|  |  |
| --- | --- |
|  | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ \_\_\_\_\_\_ИНФОРМАТИКА И СИСТЕМЫ УПРАВЛЕНИЯ\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА \_\_\_\_СИСТЕМЫ ОБРАБОТКИ ИНФОРМАЦИИ И УПРАВЛЕНИЯ (ИУ5)\_\_\_\_

**РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА**

***К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ***

по дисциплине \_\_\_\_\_\_\_\_\_\_\_\_\_ Оперативный анализ данных \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

по теме\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ «Movies» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент ИУ5-52Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Е. В. Табахов

(Группа) (Подпись, дата) (И.О.Фамилия)

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**  К. Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

Консультант **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

(Подпись, дата) (И.О.Фамилия)

*2024 г.*

**Министерство науки и высшего образования Российской Федерации**

**Федеральное государственное бюджетное образовательное учреждение**

**высшего образования**

**«Московский государственный технический университет имени Н.Э. Баумана**

**(национальный исследовательский университет)»**

**(МГТУ им. Н.Э. Баумана)**

УТВЕРЖДАЮ

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_

(Индекс)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(И.О.Фамилия)

« \_\_\_\_\_ » \_\_\_\_\_\_\_\_\_\_\_\_ 20 \_\_\_\_ г.

**ЗАДАНИЕ**

**на выполнение научно-исследовательской работы**

по теме \_\_\_\_\_\_Анализ данных по базе данных «Movies»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Студент группы \_\_\_\_\_\_\_ИУ5-52Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Табахов Евгений Вячеславович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(Фамилия, имя, отчество)

Направленность НИР (учебная, исследовательская, практическая, производственная, др.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Исследовательская\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Источник тематики(кафедра,предприятие,НИР) \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_НИР\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

График выполнения НИР: 25% к 3 нед., 50% к 9 нед., 75% к 12 нед., 100% к 15 нед.

***Техническое задание***  *Спроектировать систему анализа и визуализации по теме «*Movies*», провести визуализацию полученных данных посредством языка программирования Python*

***Оформление научно-исследовательской работы:***

Расчетно-пояснительная записка на 26 листах формата А4.

Перечень графического (иллюстративного) материала (чертежи, плакаты, слайды и т.п.)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Дата выдачи задания «10» сентября 2024 г.

**Руководитель НИР**  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ К. Ю. Маслеников

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** Е. В. Табахов\_\_\_\_\_

(Подпись, дата) (И.О.Фамилия)

Примечание: Задание оформляется в двух экземплярах: один выдается студенту, второй хранится на кафедре.

**Аннотация**

В этой работе требуется проанализировать базу данных "Movies", очистить её от ненужных данных, выдвинуть гипотезы, проанализировать их и составить графики по полученным данным. Для выполнения задания была выбрана база данных "Movies" с информацией о 3975 фильмах. В ходе работы будут использоваться такие библиотеки, как «Pandas» и «Seaborn». Анализ будет проводиться через программу «Google Colab» на языке программирования Python.

Также будет производиться корреляционный анализ, агрегирование данных, оптимизация памяти, удаление дубликатов, очистка данных. Будет осуществлён анализ гипотез по известным данным и построены соответствующие графики и схемы.

В работе будут рассмотрены данные о фильмах в различных разрезах и сравнениях. В конце будет сделан вывод о проделанной работе с данным датасетом.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_Toc181202094)

[Определение данных для анализа](#_Toc181202095) 5

[1. Описание данных 5](#_Toc181202096)

[1. Формулирование гипотез 6](#_Toc181202097)

[2. Подготовка данных для работы](#_Toc181202098) 7

[3. Изучение общей информации 8](#_Toc181202099)

[4. Исследовательский анализ данных 15](#_Toc181202100)

[Гипотеза 1 - Фильмы с большим бюджетом чаще получают больше наград 15](#_Toc181202101)

[Гипотеза 2 – Режиссерский опыт увеличивает прибыльность фильма 17](#_Toc181202102)

[Гипотеза 3 – Влияет ли жанр на среднюю прибыль 19](#_Toc181202103)

[Гипотеза 4 – Длина фильма влияет на кассовые сборы 20](#_Toc181202104)

[Гипотеза 5 – Более новые фильмы имеют лучшую доходность 22](#_Toc181202105)

[Анализ корреляции 24](#_Toc181202106)

[ЗАКЛЮЧЕНИЕ 25](#_Toc181202107)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 26](#_Toc181202108)

# **ВВЕДЕНИЕ**

Необходимо исследовать базу данных и выявить закономерности, связанные с фильмами.

Цели:

1. определение данных;
2. формулирование гипотез;
3. загрузка данных в Python;
4. проверка данных;
5. очистка данных;
6. преобразование данных;
7. выбор данных для анализа;
8. агрегирование данных;
9. визуализация данных;
10. подтверждение или опровержение поставленных гипотез;
11. формулирование ограничений и выводов.

# **Определение данных для анализа**

В качестве данных для анализа был выбран Dataset "Movies".

В датасете представлены данные о 3975 фильмах, включающие информацию о каждом произведении и описанием их характеристик.

# **Описание данных**

Для анализа были собраны данные о здоровье сна и образе жизни людей, включенных в датасет. В наборе данных содержатся:

* + - * **Movie –** Название фильма.
      * **Director –** Режиссёр фильма, который контролирует процесс создания фильма и руководит актёрами и технической командой.
      * **Running time –** Длительность фильма в минутах.
      * **Actor 1, Actor 2 и Actor 3 –** Три разных актёра, участвующих в фильме.
      * **Genre –** Жанр фильма.
      * **Budget –** Сумма, потраченная на создание фильма и его продвижение.
      * **Box Office –** Доходы от продаж билетов.
      * **Actors Box Office % –** Процент, отражающий, сколько раз актёры смогли как минимум удвоить бюджет в своих других фильмах.
      * **Director Box Office % –** Процент, отражающий, сколько раз режиссёр смог как минимум удвоить бюджет в своих других фильмах.
      * **Earnings –** Разница между доходами от проката и бюджетом.
      * **Oscars and Golden Globes nominations –** Количество номинаций, которые фильм получил на "Оскар" и "Золотой глобус".
      * **Oscars and Golden Globes awards –** Количество наград, которые фильм получил на "Оскар" и "Золотой глобус".
      * **Release year –** Год, когда фильм был впервые выпущен.
      * **IMDb score –** Оценка фильма на IMDb по шкале от 1 до 10, рассчитываемая на основе голосов зарегистрированных пользователей IMDb.

# **Формулирование гипотез**

Гипотеза 1:Фильмы с большим бюджетом чаще получают больше наград.

Гипотеза 2: Режиссерский опыт увеличивает прибыльность фильма.

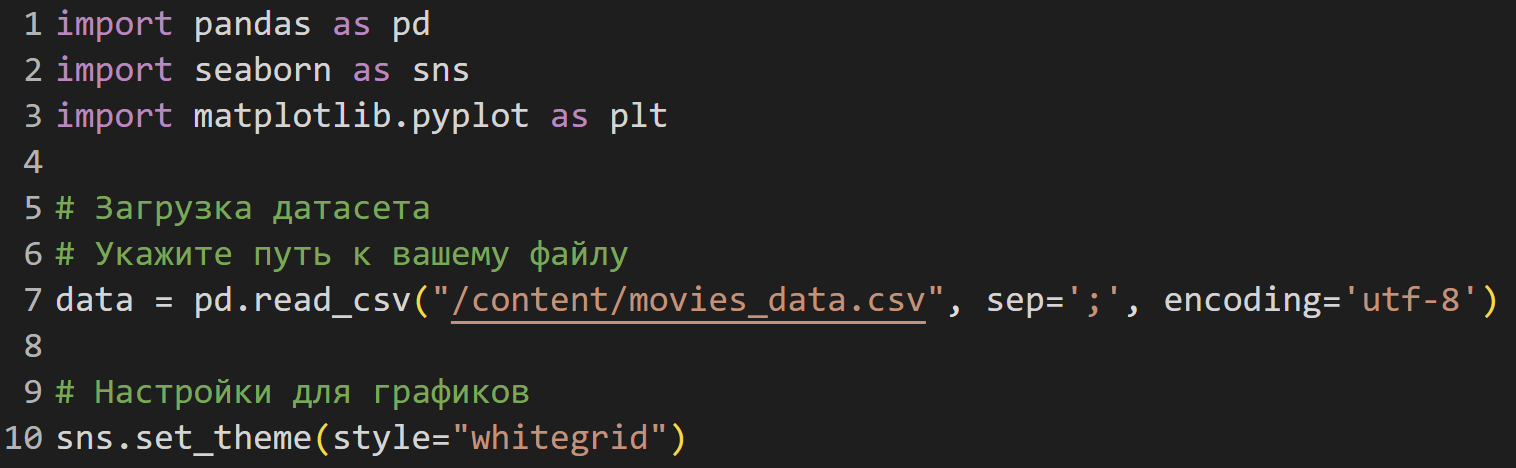
Гипотеза 3: Влияет ли жанр на среднюю прибыль.

