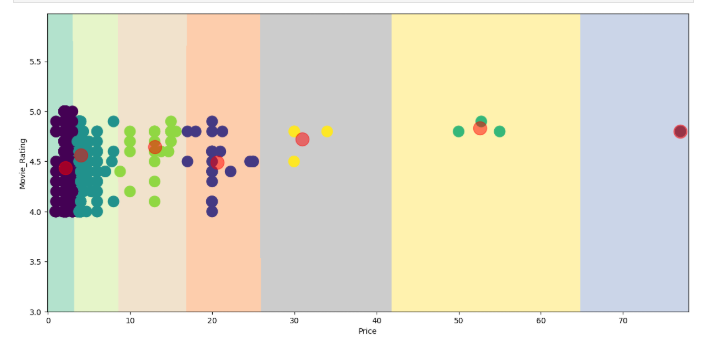


Рисунок 10 – Кластеризация с использованием оценки фильма и цены

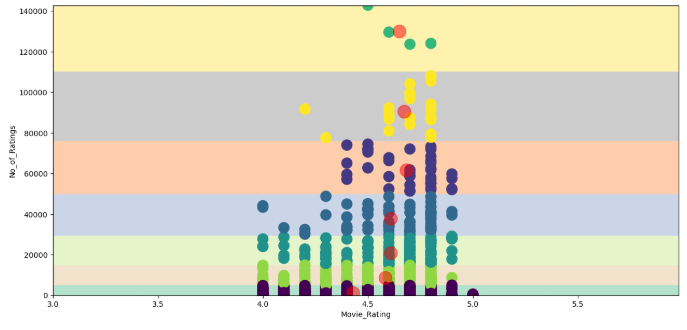


Рисунок 11 - Кластеризация с использованием количества оценок и оценка фильма

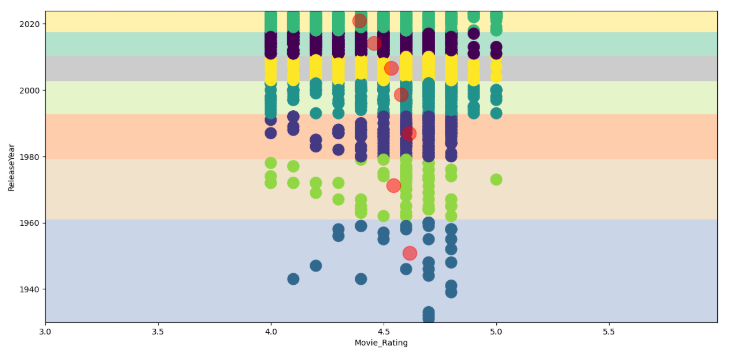


Рисунок 12 - Кластеризация с использованием года выхода и оценки фильма

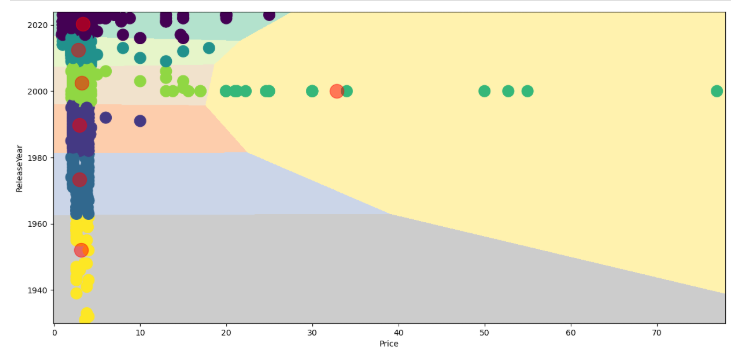


Рисунок 13 - Кластеризация с использованием года выхода и цены

**Выводы**

В ходе работы была проведена первичная обработка набора данных: были отредактированы аномальные значения и нулевые значения. Изучен k метода ближайших соседей. Также были обнаружены пропущенные значения и заменены на новые значения по умолчанию. Дубликатов в представленном наборе данных не оказалось.

Также были построены графики зависимостей в зависимости от возрастного рейтинга, по результат которым можно сделать вывод что в данном наборе данных фильмы, которые не имеют возрастного рейтинга намного больше, чем фильмов с выставленным возрастным ограничением.

Был изучен набор KMeans, с помощью которого разделили набор данных на кластеры.