**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра САПР**

**отчет**

**по лабораторной работе №1**

**по дисциплине «Машинное обучение»**

**Тема: Исследование набора данных**

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

Санкт-Петербург

2024

**Цель работы**

Исследовать алгоритмы кластеризации и классификации в набор данных.

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. **Определение параметров**
   1. Среднее значение, СКО

Среднее значение и СКО атрибутов были определены с помощью функций библиотеки numpy ‘np.mean’ ‘np.std’.

Price: среднее = 2,31, СКО = 5,23

No of Ratings: среднее = 8090.98, СКО = 16153.64

Release Year: среднее = 1971 , СКО = 271

Movie Rating: среднее = 4.484, СКО = 0.255

* 1. Построить гистограммы распределения значений, определить если ли выбросы:

Поиск выбросов был осуществлён с помощью следующего алгоритма:

Отсортировать данные и найти num1 и num2 и найти размах между ними.

Проверить какие наблюдения вышли за границы

Наличие выбросов:

Price: Да

No of Ratings: Да

Release Year: Да

Movie Rating: Да

Гистограммы распределений представлены ниже:

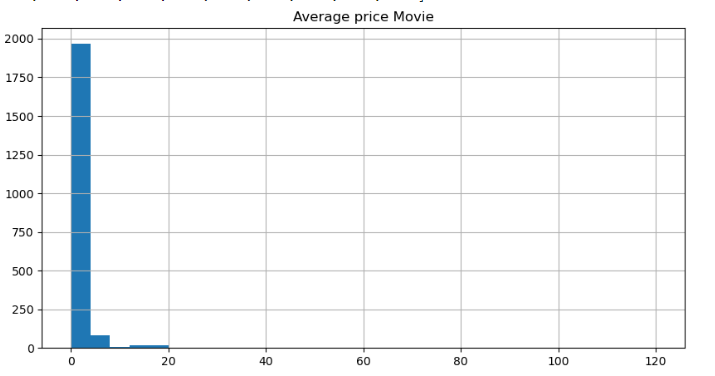


Рисунок 2.1. – Распределение средней цены за фильм

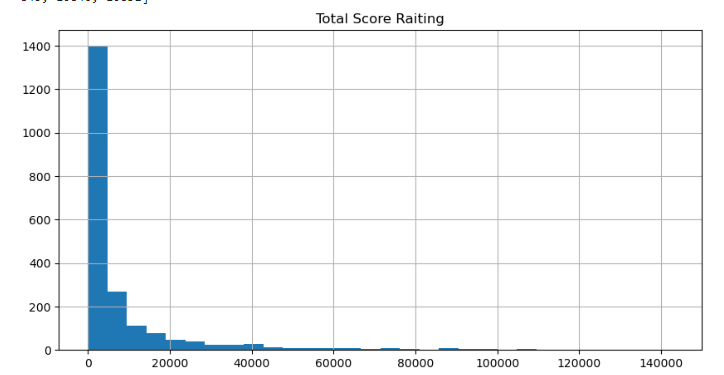


Рисунок 2.2. – Распределение всего оценок за фильмы

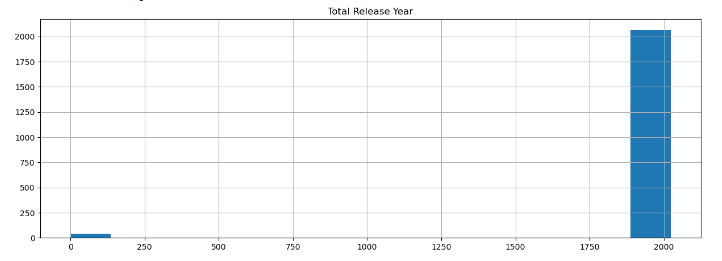


Рисунок 2.3. – Распределения года выпуска фильма

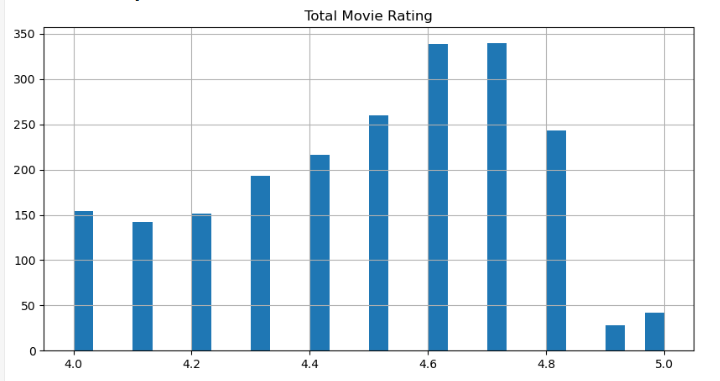


Рисунок 2.4. – Распределение оценки за фильм

* 1. Определить если пропущенные значения

Наличие пропущеных значений определялось с помощью функции df[‘название атрибута’].isna().sum()

В ходе работы были обнаружены пропущенные данные у двух атрибутов.

* 1. Предложить вариант обработки пропущенных значений

Так пропущенные значения были у атрибутов Price и ReleaseYear было обнаружено что у этих атрибутов пропущенные данные имели пропуск и обозначались NaN. Для замены значения NaN на 0 была использована функция df.fillna().В ранее указанных вычисления функция используется.

1. **Определение корреляции**

В ходе работы были рассмотрены 4 зависимости: средних оценок от общего количества выставленных, средней цены за фильм от даты релиза, выставленных средних оценок от года выхода фильма, оценок фильма от цены за фильм.