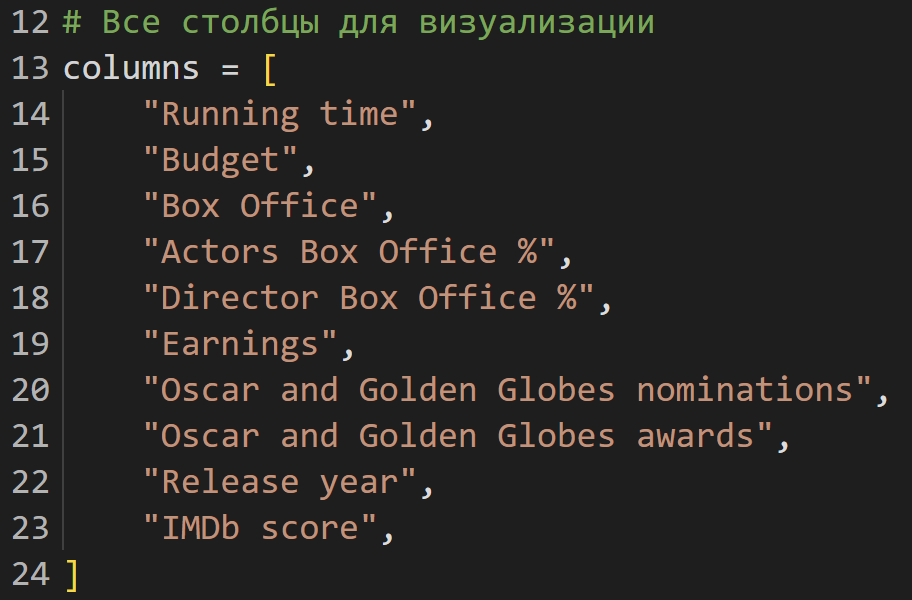
Гипотеза 4: Длина фильма влияет на кассовые сборы.

Гипотеза 5: Современные фильмы имеют лучшую доходность.

# **Подготовка данных для работы**

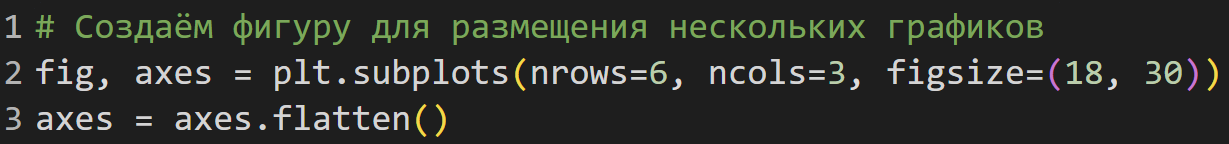
Загружаем датасет и подключаем необходимые библиотеки, выбираем из датасета поля, необходимые для анализа, и проводим очистку данных.

