ГУАП

КАФЕДРА № 41

ОТЧЕТ   
ЗАЩИЩЕН С ОЦЕНКОЙ

ПРЕПОДАВАТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| старший преподаватель |  |  |  | В.В. Боженко |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №2 |
| ПРЕДВАРИТЕЛЬНЫЙ АНАЛИЗ ДАННЫХ |
| по курсу: ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ |
|  |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ гр. № | 4314 |  |  |  | Д.А. Паюнен |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2025

**Цель работы**

Осуществить предварительную обработку данных csv-файла, выявить и устранить проблемы в этих данных.

**Вариант 12**

Задание на лабораторную работу

1. Загрузить датасет с помощью библиотеки pandas.
2. Вывести первые 20 строк с помощью метода head.
3. Выполнить обзор данных - описать столбцы и вашу предметную область.
4. С помощью метода info оценить данные (есть ли пропуски, сколько всего строк, какие типы данных у столбцов).
5. Применить describe. С помощью describe оценить числовые столбцы (если они есть).
6. Вывести на экран названия столбцов с помощью df.columns. Выявить проблемы с названиями, если они есть. При необходимости переименовать столбцы. Если проблемы не обнаружены также дать пояснения.
7. Проверить данные на наличие пропусков и устранить их, если они есть (пропуски необходимо либо удалить, либо заменить каким-то значением).
8. Проверьте данные на наличие явных и неявных дубликатов. Удалите дубликаты, если они есть.
9. Проверьте типы данных, при необходимости измените типы данных, чтобы они соответствовали действительности.
10. Осуществите группировки и создайте сводные таблицы в соответствии с вариантом.
11. Сделайте выводе по работе.

**Ход работы**

**1. Предметная область**

Данный набор данных содержит информацию о фильмах, включая название, год выпуска, рейтинг, количество голосов, продолжительность, основной жанр и страну производства. Анализ этих данных позволяет исследовать тенденции в киноиндустрии, выявлять популярные жанры в разные годы, оценивать успешность фильмов на основе рейтингов и количества голосов, а также сравнивать характеристики фильмов в зависимости от страны производства. Такие данные могут быть полезны для кинокритиков, исследователей кино, стриминговых платформ для формирования рекомендаций и принятия решений о закупке контента, а также для широкой аудитории, интересующейся статистикой фильмов.

**2. Набор данных**

Согласно методическому пособию: movies.csv

**3. Атрибуты**

* title — Название фильма
* release\_year — Год выхода фильма
* score — Рейтинг фильма
* number\_of\_votes — Количество голосов за фильм
* duration — Продолжительность фильма в минутах
* main\_genre — Основной жанр фильма
* main\_production — Основная страна производства фильма

**4. Чтение файла (набора данных)**

Импорт библиотек, чтение файла с помощью pandas (рис. 1).

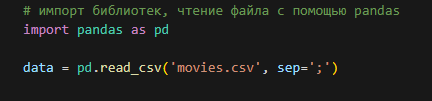


Рисунок 1 – Импорт библиотеки pandas

**5. Обзор данных**

**5.1. Метод head**

Вывод первых 20 строк с помощью метода head (рис. 2).

