**МИНОБРНАУКИ РОССИИ**

**Санкт-Петербургский государственный**

**электротехнический университет**

**«ЛЭТИ» им. В.И. Ульянова (Ленина)**

**Кафедра информационных систем**

отчет

**по лабораторной работе №2**

**по дисциплине «Машинное обучение»**

Тема: Метод Кластеризации K-Mean

|  |  |  |
| --- | --- | --- |
| Студент гр. |  | Полуянов В. Н. |
| Преподаватель |  | Татчина Я.А. |

Санкт-Петербург

2024

**Цель работы.**

Знакомство с методом кластеризации K-mean с помощью пакета sklearn.

**Постановка задачи.**

1. Добавить к выбранному набору данных новый атрибут, вычислив его из имеющихся;
2. Очистить данные, удалив выбросы и дубликаты; обработать пропущенные значения; найти искаженные данные;
3. Применить метод кластеризации K-mean к текущему набору данных
4. Проанализировать полученные результаты.

**Выполнение работы.**

Для данной лабораторной работы был выбран новый набор данных, содержащий в себе более 7000 записей о фильмах и их доходах за промежуток с 1986 по 2020 года по данным ресурса IMDb.

Были созданы новые атрибуты: год выпуска и прибыль. Год выпуска заменяет столбцы «год» и «выход в прокат», а именно – получает значение года из поля «выход в прокат» и заменяет атрибут «год» полностью, ведь год производства не так важен, как год первого показа. Прибыль считается как разница дохода и бюджета, является новым атрибутом.

Все поля в файле были записаны как вещественные, что не несёт смысла для полей «год», «количество голосов», «бюджет» и т.п., тип данных этих столбцов был заменён на целочисленный.

Далее, выяснилось, что бюджет у четверти фильмов отсутствует. Для заполнения пустых полей были взяты медианные значения бюджета для каждого года. Все записи с пустым полем «доход» были удалены (189 записей), ведь они не могут быть использованы в дальнейшем. Оставшиеся записи иногда не имели информации о возрастном рейтинге, продолжительности и оценке, но таких примеров было относительно мало (не больше 20 единиц). Они тоже были удалены.

После предобработки данных был проведён анализ атрибутов и их визуализация. Так, распределение возрастных рейтингов можно увидеть на рисунке 1.

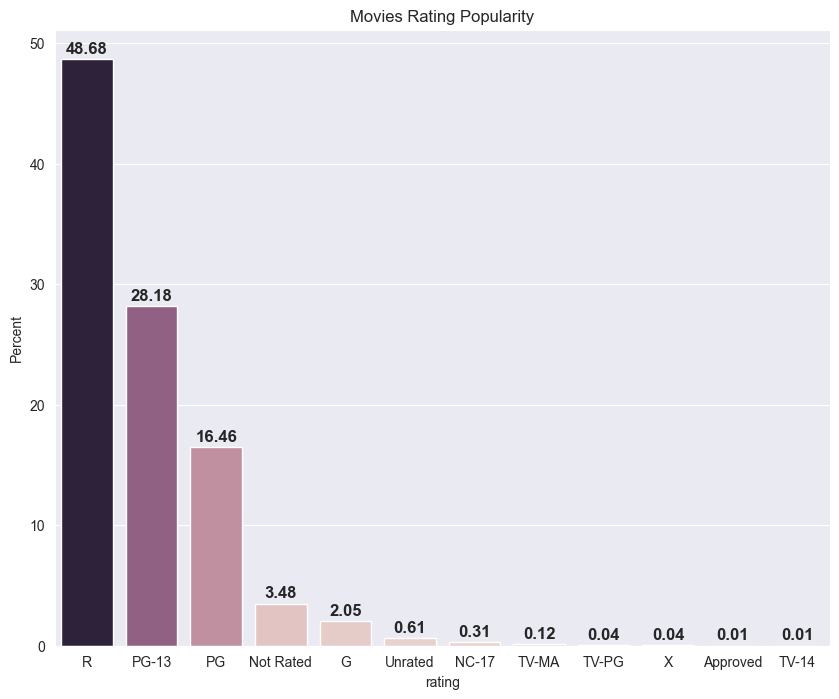


Рисунок . Распределение возрастных ограничений

Среди жанров лидируют комедия, экшн, драма (рис. 2).