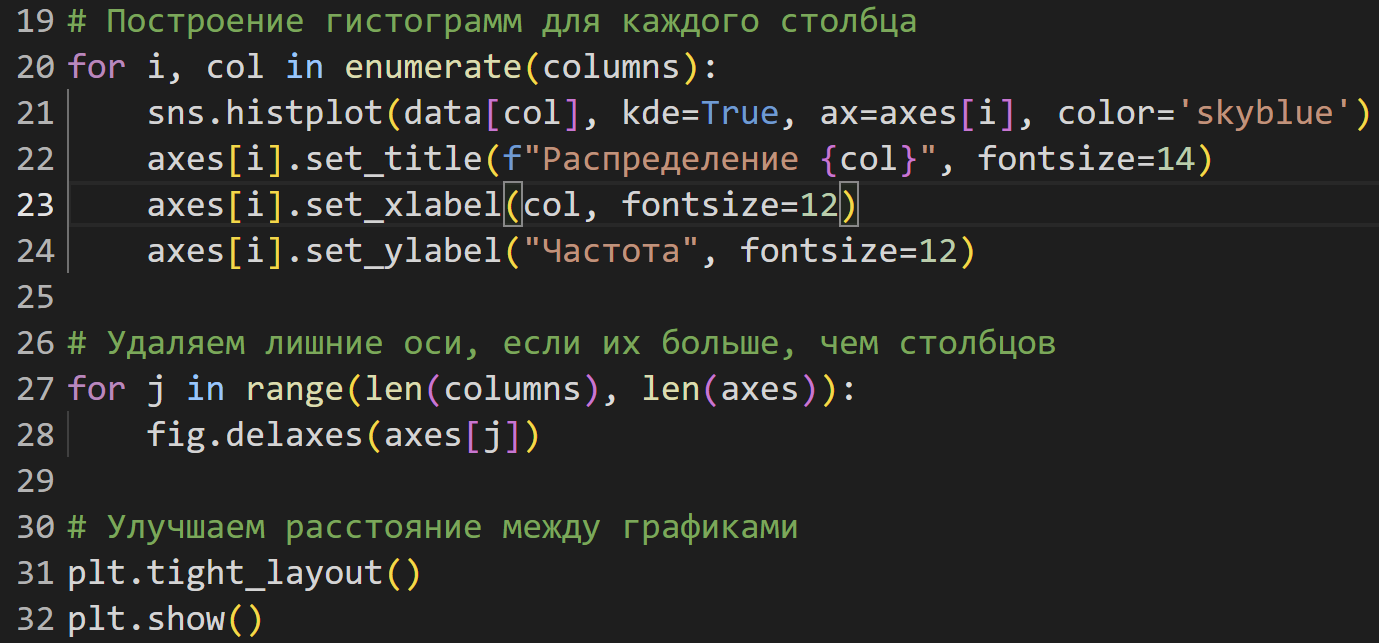


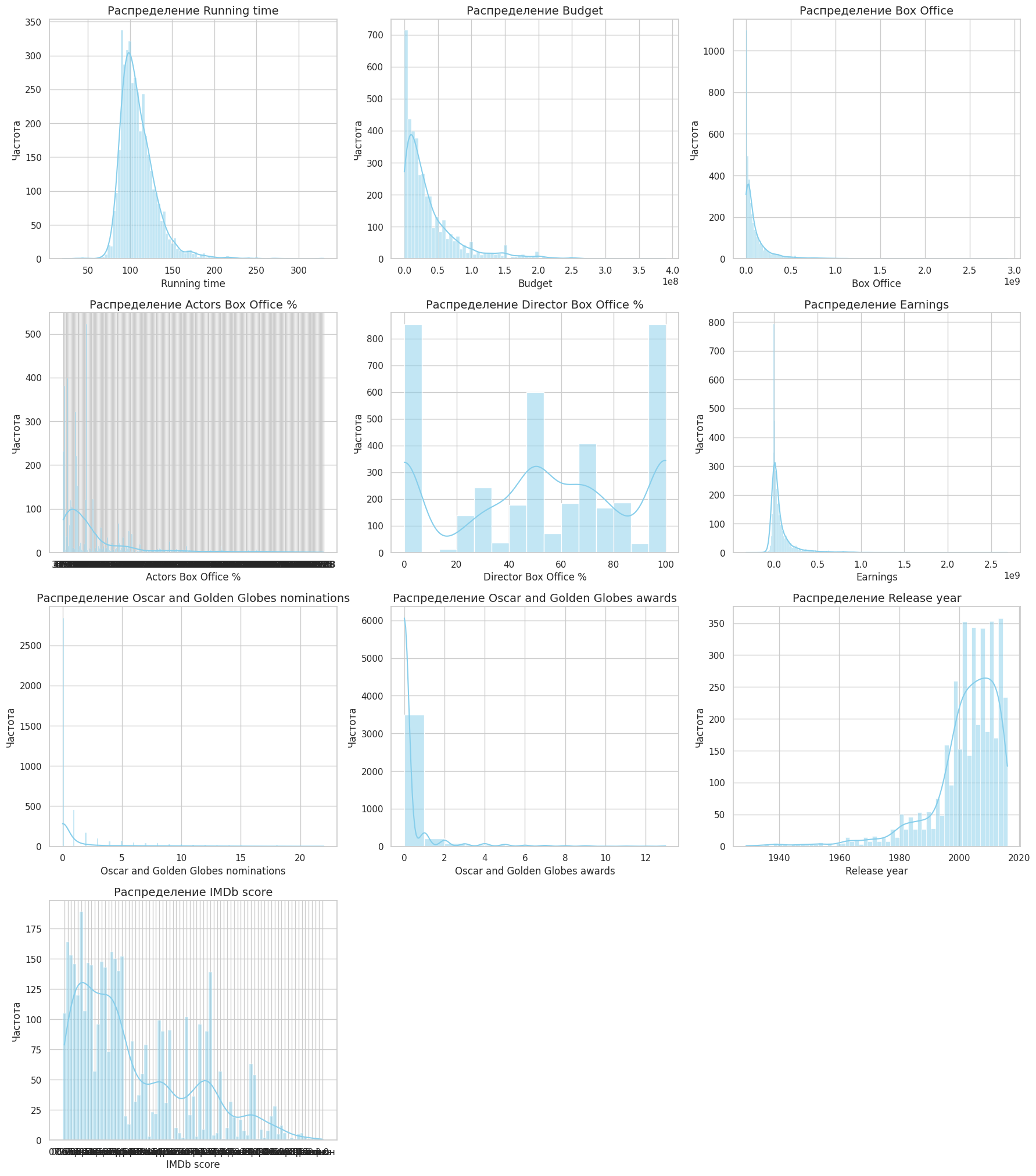


# **Изучение общей информации**

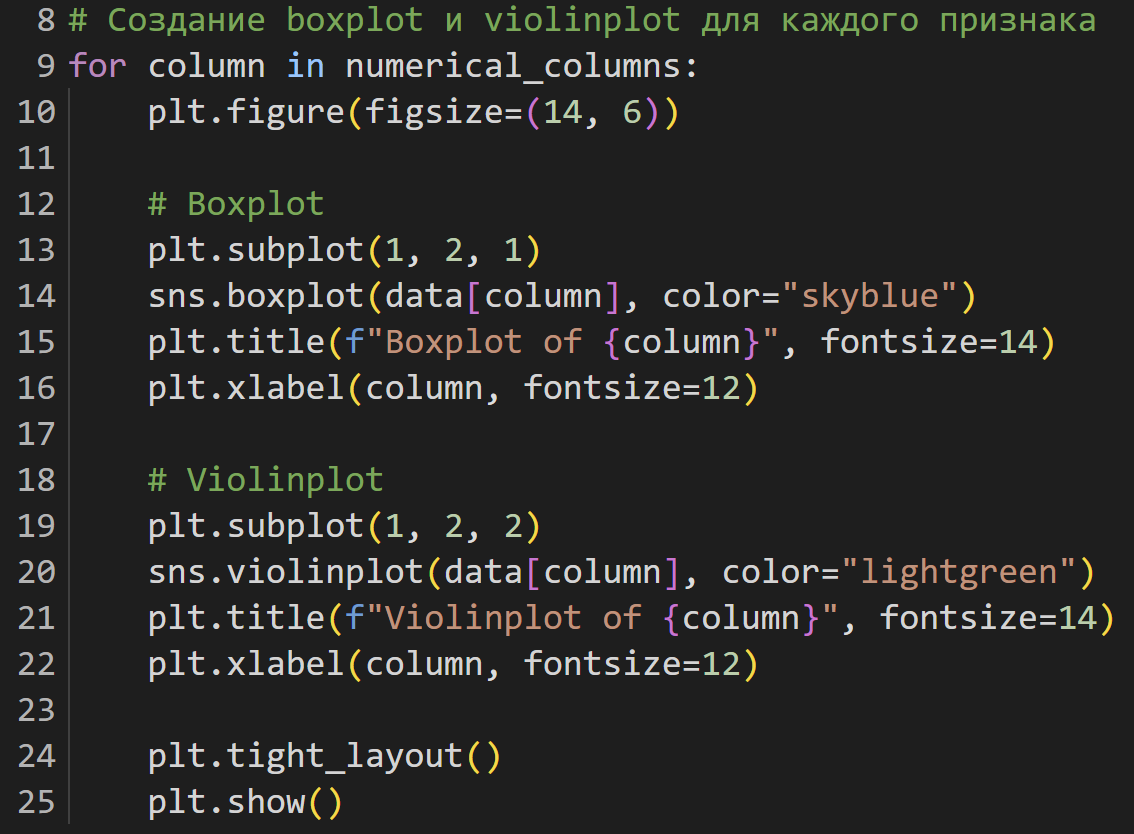
Посмотрим на распределения признаков:

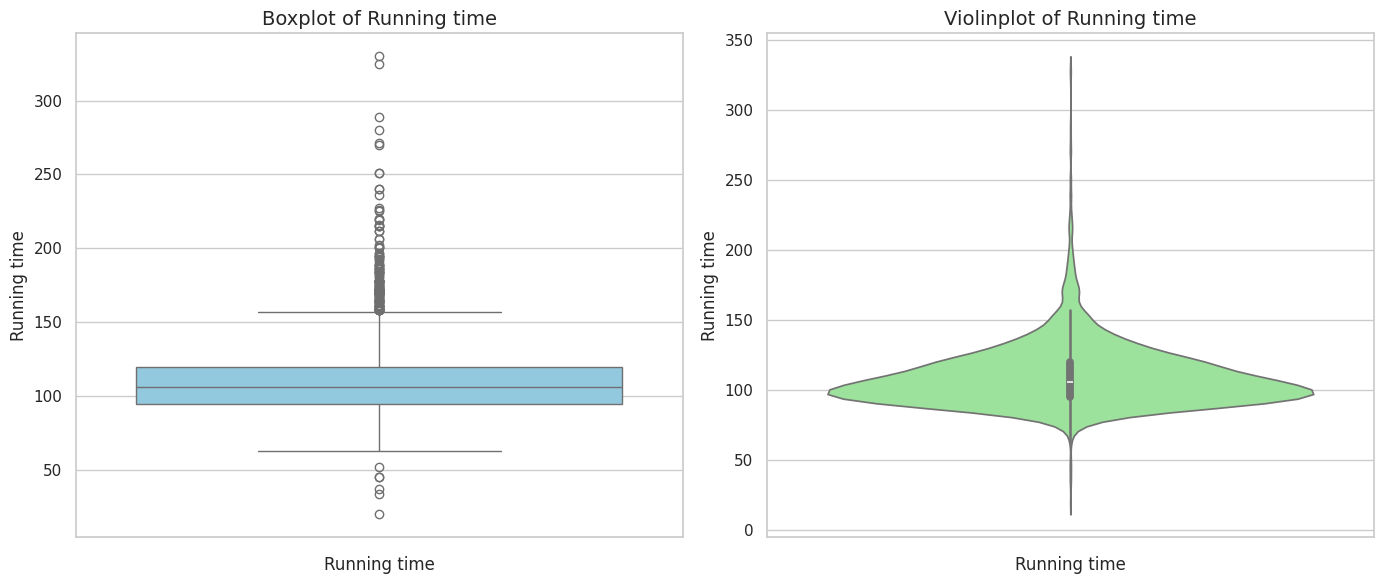


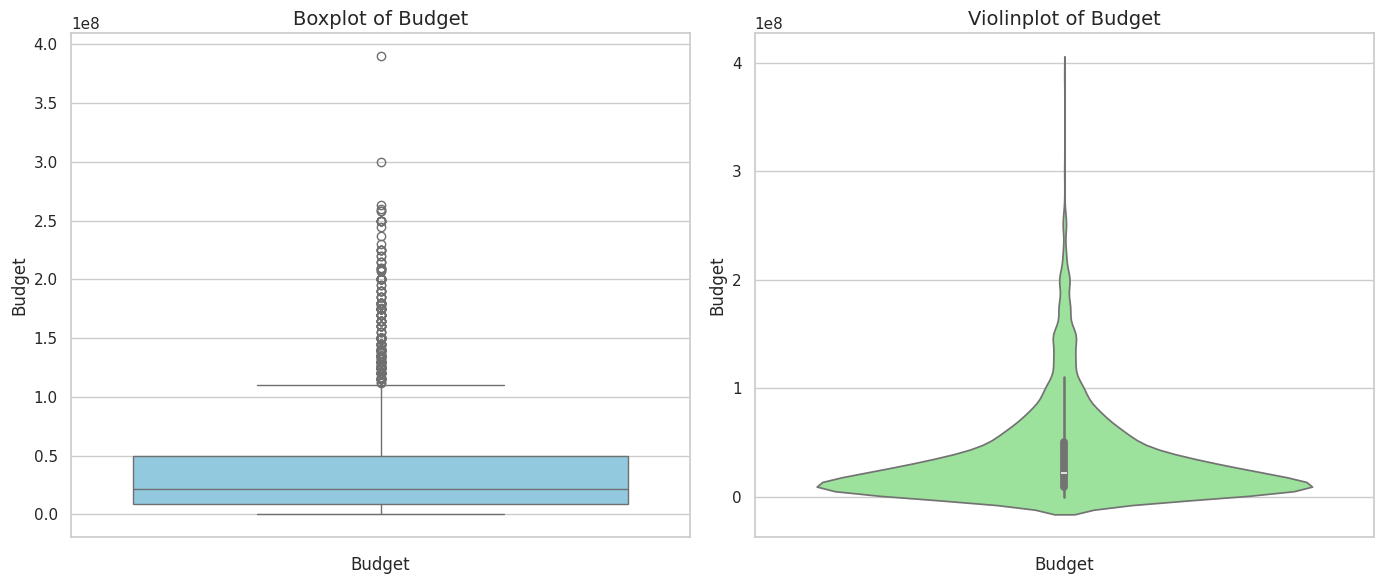


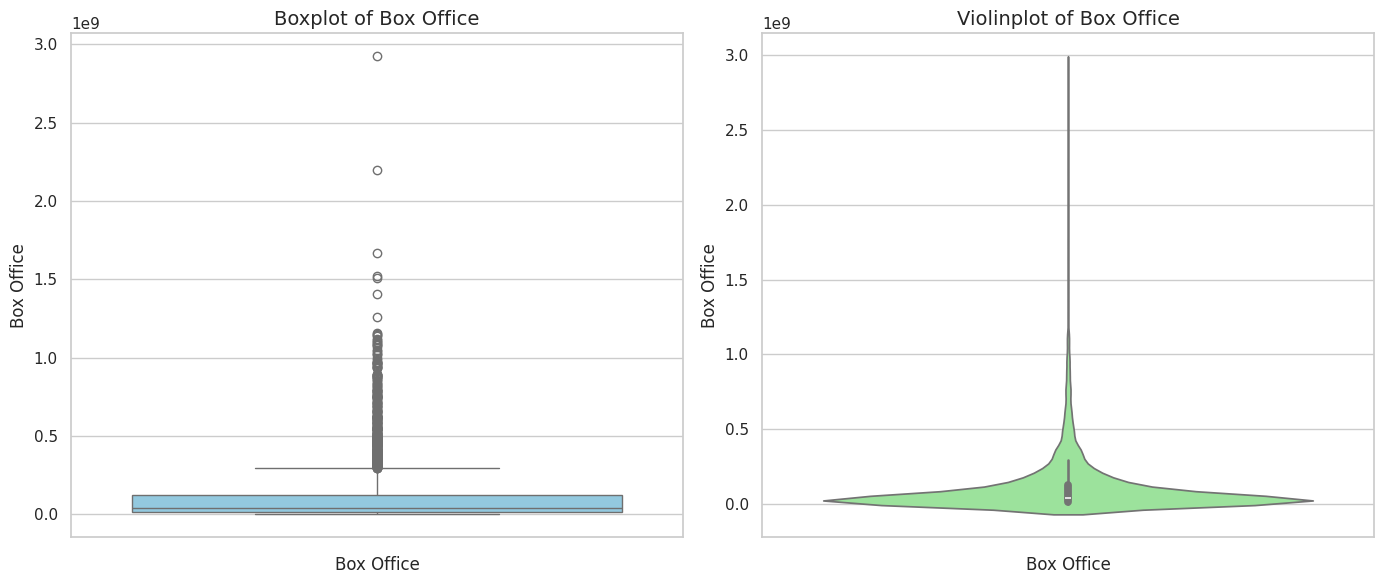


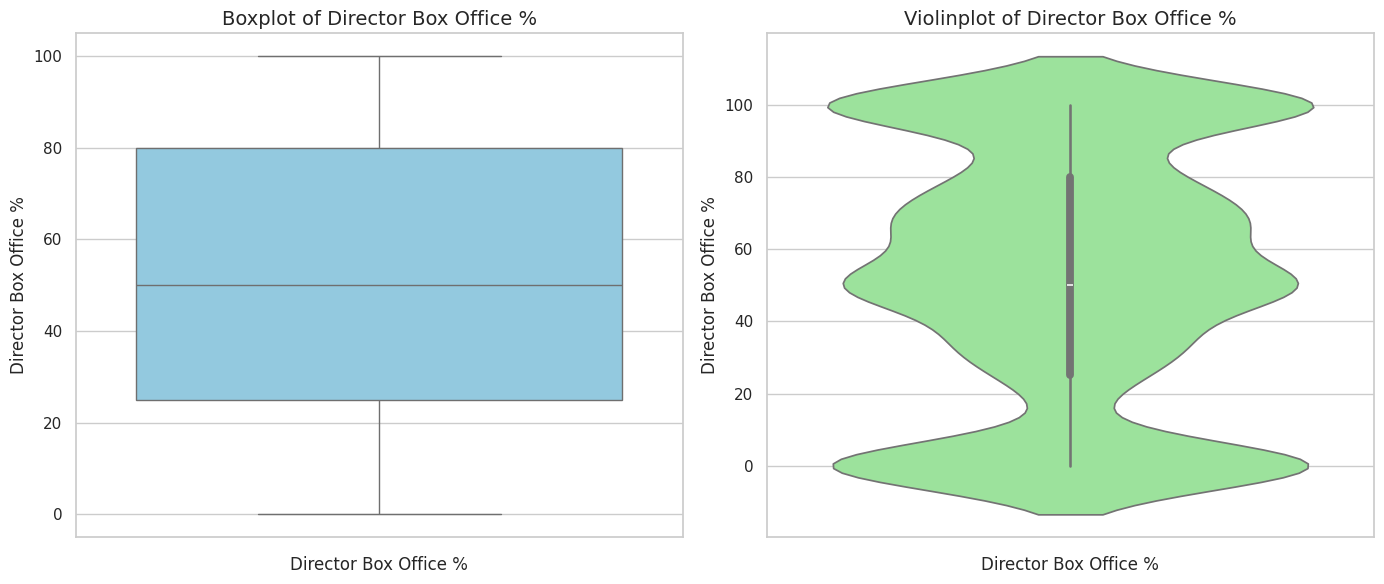
Использование Boxplot и Violinplot для анализа распределений:



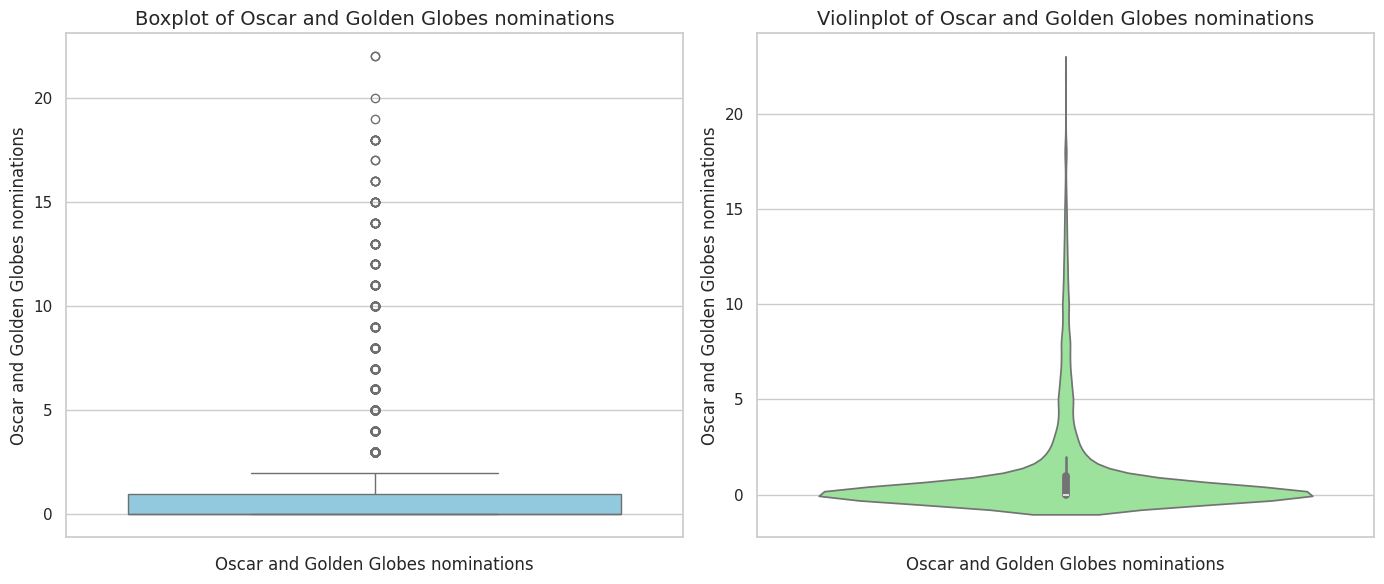


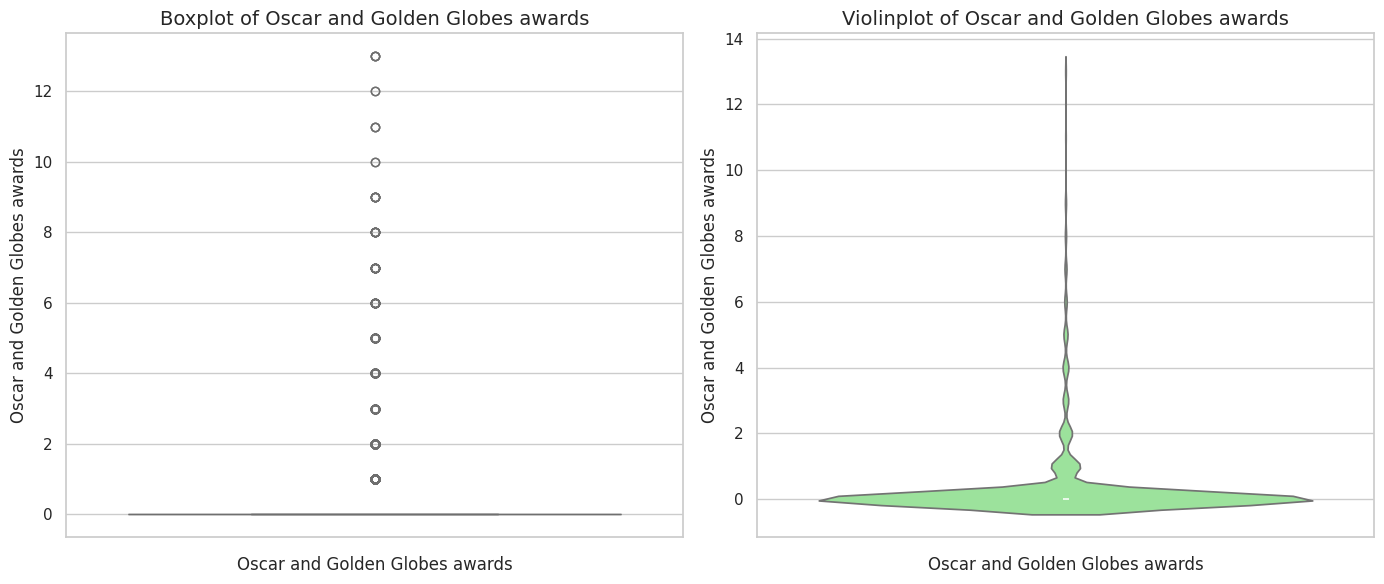




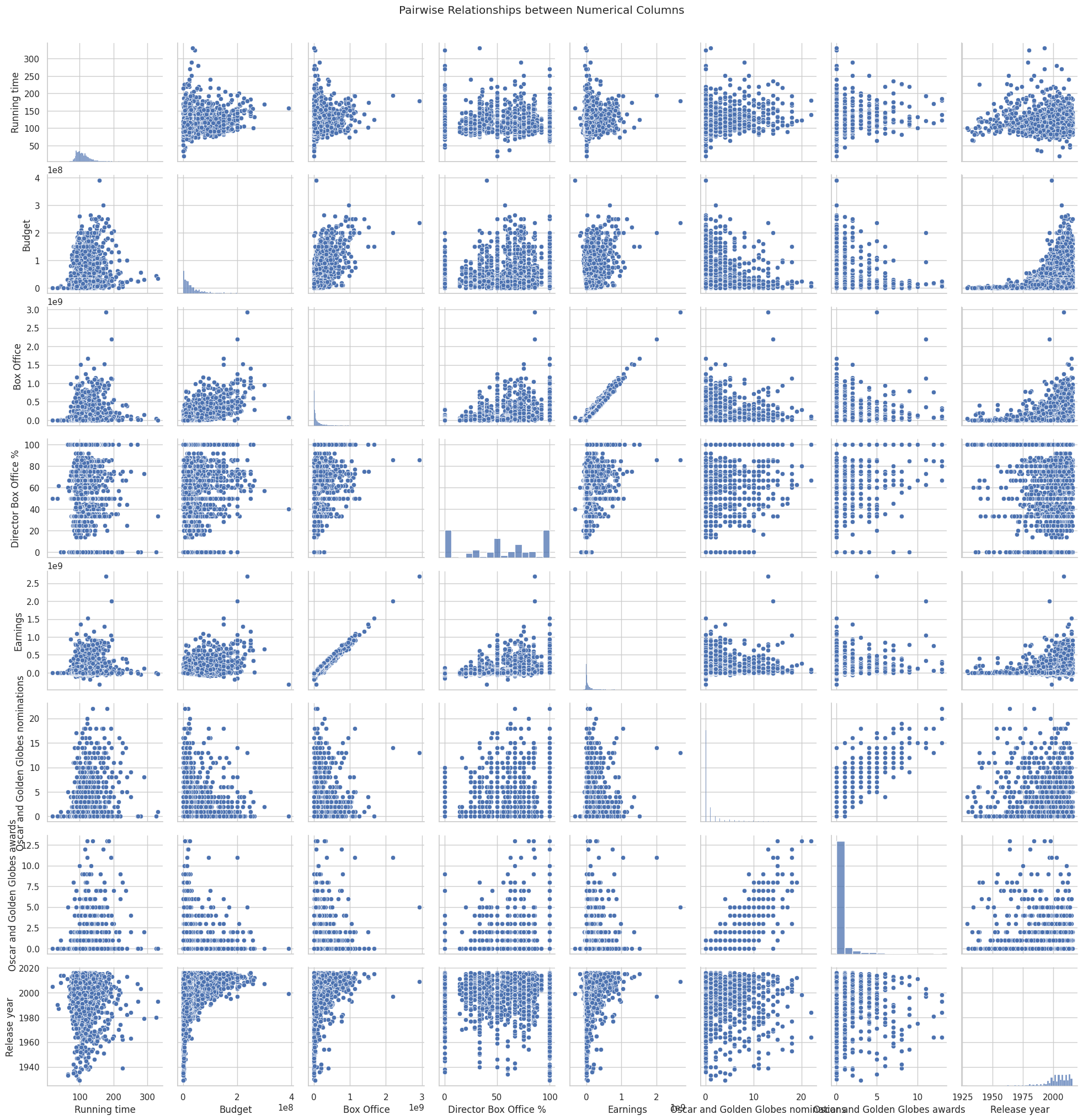






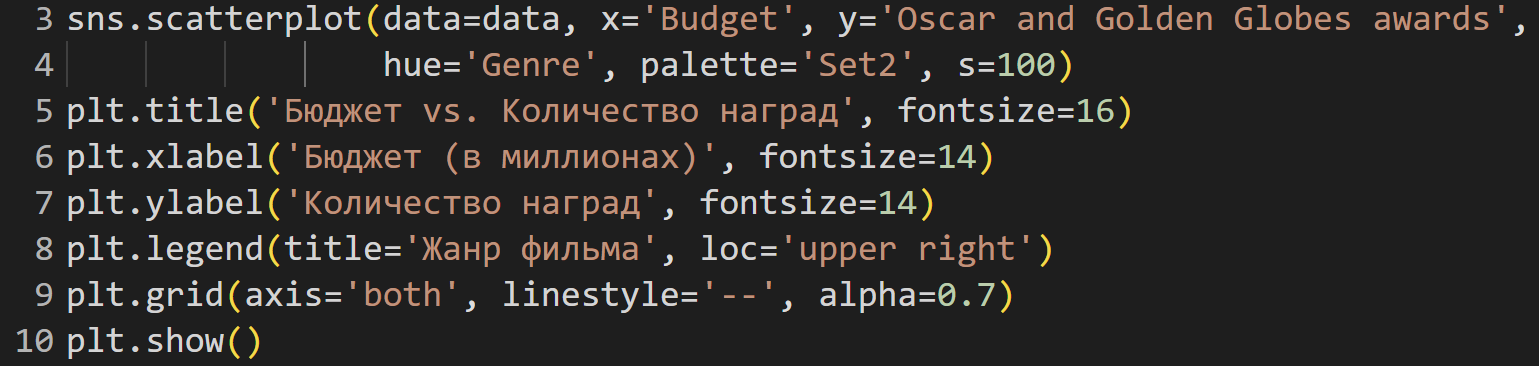


Парные взаимосвязи между признаками:

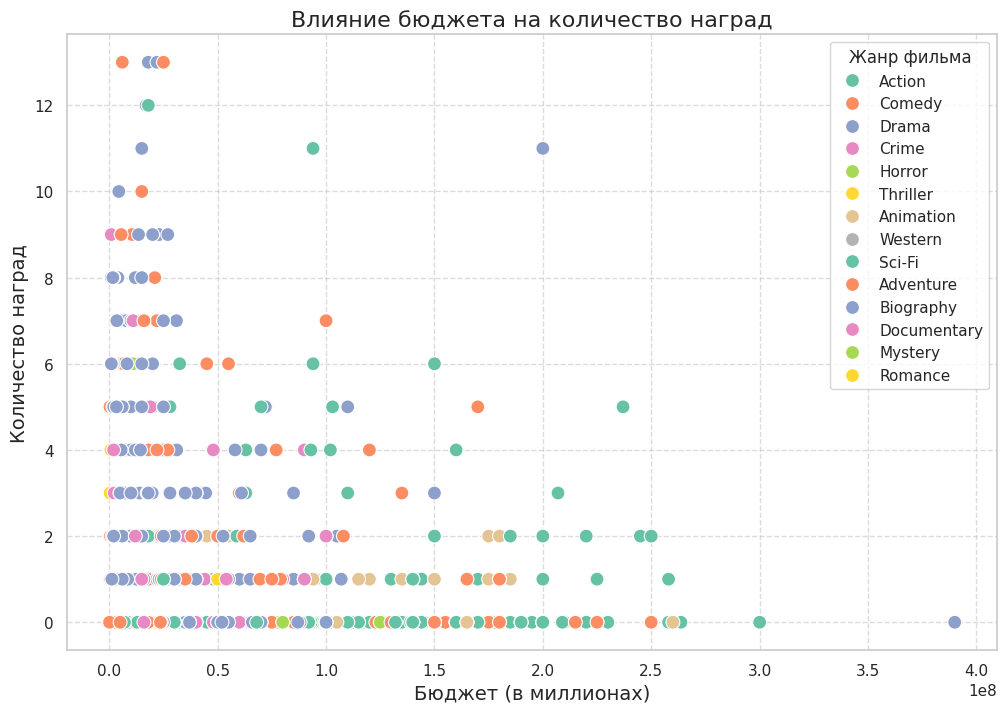


# **Исследовательский анализ данных**

## **Гипотеза 1 - Фильмы с большим бюджетом чаще получают больше наград**

****

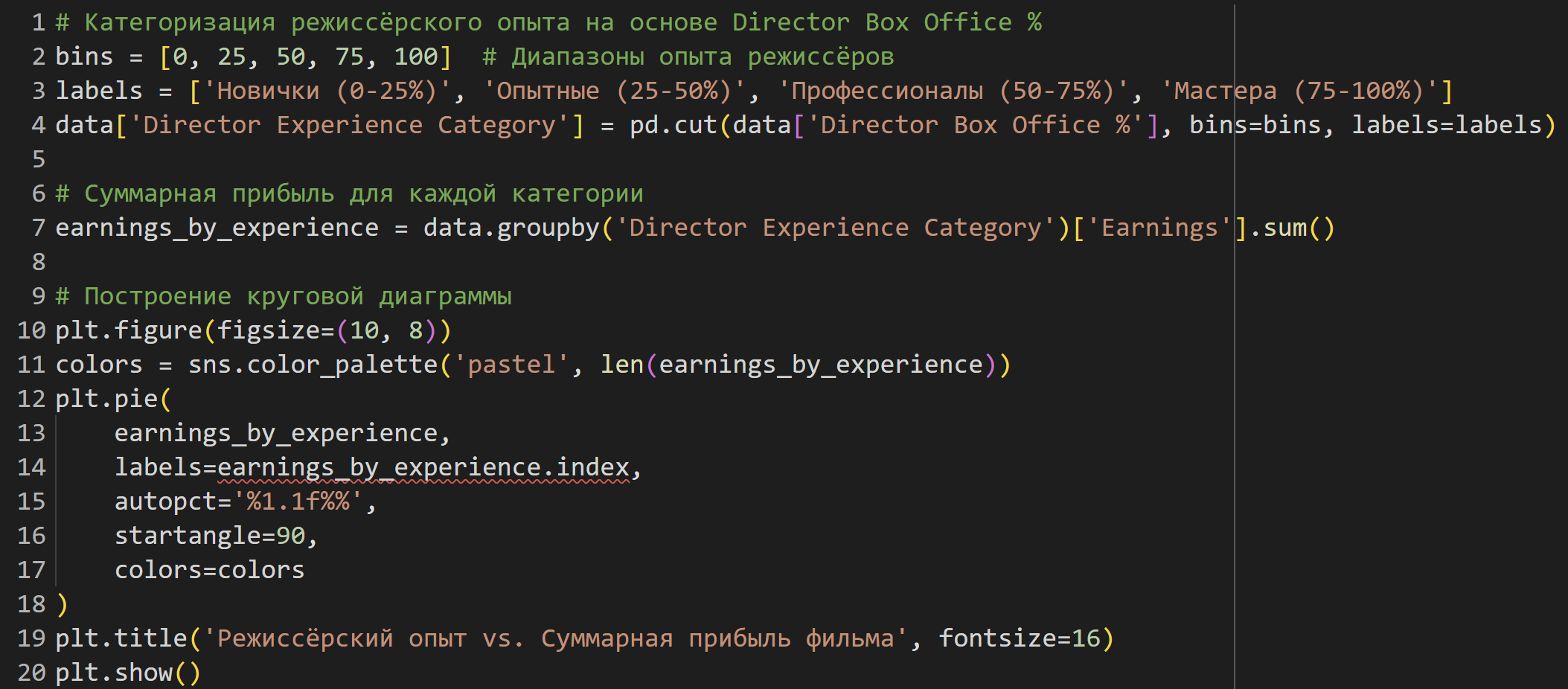
Полученный график выглядит так

****

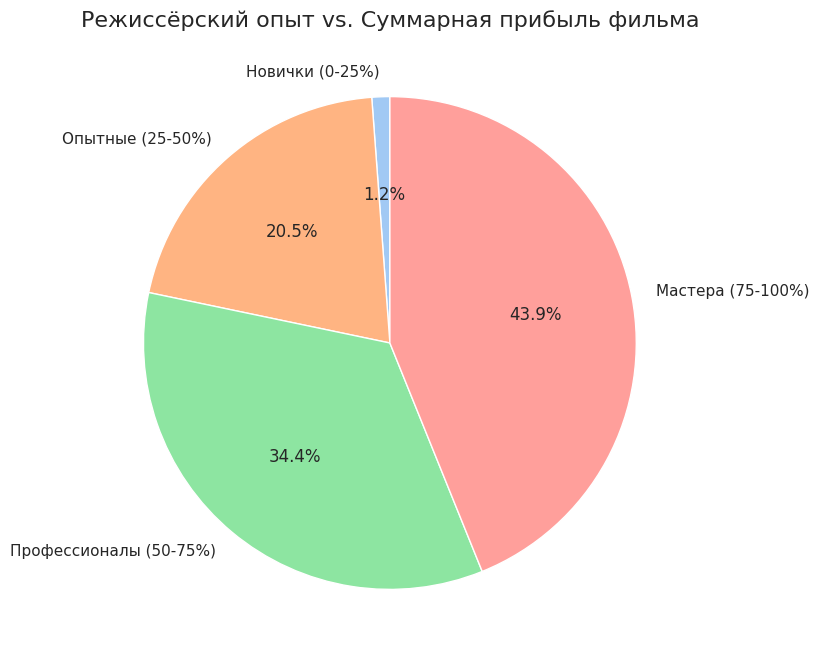
Видно, что далеко не всегда фильмы с большим бюджетом удостаиваются большого количества наград. Даже напротив, фильмы, получившие максимальное число наград, имеют очень маленький бюджет.

**Вывод:**Гипотеза 1 – не подтверждена.

## **Гипотеза 2 - Режиссерский опыт увеличивает прибыльность фильма**



Полученный график выглядит так

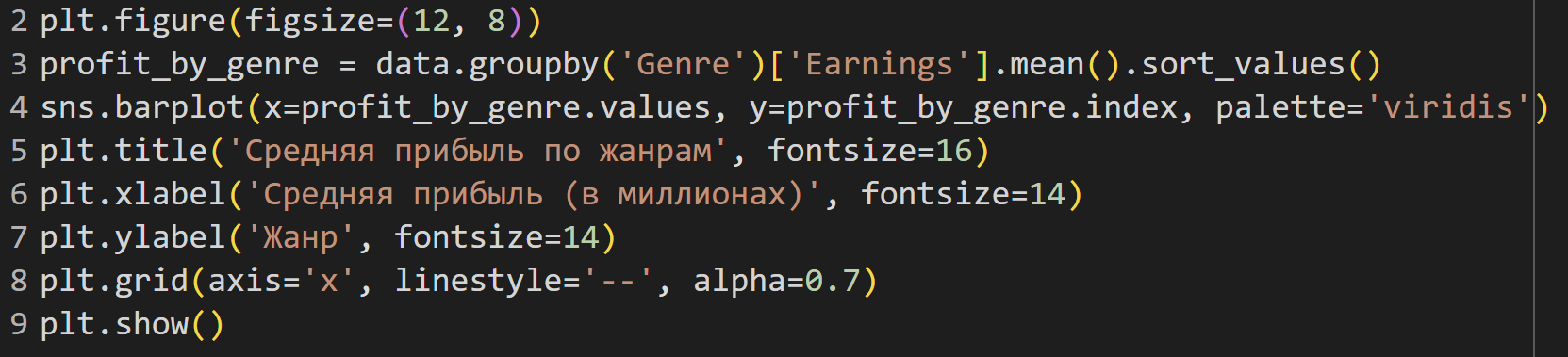


Действительно, новичкам в новом для себя амплуа режиссера очень сложно сделать большую прибыль с фильма, в то время как мастера своего дела с этим успешно справляются.

