МИНОБРНАУКИ РОССИИ

## Федеральное государственное автономное образовательное

## учреждение высшего образования

## «ЮЖНЫЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Институт математики, механики и компьютерных наук им. И.И. Воровича

01.03.02 — Прикладная математика и информатика

**Проект 2-го курса**

**ОТЧЕТ**

студентов 2 курса

*Столбовой Кристины Михайловны*

*Бережной Ярославы Вячеславовны*

**Преподаватель \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_***/к.ф.-м.н.,**доцент А.П. Мелехов* /

Ростов-на-Дону

2022 г.

**Исследовательский анализ данных.**

Наша задача: дать краткую информацию о наборе данных.

Среда разработки проекта – *kaggle.com (это специальная платформа для работы с данными)*

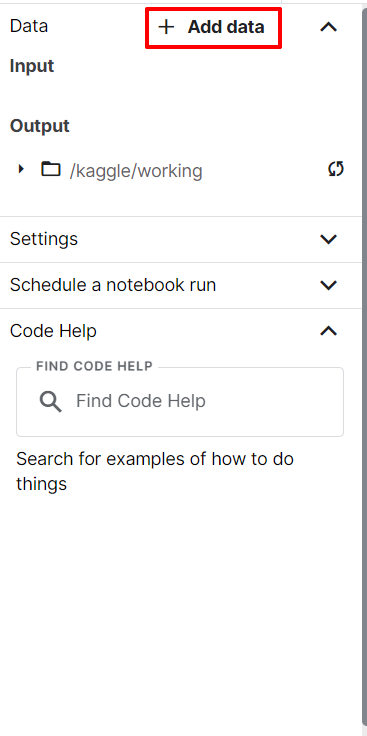
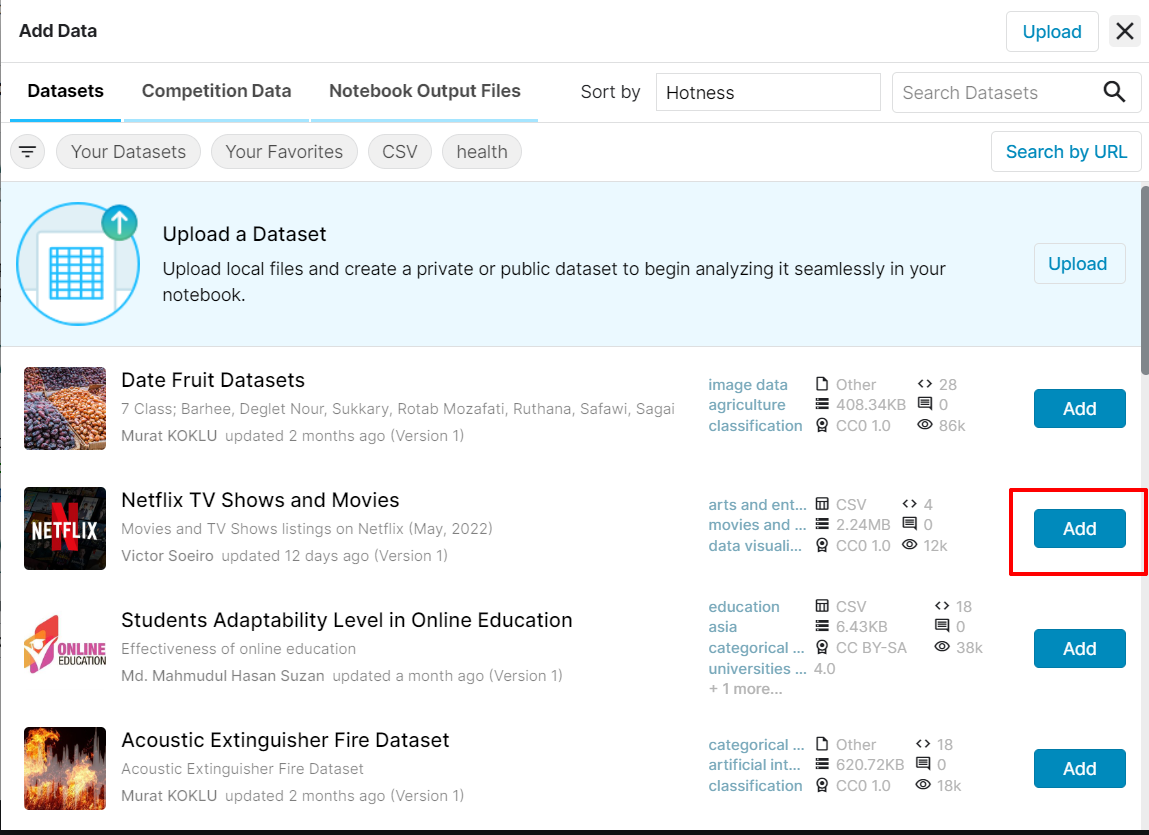
**Приступим к выполнению поставленной задачи.**

* Создаем New Notebook

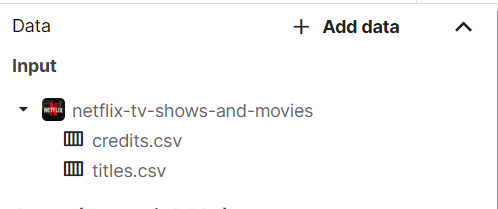


Вначале необходимо загрузить данные, с которыми нам предстоит работать (в нашем случае информация о фильмах).

* Для загрузки данных проводим следующие манипуляции:

* Выбранные нами данные подключены к проекту!