* 1. Определить, какие атрибуты высокоррелированы и характер корреляции

Для выполнения задания на потребуется воспользоваться функцией np.corrcoef()

* средних оценок от общего количества, выставленных: коэффициент = 0.25, чем меньше оценок, тем более целая общая оценка фильма, а чем больше то оценка с плавающих значением после запятой.
* средней цены за фильм от года выхода: коэффициент = - 0.47, можем заметить, что у некоторых фильмов не проставлена дата выпуска в представленном наборе данных из-за этого и коэффициент уходит в отрицательное значение
* выставленных средних оценок от года выхода фильма: коэффициент = - 0.061, можем заметить, что у некоторых фильмов не проставлена дата выпуска в представленном наборе данных из-за этого и коэффициент уходит в отрицательное значение.
* оценок фильма от цены за фильм: коэффициент = 0.13, больше всего выставляют оценки фильмам ниже 30 долларов.
  1. Какие атрибуты не имеют корреляцию

Из исследованных атрибутов все имеют корреляцию.

* 1. Построить графики рассеивания

Графики рассеивания представлены на рисунках 3.1–3.4

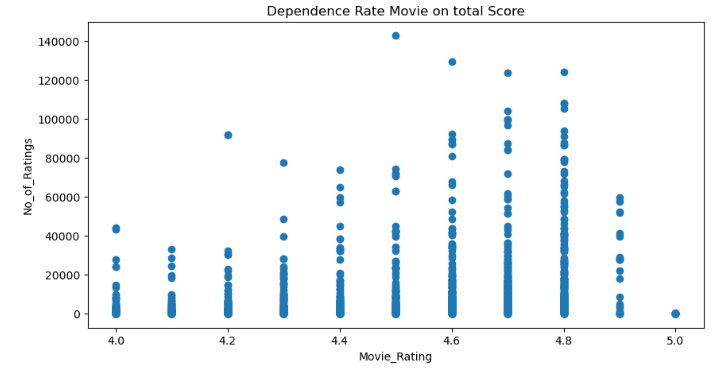


Рисунок 3.1. – Зависимость средних оценок от общего количество выставленных

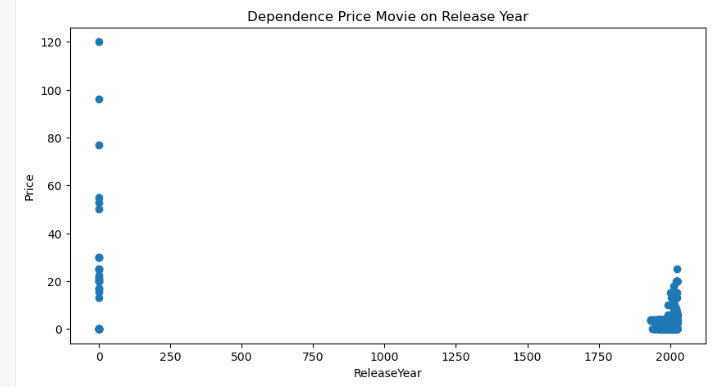


Рисунок 3.2. – Зависимость средней цены за фильм от года выхода

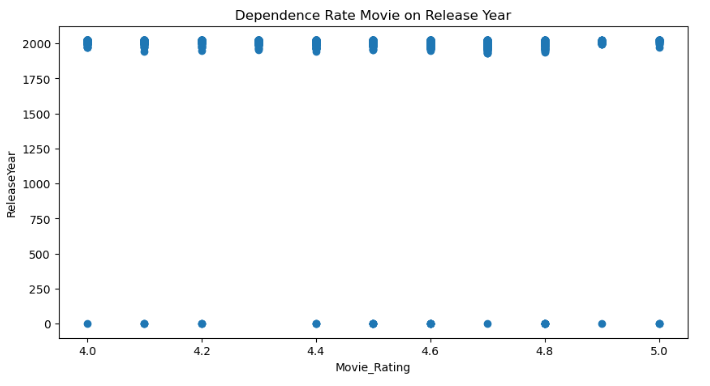


Рисунок 3.3. – Зависимость выставленных средних оценок от года выхода фильма

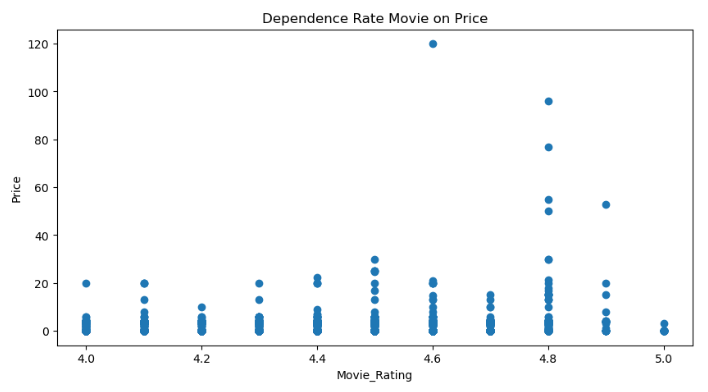


Рисунок 3.4. –Зависимость оценок фильма от цены за фильм

* 1. Проанализировать полученные результаты

Из полученных результатов можем сделать вывод, что в текущем наборе данных представлены фильма с оценкой 4 и до 5. И в основном фильмы, вышедшие в период со 1950 до 2023.