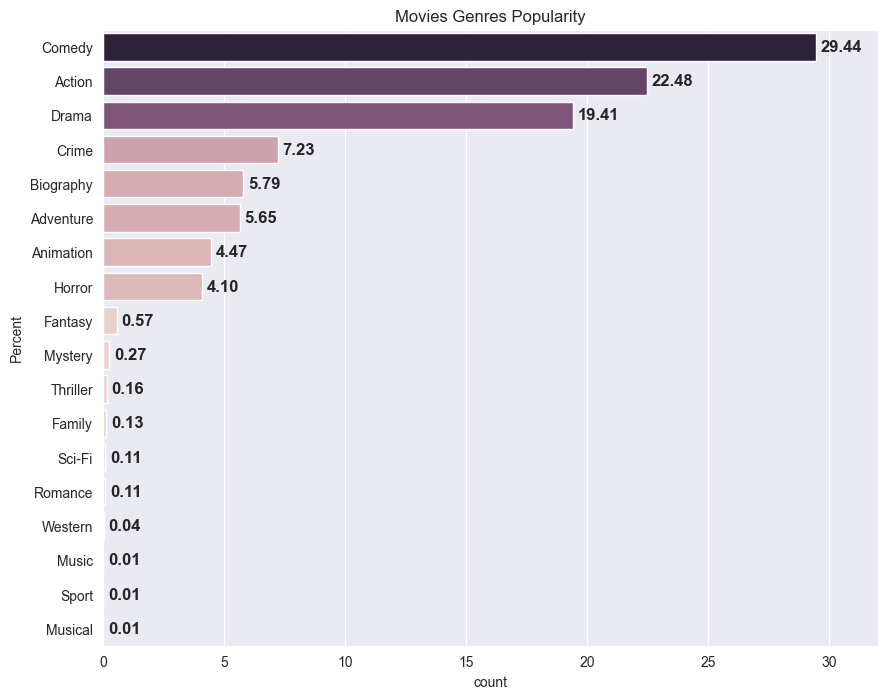


Рисунок . Распределение жанров

Большинство фильмов сняты в США (рис. 3). Далее, с большой разницей, идут Великобритания, Франция, Канада.