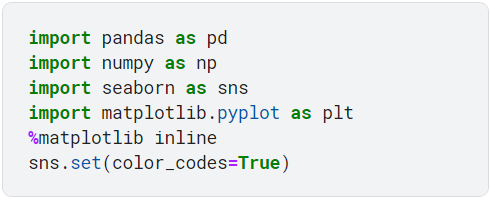
* Импорт необходимых библиотек.

В своей работе мы использовали следующие библиотеки ***Python:***

***Pandas*** *#библиотека, которая предоставляет инструменты для хранения данных и работе с ними.*

***Numpy*** *#библиотека, добавляющая поддержку больших многомерных массивов и матриц.*

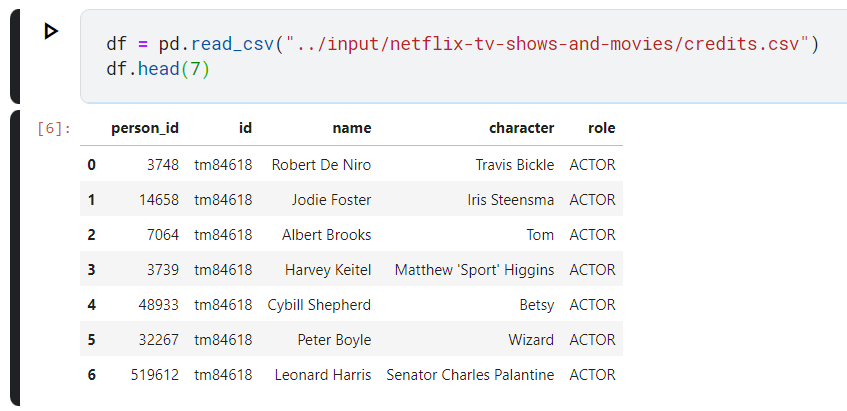
***Seabor, Matplotlib*** *#библиотеки для визуализации*



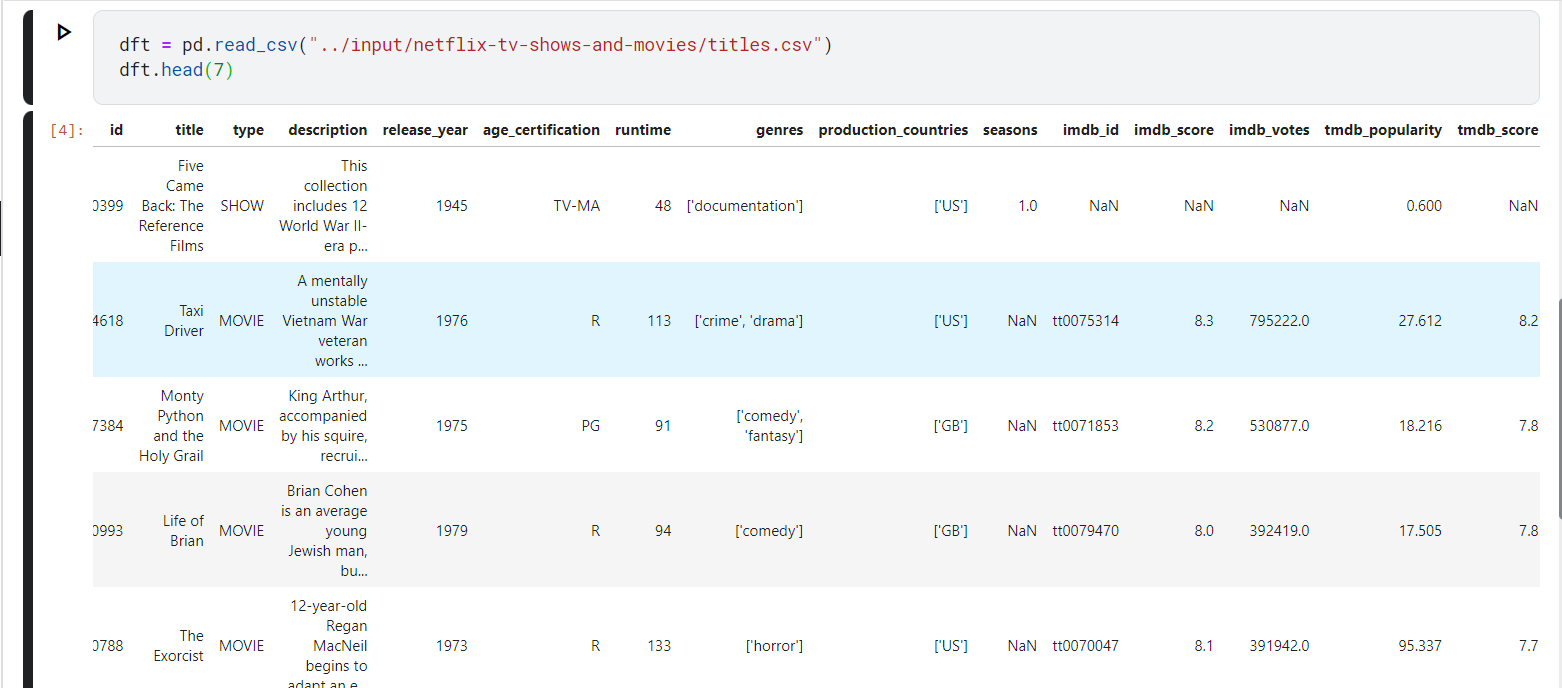
* Посмотрим на данные, которые предоставляет нам kaggle

Мы имеем 2 файла:

* credits.csv



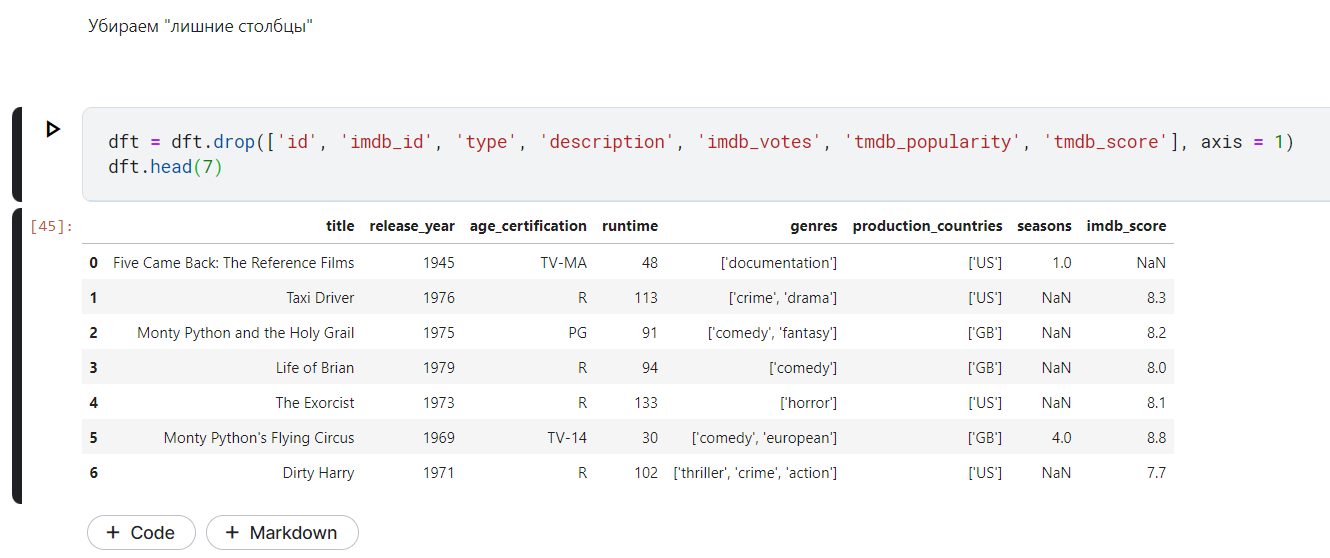
* ttitles.csv



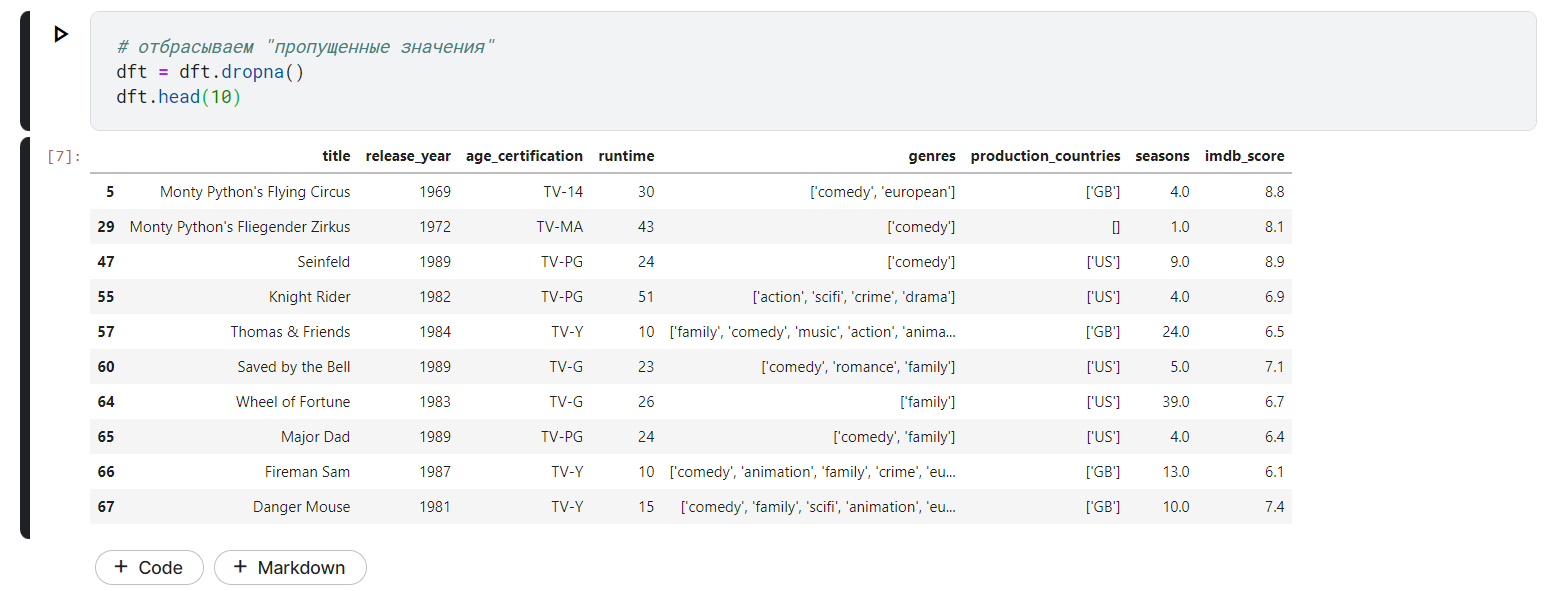
# Команда *df(dft).head(n)* показывает нам первые n строчек с информацией из таблицы данных.