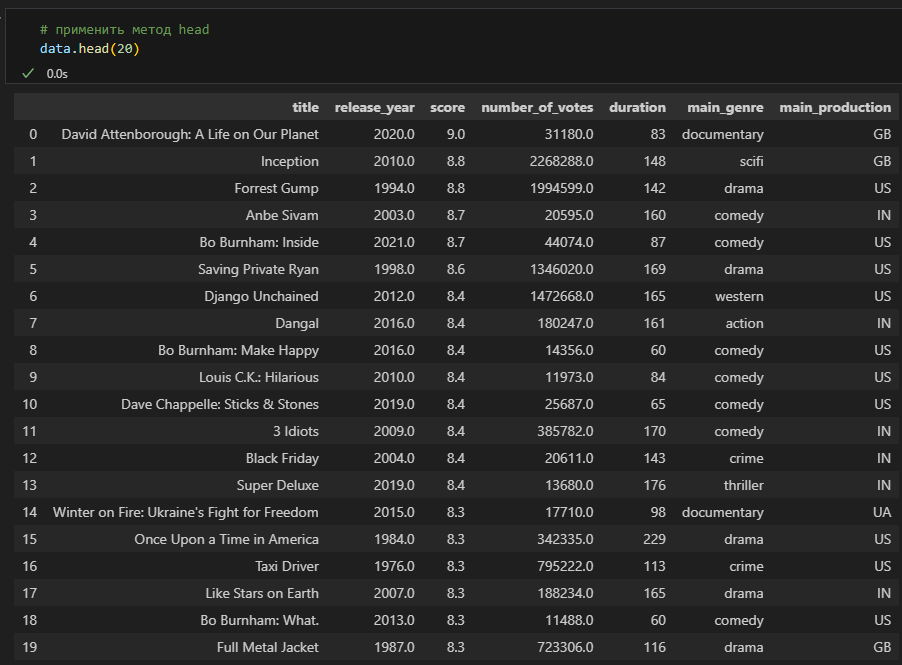


Рисунок 2 – Метод head

**5.2. Метод info**

Оценка данных с помощью метода info (рис. 3).

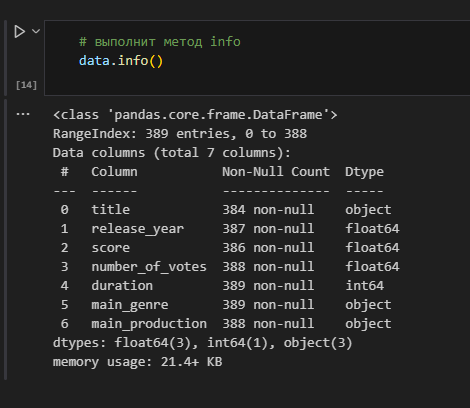


Рисунок 3 – метод info

**5.3. Метод describe**

Оценка данных с помощью метода describe (рис. 4).

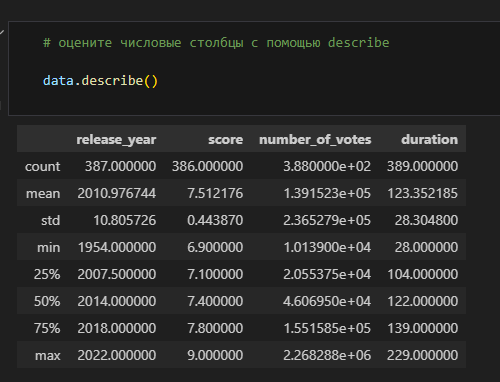


Рисунок 4 – метод describe

Данный набор данных содержит информацию о 389 фильмах. Типы данных:

* Числовые: release\_year, score, number\_of\_votes, duration.
* Категориальные/Строковые: title, main\_genre, main\_production.

Пропуски обнаружены в столбцах title, release\_year, score, number\_of\_votes и main\_production. Количество пропусков незначительно по сравнению с общим количеством записей (389).

**5.4. Оценка названий столбцов**

Были выведены на экран названия столбцов. Серьезных проблем в названиях столбцов нет.(рис. 5).

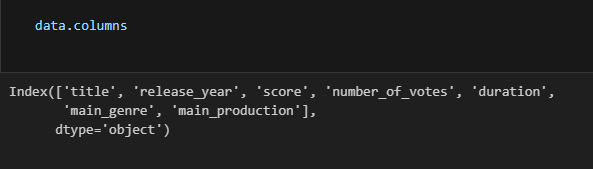


Рисунок 5 – Оценка названий столбцов

**6. Проверка пропусков**

Были удалены все записи с пропущенными значениями, тк в основном они были категориальные, а также их было не так много, соответственно на анализ они не повлиют.

Также можно было заполнить пропуски медианным значением или модой у категориальных признаков.