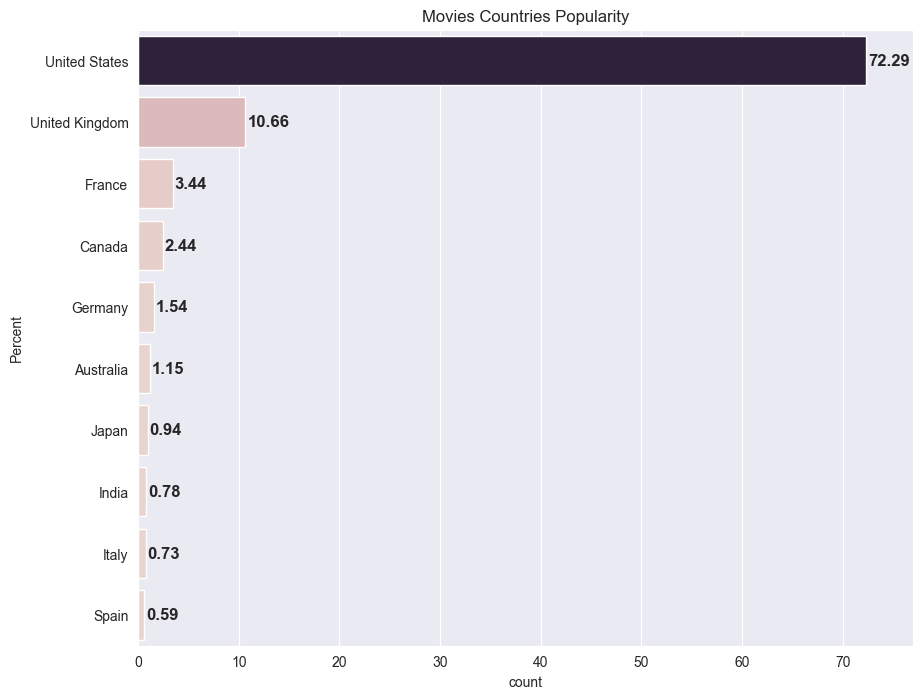


Рисунок . Распределение стран

Так же стоит выделить распределение оценок фильмов. График смещён вправо, что говорит о желании зрителей ставить более высокие оценки, причём большинство оценок находятся в диапазоне от 5.5 до 7.5 со средним значением 6.5 (рис. 4)

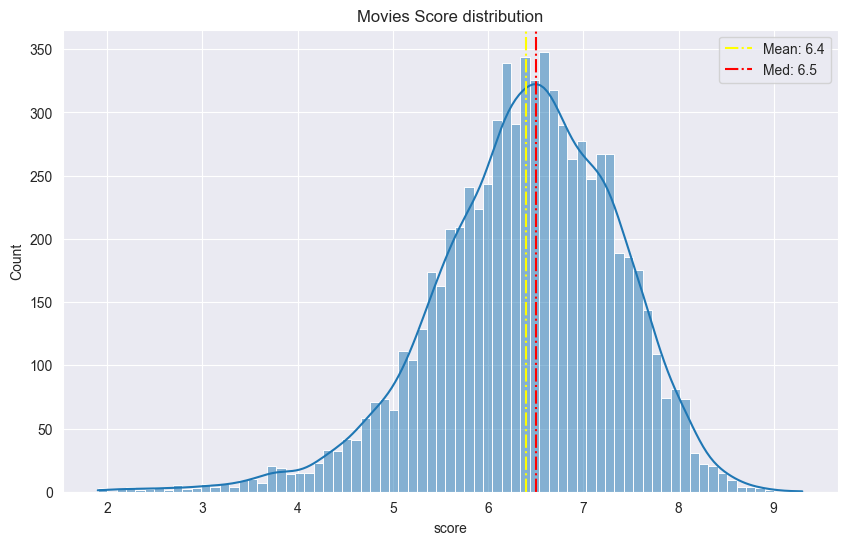


Рисунок . Распределение оценок

Самые «продуктивные» режиссёры в данном наборе: Woody Allen: 38, Clint Eastwood: 31, Steven Spielberg: 27, Ron Howard: 24.

Больше всего сценариев написано: Woody Allen: 37, Stephen King: 31, Luc Besson: 25, John Hughes: 25.

Больше всего фильмов выпущено компаниями: Universal Pictures: 376, Columbia Pictures: 332, Warner Bros.: 332.

Были обнаружены две сильные зависимости. Доход коррелирует с бюджетом на 71% и на 63% с количеством голосов, что ожидаемо. Чем больше вложено денег в съёмку, тем более ожидаемы высокие доходы. Чем больше зрителей оставило отзыв, тем больше показов, тем больше доходы.

Для кластеризации методом «локтя» было выбрано число k = 4. После применения метода было выявлено вполне объяснимое распределение по группам, а именно:

1. Наименьший кластер фильмов с наилучшими финансовыми показателями;
2. Два наибольших кластера с хорошими показателями. Разница между средними доходами в этих группах – один десяток;
3. Небольшой кластер фильмов с плохими показателями доходности. Сюда попало множество старых фильмов.

**Выводы.**

В ходе работы были изучены и обработаны данные киноиндустрии за 1986-2020 года. Были выявлены зависимости и проведена кластеризация. В результате были получены навыки кластеризации с помощью метода K-means.