Мы видим, что 2ая таблица содержит в себе больше какой-либо информации, поэтому будем работать именно с ней.

Прежде чем приступить к работе уберем из таблицы «лишние» столбцы (id, imdb\_id, type, description, imdb\_votes, tmdb\_popularity, tmdb\_score)



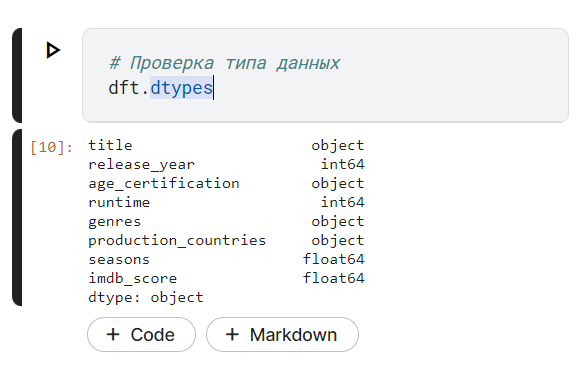
Мы видим, что в таблице есть значения *NaN* впоследствии при анализе строк таблицы, которые содержат данное значения, могут возникнуть трудности, поэтому уберем эти строчки.



При последующей работе с данными нам может помешать следующий факт: в таблице могут быть повторяющиеся строчки/данные. Решим эту проблему:



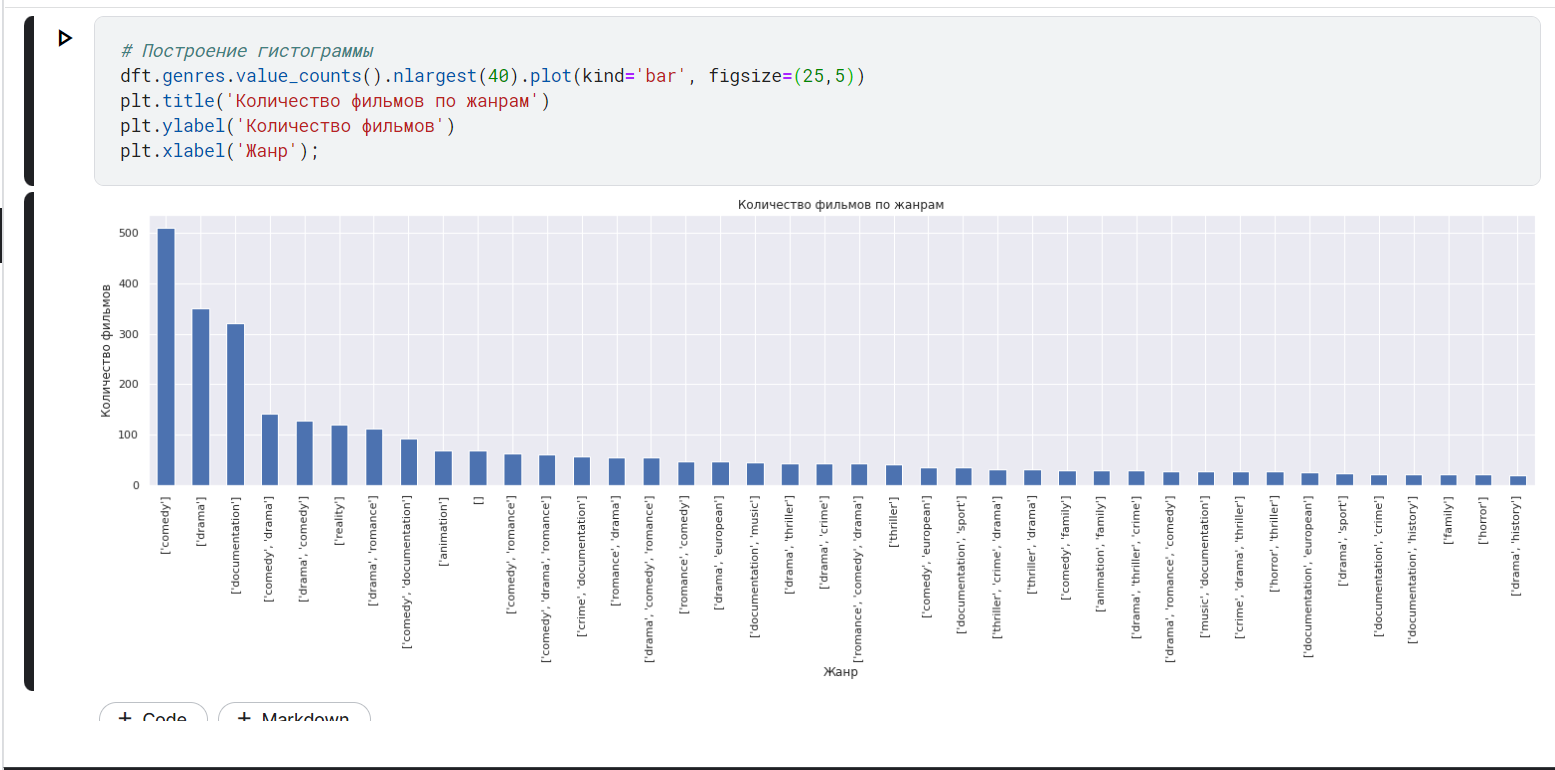
Как мы видим, данные попались «хорошие». Переходим к следующему шагу проверки данных на качество. Для этого проверим все ли данные такого типа, какого мы планируем (т.е. например, длительность фильма хранится в числовом формате)