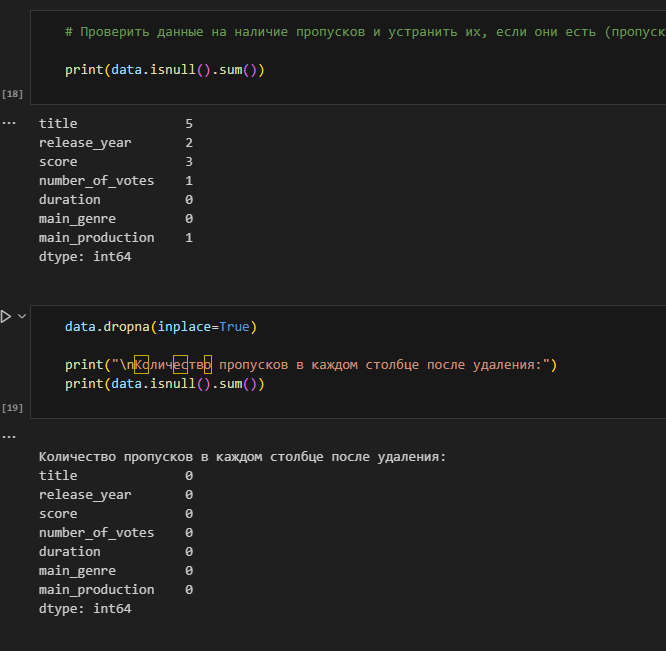


Рисунок 6 – Проверка пропусков

**7. Проверка дубликатов**

**7.1*.* Проверка явных дубликатов**

Команда print(df.duplicated().sum()) выводит на экран количество строк-дубликатов.(рисунок 7). Было обнаружено 2 дубликата и дальше они были удалены.

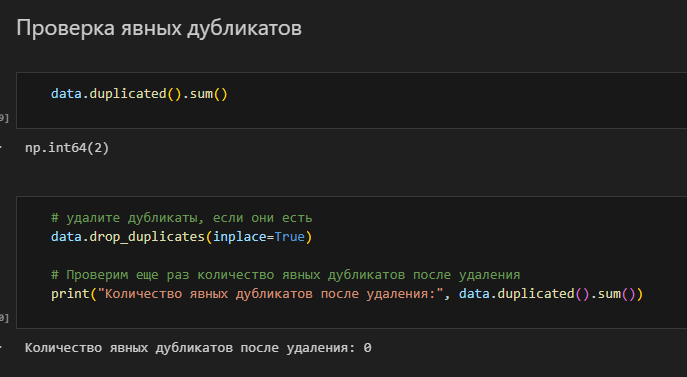


Рисунок 7 – Проверка явных дубликатов

**7.2. Проверка неявных дубликатов**

Была сделана проверка явных дубликатов (рис. 8).

В данных были обнаружены неявные дубликаты:

* COMEDY – дублируется comedy;
* drama – дублируется dramaa.

Эти проблемы нужно исправить на этапе предобработки, чтобы анализ был корректным.

Все неявные дубликаты устранены. Данные приведены к корректному виду, теперь с ними можно уверенно работать дальше: строить группировки, сводные таблицы и анализировать.

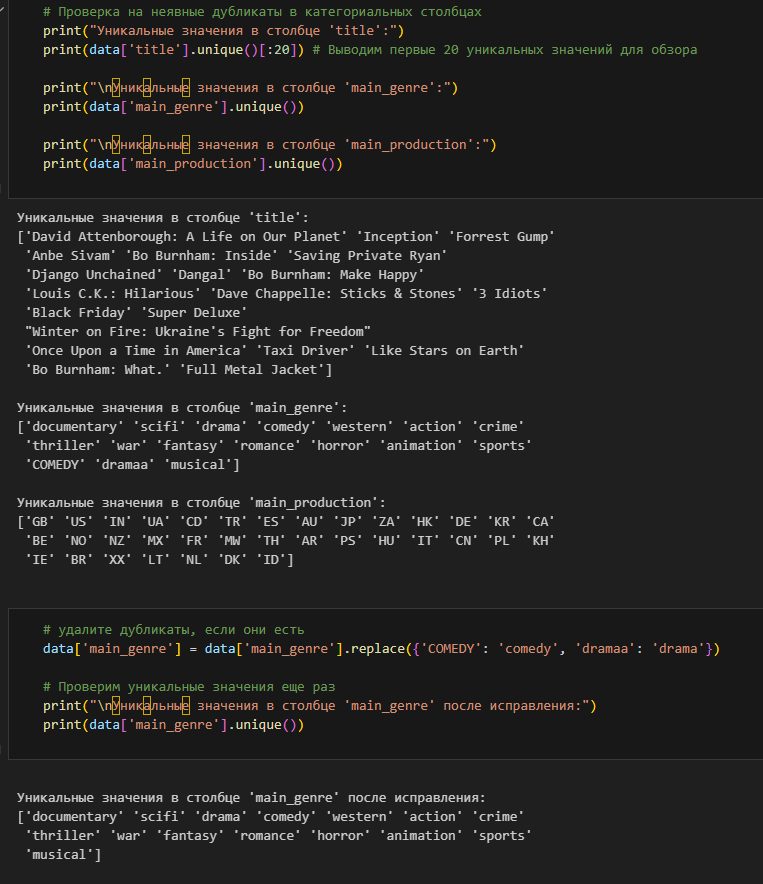


Рисунок 8 – Проверка неявных дубликатов

**8. Проверка типов данных**

Были проверены типы данных(рис. 9):