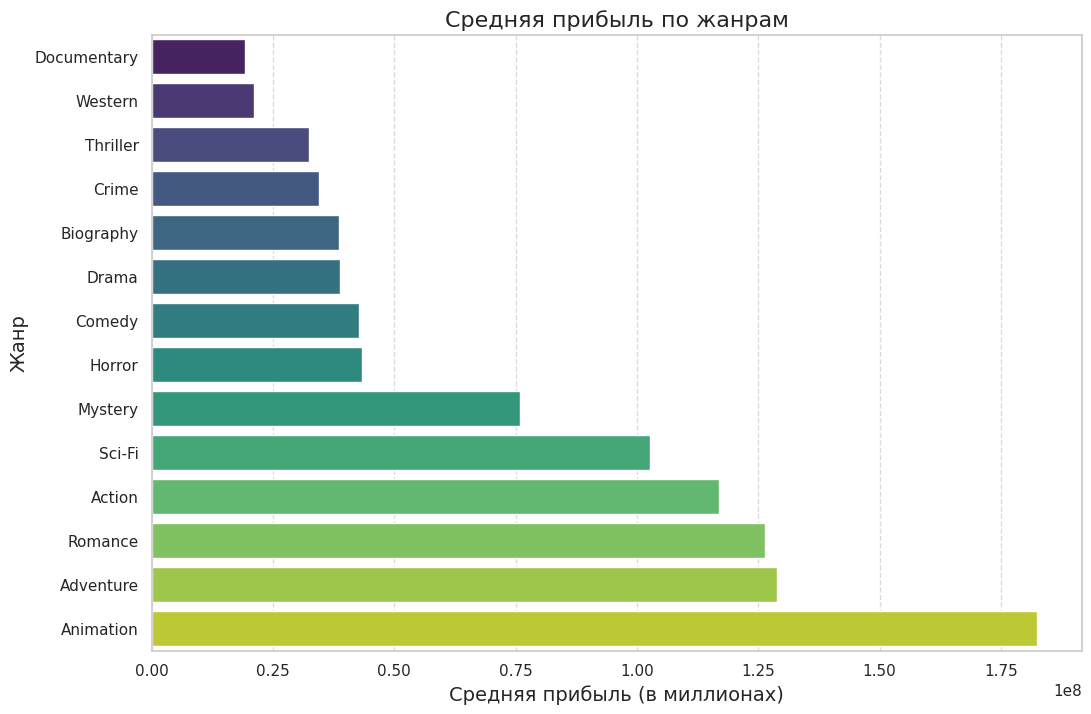
**Вывод:**

Гипотеза 2 – подтверждена

## **Гипотеза 3 - Влияет ли жанр на среднюю прибыль**



Полученный график выглядит так



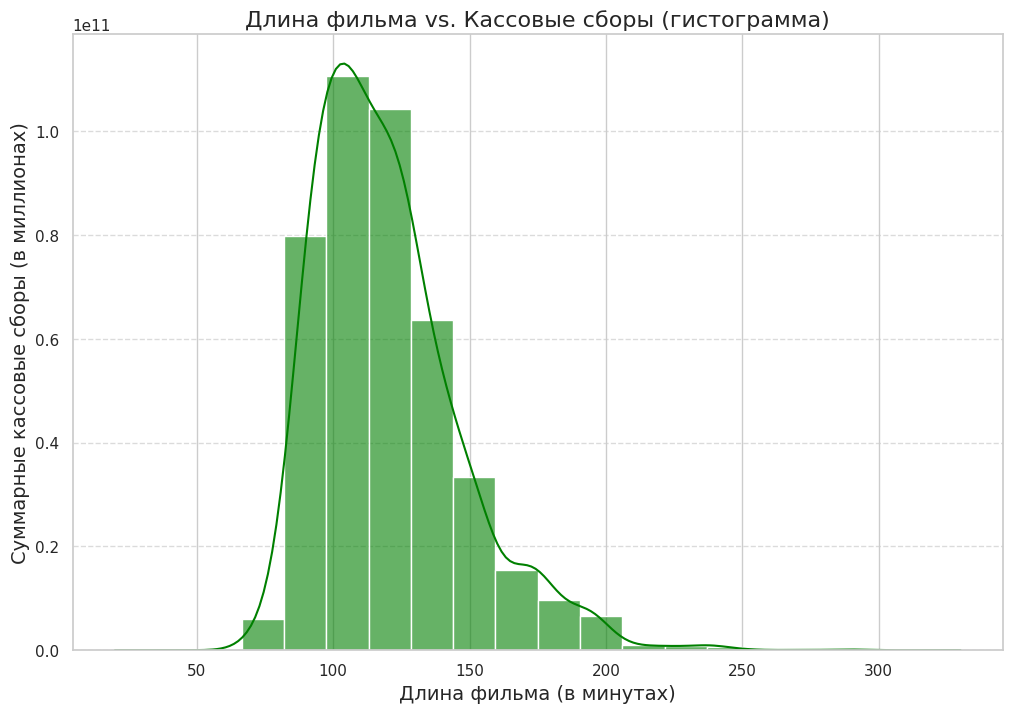
Как видно из графика, жанр действительно влияет на прибыльность фильма. Например, анимированные фильмы(мультфильмы) делать в разы выгоднее, нежели серьезные документальные фильмы.

**Вывод:**

Гипотеза 3 – подтверждена.

## **Гипотеза 4 - Длина фильма влияет на кассовые сборы**

Полученный график выглядит так

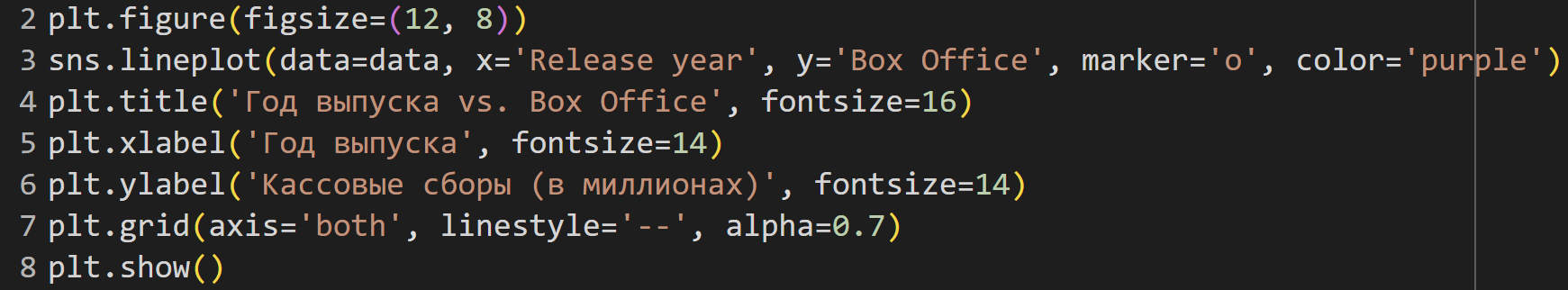


Как мы можем заметить, влияние длины фильма на его кассовые сборы значительно. Если есть амбиции хорошо заработать на фильме, то следует его делать продолжительностью около 100-а с небольшим минут.

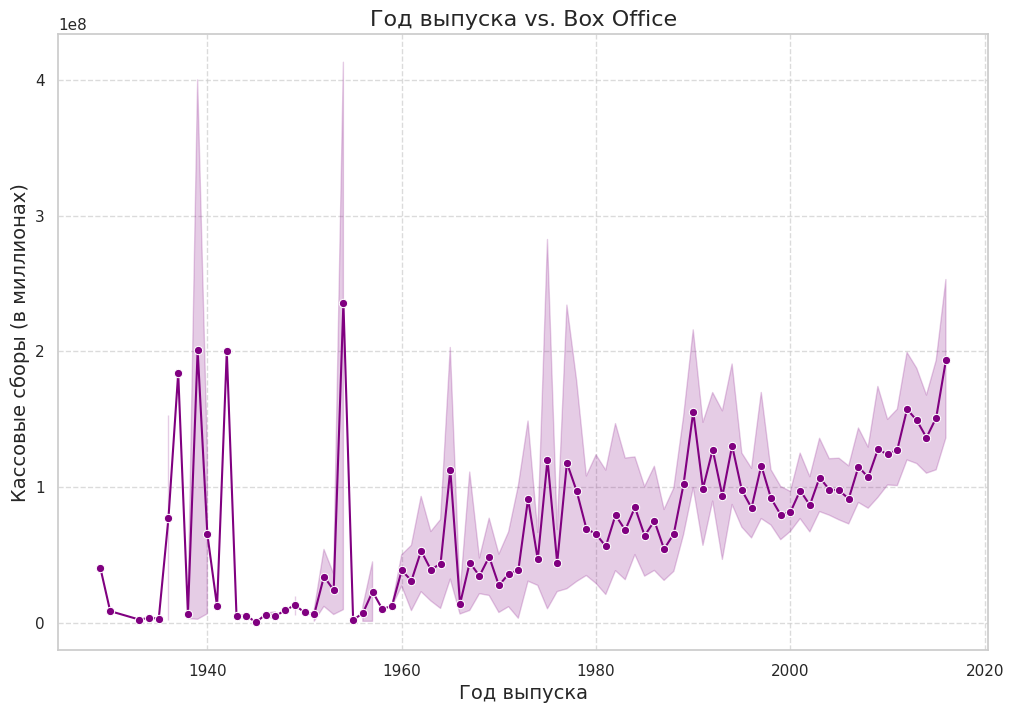
**Вывод:**

Гипотеза 4 – подтверждена.