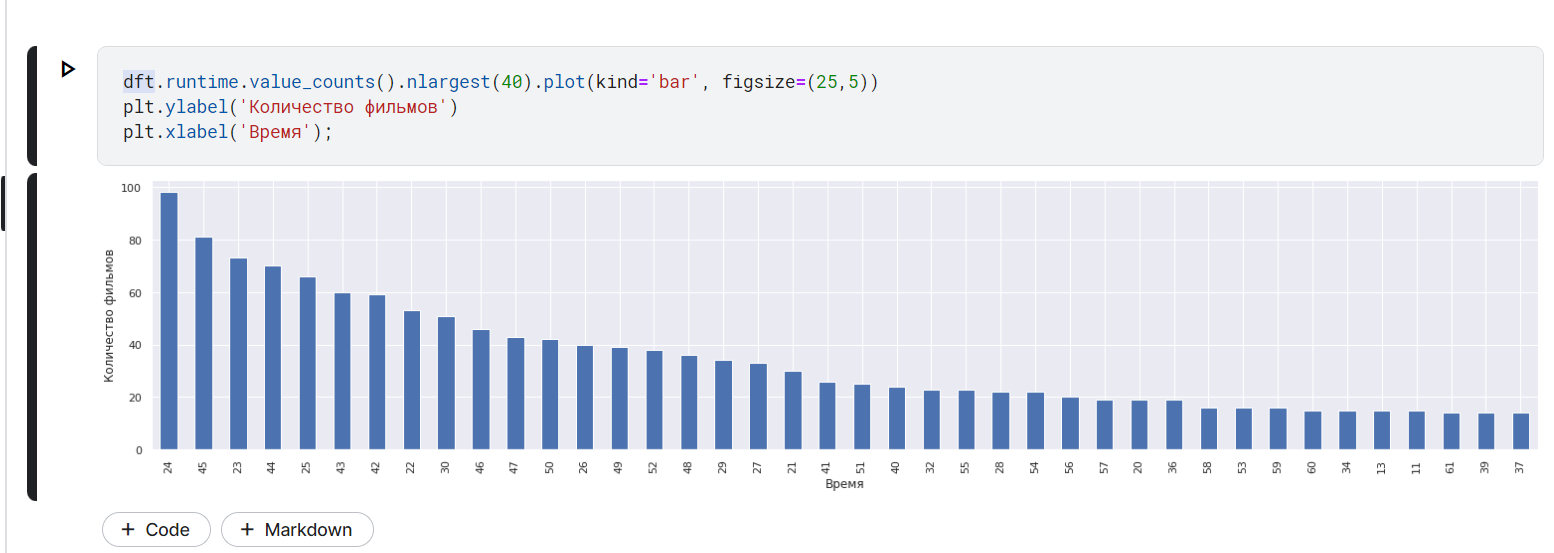
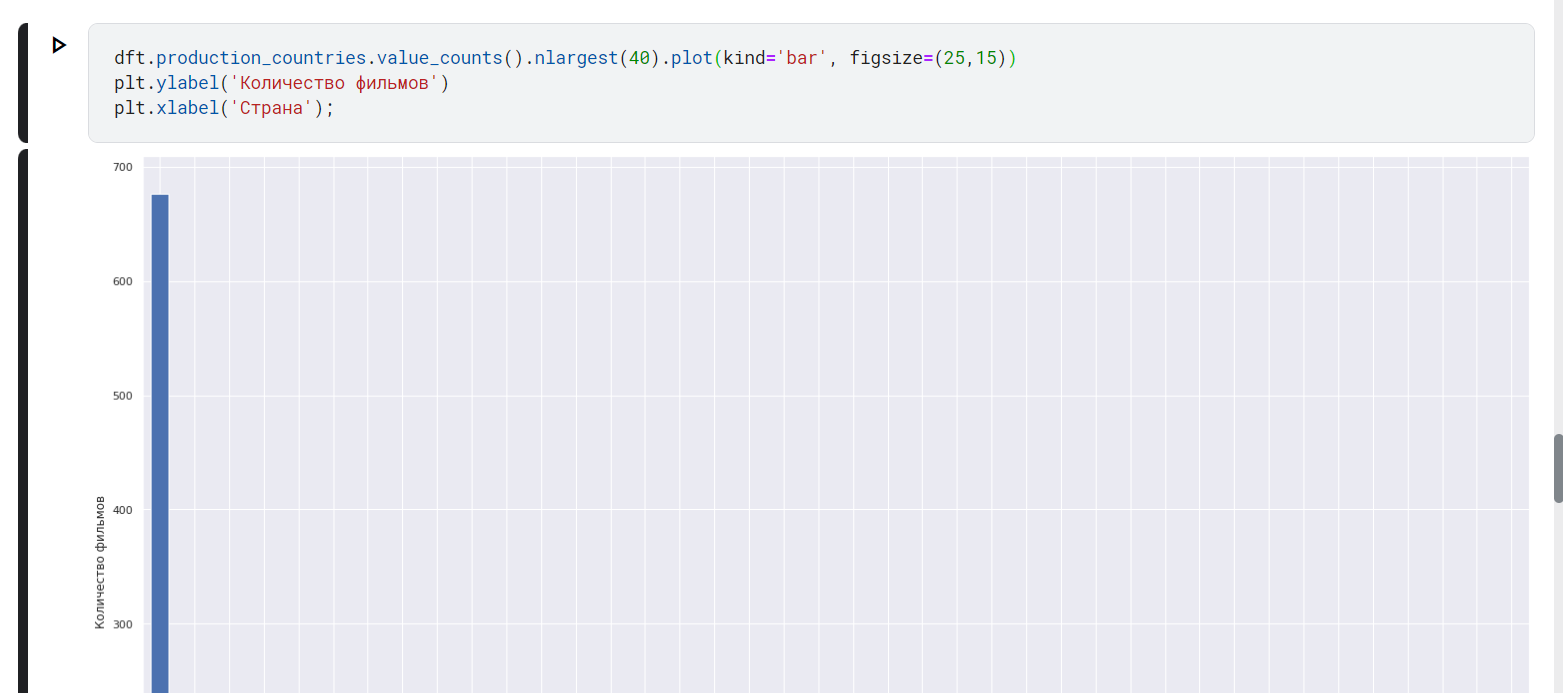
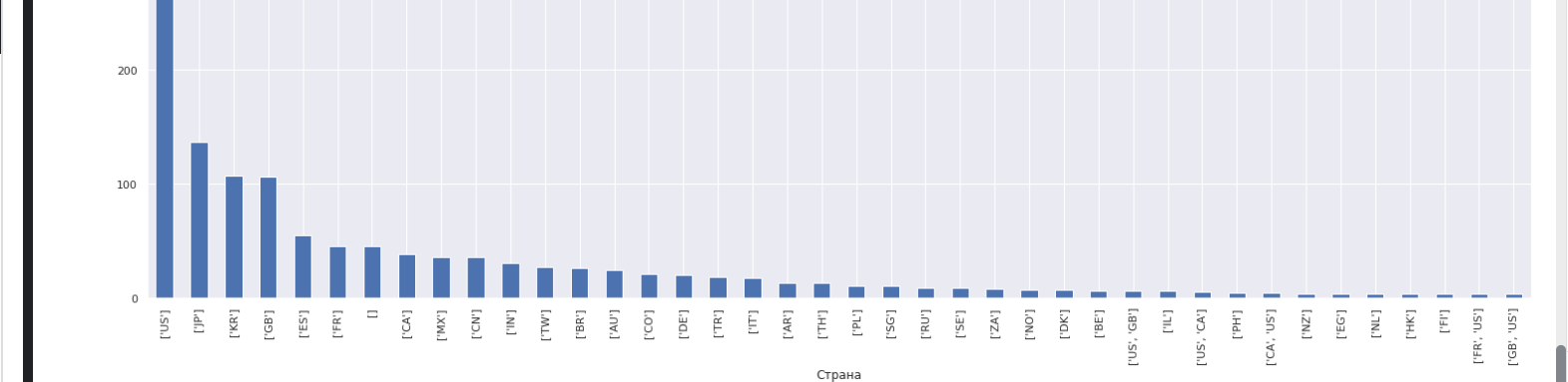