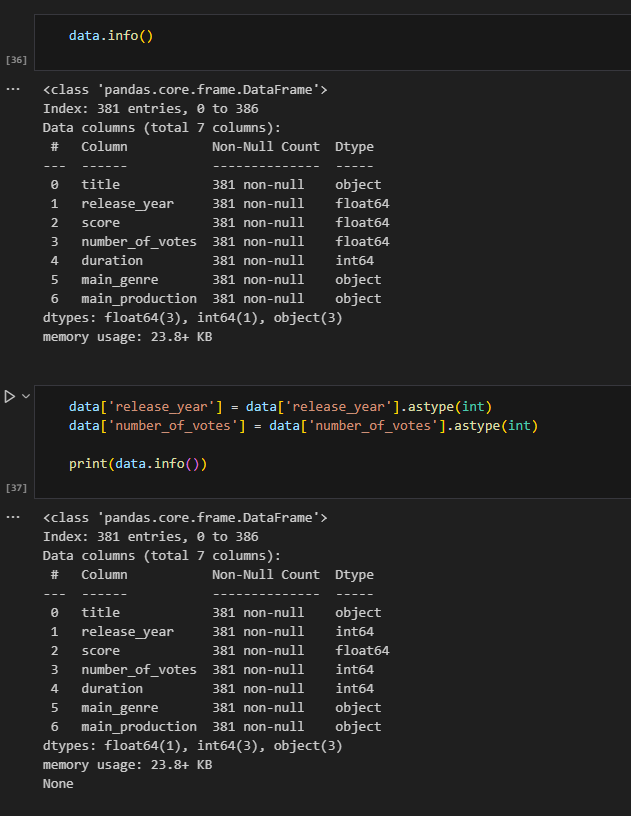


Рисунок 9 – Проверка типов данных

Были изменены типы данных для параметров - release\_year, number\_of\_votes, тк они представляют целочисленные значения.

**9 Группировка данных**

**9.1. Задание 1**

Группировка - год выхода и количество фильмов каждого жанра

(main\_genre). (рис. 10).

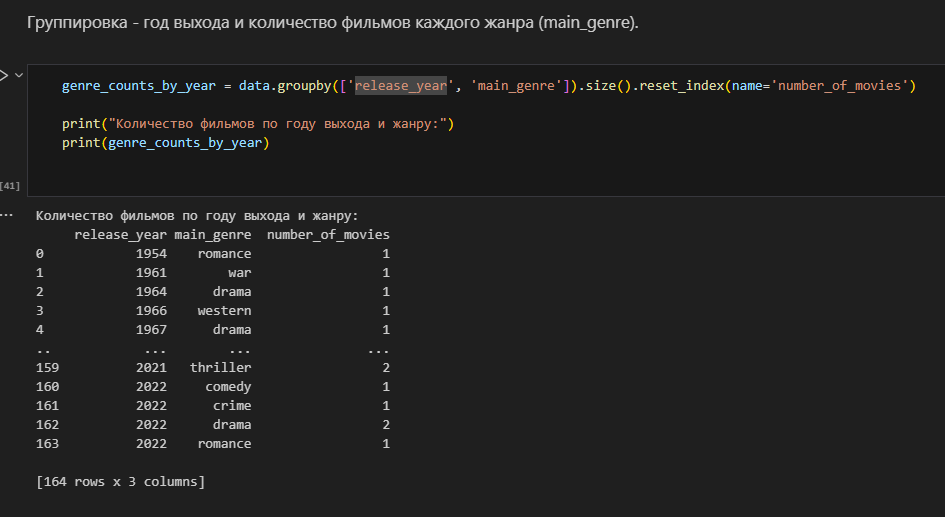


Рисунок 10 – Задание 1

На основе результатов группировки по году выхода и жанру (genre\_counts\_by\_year) можно увидеть распределение фильмов по жанрам в разные годы. Например, можно определить:

* Какие жанры были популярны в определенные годы.
* Как менялось количество фильмов определенного жанра с течением времени.
* В какие годы выходило больше всего фильмов в целом.

Этот результат помогает понять тенденции в кинематографе с точки зрения жанров и их динамики по годам.

**9.2*.* Задание 2**

Группировка - год выхода и количество фильмов каждого жанра

(main\_genre). Создать датафрейм. Переименовать столбец с количеством в “сount”. Отсортировать по убыванию столбца “count”. (рис. 12).

