Введение

Файл «movies.csv» – это база данных, содержащая информацию о киноиндустрии за период 1980-2020 годы.

Цель данного проекта – произвести очистку данных (Data Cleaning) и анализ данных (Data Analysis) с визуализацией (Data Visualization).

Метаданные

Набор данных представляет собой файл в формате CSV и представлен следующими столбцами:

- name (object) – название фильма
- rating (object) – рейтинг фильма (R, PG и т.д.)
- genre (object) - жанр
- year (int64) - год
- released (object) – дата релиза
- score (float64) – рейтинг пользователей IMDb
- votes (float64) – количество проголосовавших
- director (object) - режиссер
- writer (object) - сценарист
- star (object) – основной актер/актриса
- country (object) - страна
- budget (float64) - бюджет
- gross (float64) - доход
- company (object) - компания
- runtime (float64) – продолжительность фильма

Файл содержит 7668 строк и 15 столбцов. Все операции были выполнены в Jupyter Notebook. Все файлы доступны по ссылке: enupilov/MovieIndustry (github.com).

Очистка данных (Data Cleaning)

При очистке данных были выполнены следующие шаги:

1. Удалены пустые строки;
2. Создан столбец с годом, откорректированным по дате релиза;
3. Удалены дубликаты;
4. Проверка аномальных значений в столбце budget и gross.

Анализ данных (Data Analysis)

При анализе данных были визуализированы следующие данные:

1. Создана точечная диаграмма для визуализации зависимости столбцов budget и gross;
2. Создана точечная диаграмма при помощи Seaborn;
3. Создана корреляция числовых столбцов;
4. Создана корреляция всех столбцов путем перевода значений в числа.