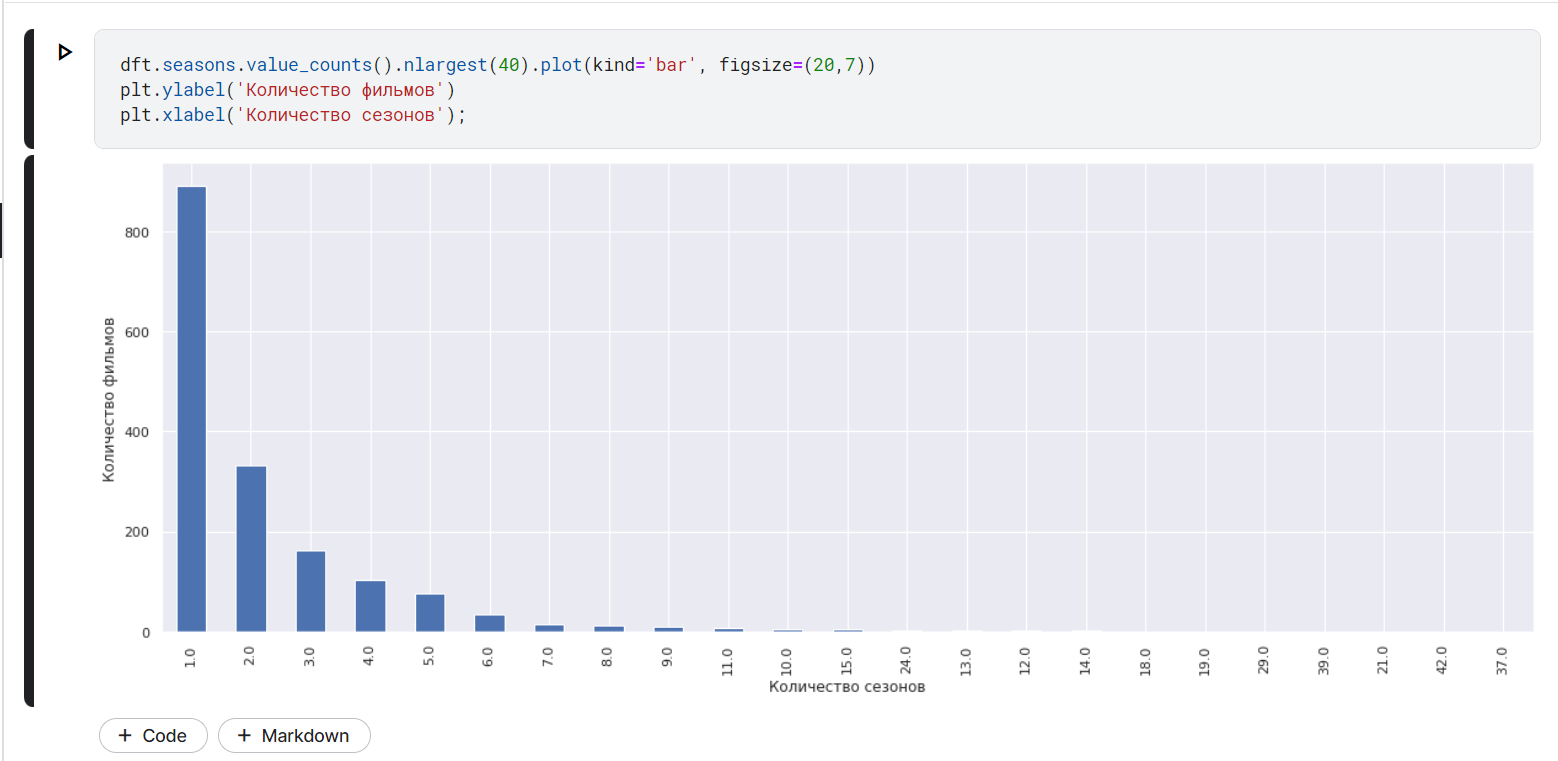
Все хорошо. Наше вмешательство не требуется.