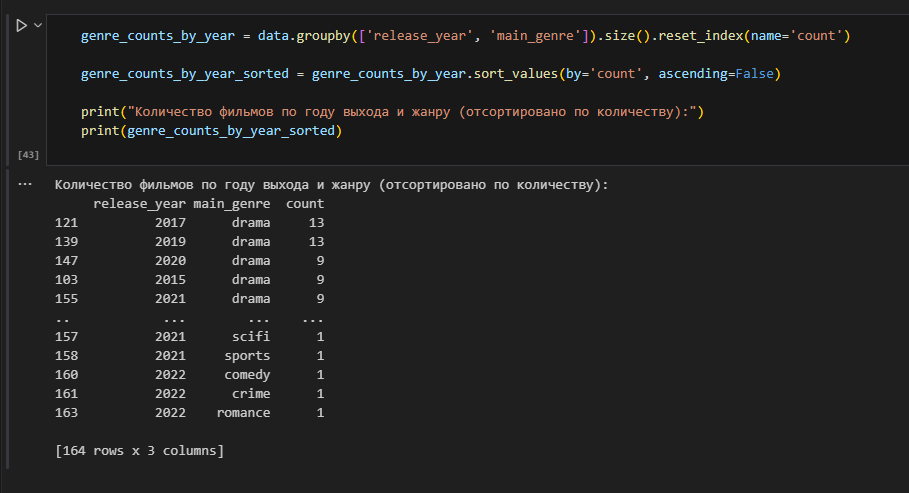


Рисунок 12 – Задание 2

Результат этой группировки (genre\_counts\_by\_year\_sorted), отсортированный по убыванию количества фильмов, позволяет быстро увидеть наиболее "продуктивные" сочетания года выпуска и жанра. Можно определить:

* В какие годы и в каких жанрах было выпущено наибольшее количество фильмов.
* Наиболее популярные жанры в определенные периоды времени с точки зрения объема производства.

Например, верхние строки отсортированной таблицы показывают, что в 2017 и 2019 годах жанр "drama" был наиболее представлен с 13 фильмами в каждом году.

**9.3. Задание 3**

Сводная таблица (pivot\_table) - средний рейтинг (score) фильмов по

жанрам. Отсортировать по убыванию рейтинга. Округлить до двух знаков. (рис. 13).

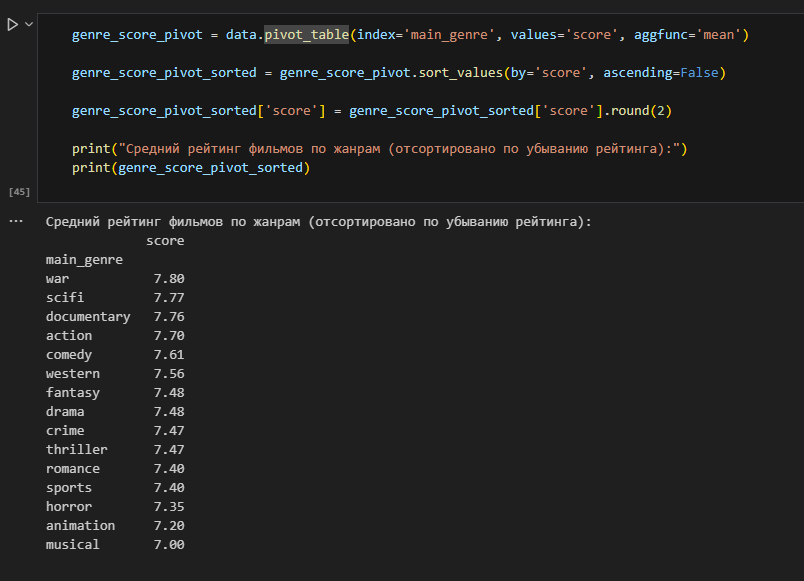


Рисунок 13 – Задание 3

На основе результатов сводной таблицы (genre\_score\_pivot\_sorted) можно увидеть средний рейтинг фильмов для каждого жанра, отсортированный от самого высокого к самому низкому. Это позволяет определить:

* Какие жанры имеют наивысший средний рейтинг среди зрителей (например, "war" и "scifi" в данном случае).
* Какие жанры имеют более низкий средний рейтинг (например, "animation" и "musical").
* Сравнить средние рейтинги разных жанров между собой.

**9.4. Задание 4**

Сводная таблица (pivot\_table) - средний рейтинг (score) фильмов по

жанрам - строки и годам - столбцы. Отсортировать по возрастанию main\_genre. (рис. 14).