## **Гипотеза 5 - Современные фильмы имеют лучшую доходность**



Полученный график выглядит так

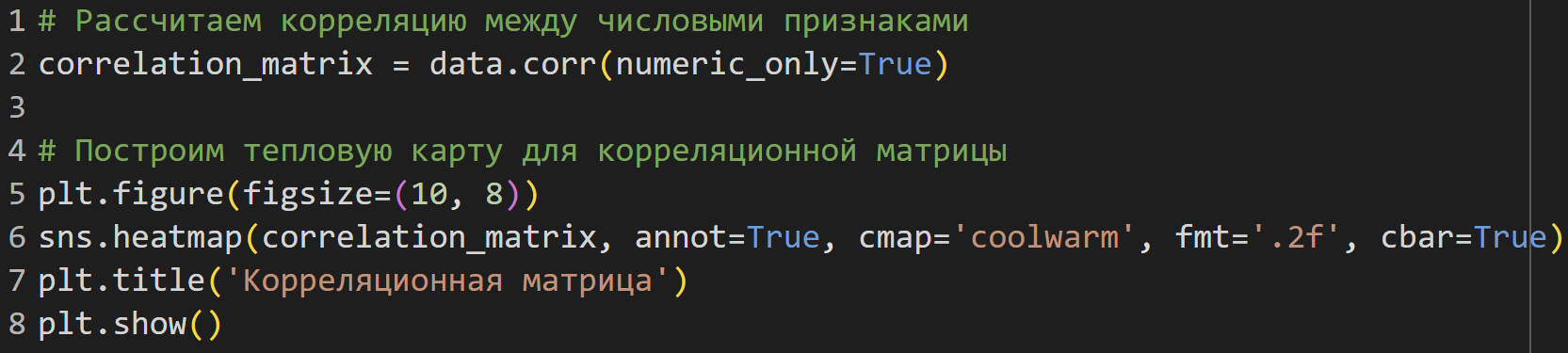


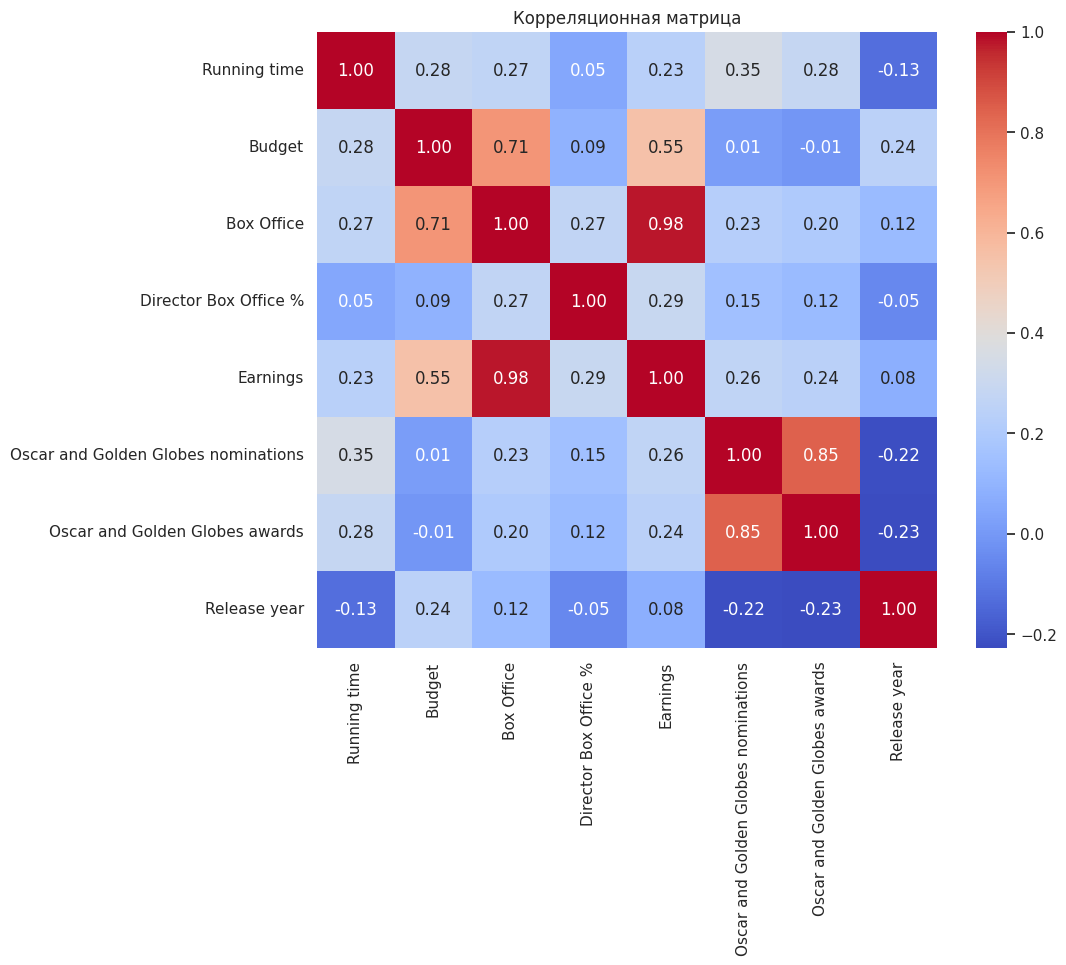
Глядя на график, можно сделать неоднозначные выводы. Глобально, средние кассовые сборы современных фильмов действительно увеличились, однако находятся старые классические фильмы, сборы которых до сих пор не удалось перебить.

**Вывод:**

Гипотеза 5 – подтверждена частично.

## **Анализ корреляции**

****

****

**Выводы корреляционного анализа:**

# **Сильная связь между Box Office и Earnings (0.98):** Это логично, так как разница между доходами (Box Office) и бюджетом напрямую формирует прибыль (Earnings).

# **Умеренная корреляция между Budget и Box Office (0.71):** Чем больше бюджет фильма, тем выше вероятность получения большого дохода в прокате, хотя это не гарантирует успех.

# **Связь между Oscar and Golden Globes nominations и Oscar and Golden Globes awards (0.85):** Фильмы, которые получают больше номинаций, часто выигрывают больше наград.

# **Слабая отрицательная корреляция между Release year и Oscar and Golden Globes awards/nominations (-0.22 и -0.23):** Фильмы, выпущенные в более ранние годы, чаще получали номинации и награды.

# **Отсутствие значимой связи между Director Box Office % и другими параметрами:** Этот показатель слабо коррелирует с другими метриками, такими как бюджет или доходы, что говорит о том, что успешность режиссёра не всегда влияет на результаты конкретного фильма.

# **Связь между Running time и Oscar and Golden Globes nominations (0.35):** Фильмы с большей длительностью, как правило, имеют больше номинаций, возможно, из-за более сложных сюжетов или серьёзного подхода к созданию.

# **Слабая положительная связь между Budget и Earnings (0.55):** Это говорит о том, что более дорогие фильмы чаще оказываются прибыльными, но не всегда.

# **ЗАКЛЮЧЕНИЕ**

В ходе исследовательской работы был проведен анализ базы данных "Movies", выявлены важные статистические данные, которые могут помочь в понимании факторов, влияющих на киноиндустрию.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas», «Seaborn» и «Matplotlib» на базе языка программирования Python. Также использовался «Google Colab»: в нём был написан код и анализированы диаграммы. В ходе работы я получил новые навыки в области анализа данных и познакомился с вышеуказанными программами.

Была осуществлена оптимизация использования памяти, агрегирование данных, что в дальнейшем позволило провести необходимые в ходе работы исследования и получить ценные знания в предметной области.

# **СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ**

1. БУРНАШЕВ Р.А. Методические указания по программному обеспечению «Pandas». - 1 изд. - Казань: КФУ, 2022. - 25 с.
2. Методические указания по программному обеспечению «Seaborn» // [Электронный ресурс]: seaborn.pydata URL: https://seaborn.pydata.org/tutorial.html (дата обращения: 14.09.2024).
3. Методические указания по программному обеспечению «Plotly» // [Электронный ресурс]: habr.com URL: https://habr.com/ru/companies/skillfactory/articles/506974/ (дата обращения: 15.09.2024).
4. «Pandas. Работа с данными» // [Электронный ресурс]: vk.com URL: https://vk.com/wall-159224823\_92912 (дата обращения: 15.09.2024).
5. «Python. Визуализация данных: Matplotlib, Seaborn, Mayavi» // [Электронный ресурс]: vk.com URL: https://vk.com/wall-192648009\_320 (дата обращения: 15.09.2024).