Теперь, когда данные проверены, мы можем с ними работать.

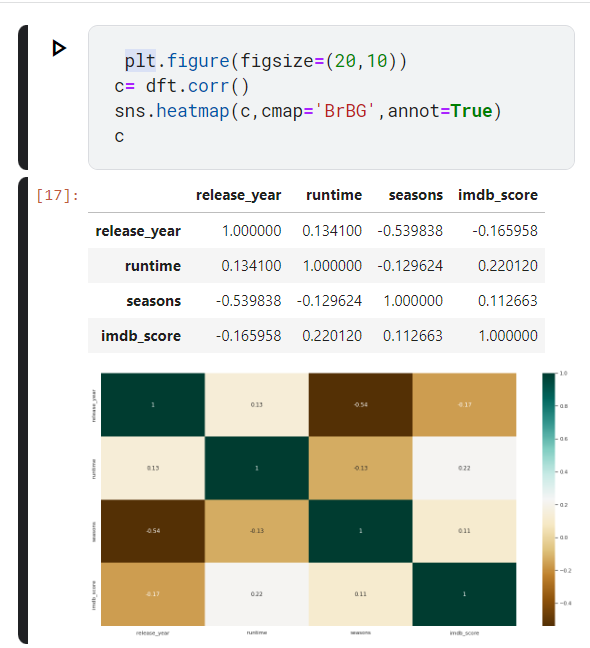
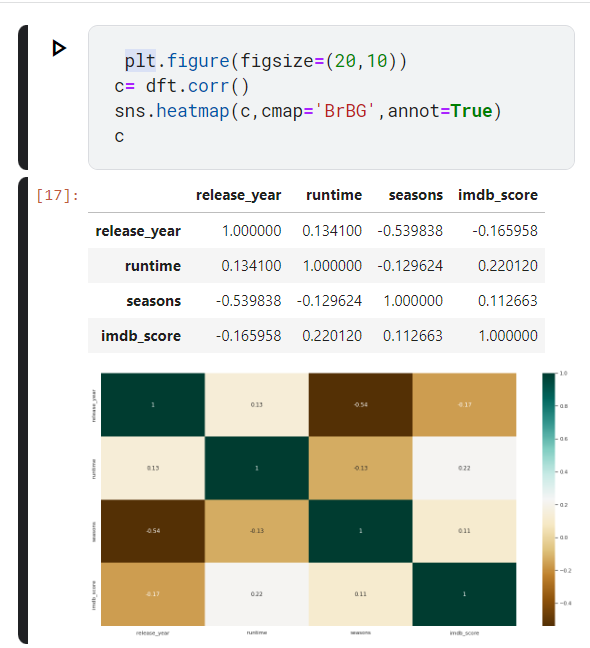
Посмотрим, на жанры фильмов, их длительность и страну выпуска.