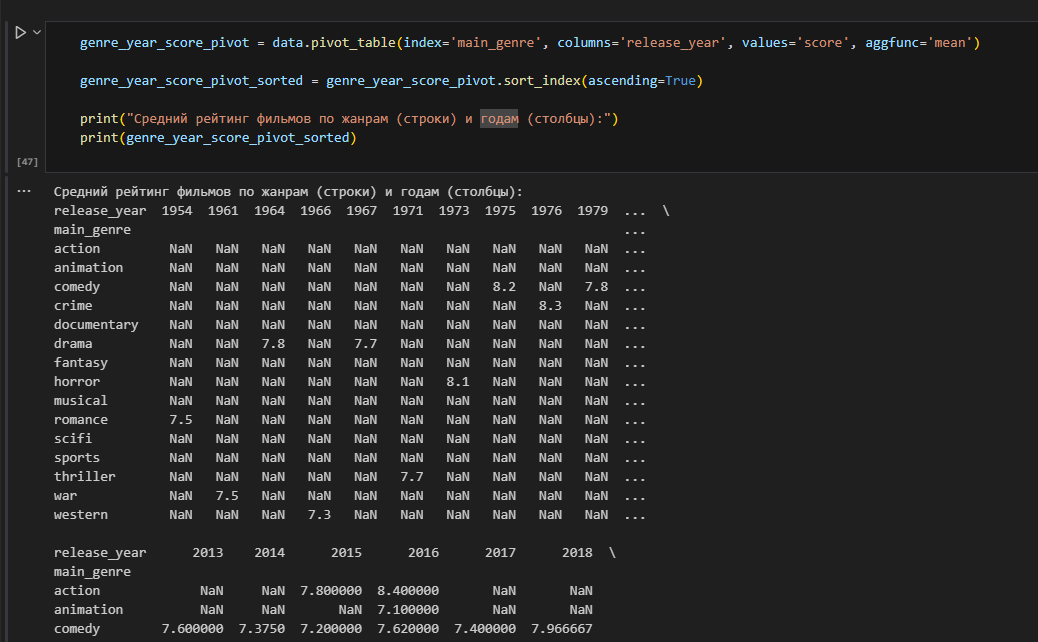


Рисунок 14 – Задание 4

На основе результатов выполнения кода, который создает сводную таблицу средних рейтингов фильмов по жанрам (строки) и годам выпуска (столбцы), можно увидеть:

* Динамику среднего рейтинга по жанрам с течением времени: можно проследить, как менялся средний рейтинг определенного жанра из года в год.
* Сравнение средних рейтингов разных жанров в конкретные годы: Можно сравнить, какие жанры имели более высокие или низкие средние рейтинги в определенный год.
* Годы с наибольшей активностью по жанрам: Ячейки с числовыми значениями указывают на наличие фильмов данного жанра в соответствующем году. Пустые ячейки (NaN) означают отсутствие данных (фильмов) для данного сочетания жанра и года в вашем наборе данных.

**Вывод**

В рамках лабораторной работы был проанализирован набор данных movies.csv, содержащий сведения о 389 фильмах: название, год выпуска, рейтинг, количество голосов, продолжительность, жанр и страну производства.

Предобработка данных включала загрузку набора в DataFrame, анализ структуры и удаление строк с пропусками и дубликатами. Были устранены неявные дубликаты в категориальных признаках (например, некорректные значения жанров), приведены типы данных к корректным (например, release\_year и number\_of\_votes — к целочисленным). После очистки в датасете осталось 381 запись без пропусков и дубликатов.

Группировки и сводные таблицы позволили выявить ключевые закономерности:

* Количество фильмов по годам и жанрам показало периоды наибольшей активности кинопроизводства и доминирующие жанры в разные годы.
* Отсортированная группировка подтвердила, что в 2017–2019 годах жанр drama был наиболее представлен.
* Сводная таблица среднего рейтинга по жанрам показала, что наивысшие оценки получили жанры war, scifi и documentary, тогда как animation и musical имели более низкие значения.
* Анализ динамики рейтингов по жанрам и годам позволил проследить изменения зрительских предпочтений.

В результате была проведена полная предобработка и базовый исследовательский анализ данных о фильмах. Полученные сводки и группировки дают представление о распределении жанров, динамике их популярности и различиях в оценках, что создаёт основу для дальнейшего анализа тенденций в киноиндустрии.

Ссылка на репозиторий в GitHub с файлами лабораторной работы: <https://github.com/Yeetmq/suai_da>