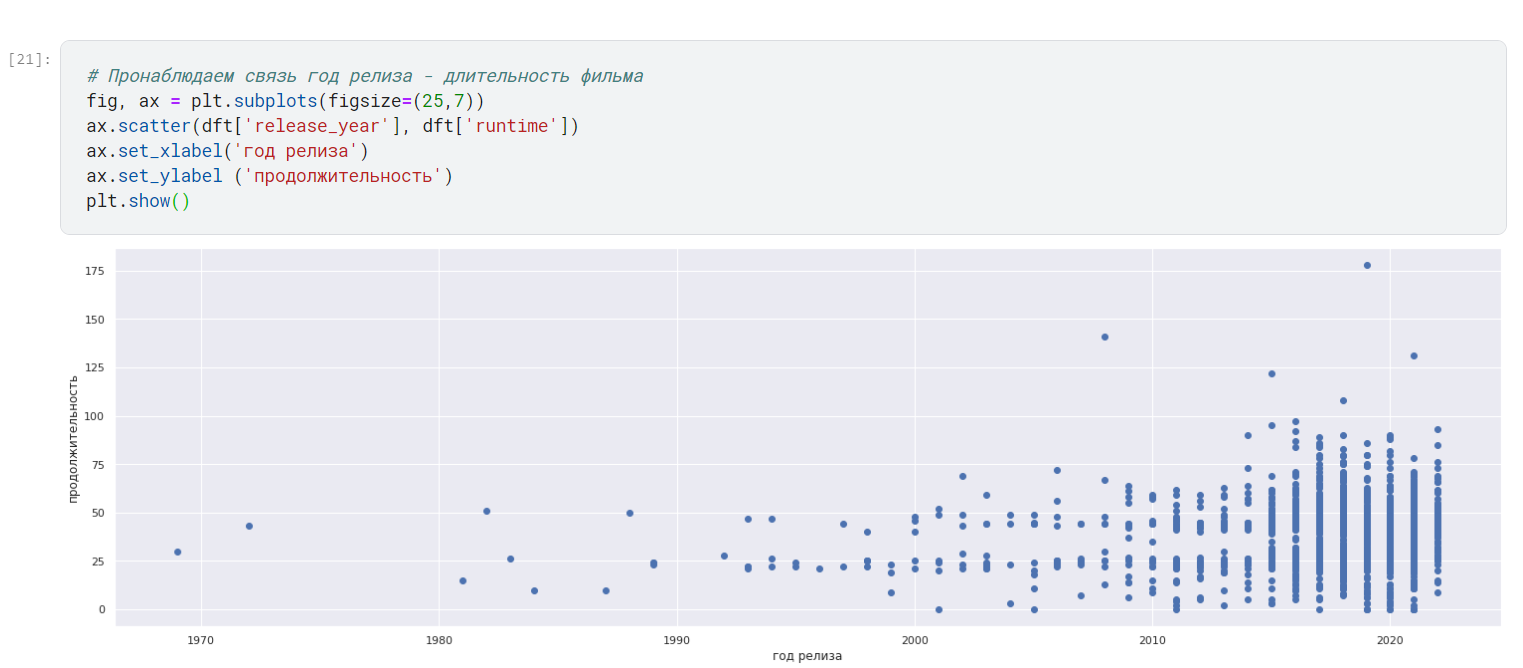


Посмотрим на взаимосвязь между годом выпуска, продолжительности, количества сезонов и рейтингом фильмов. Так как выше перечисленная информация хранится численном типе данных, нам в это поможет тепловая карта.



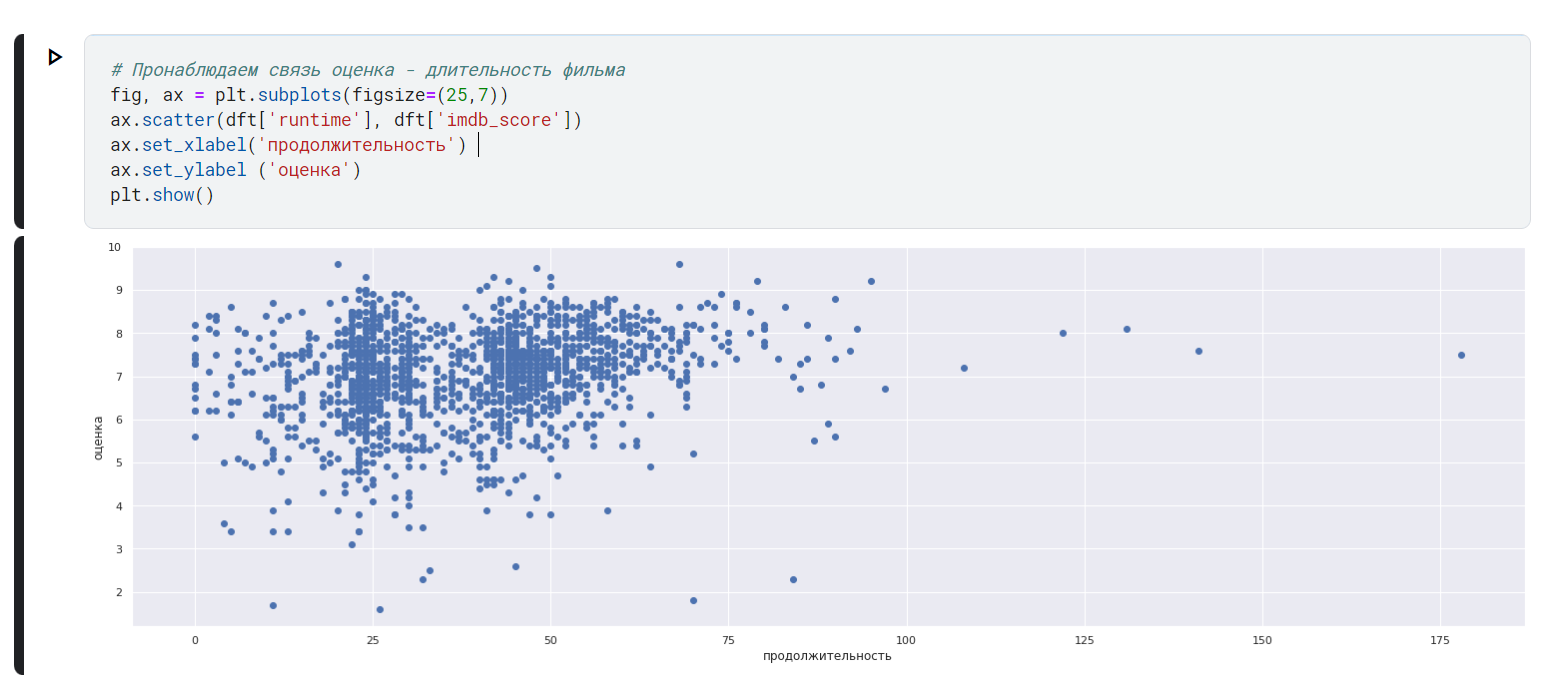
Рассмотрим имеющиеся зависимости. В этом нам поможет *диаграмма рассеяния.*

Проанализируем связь год релиза - длительность фильма:



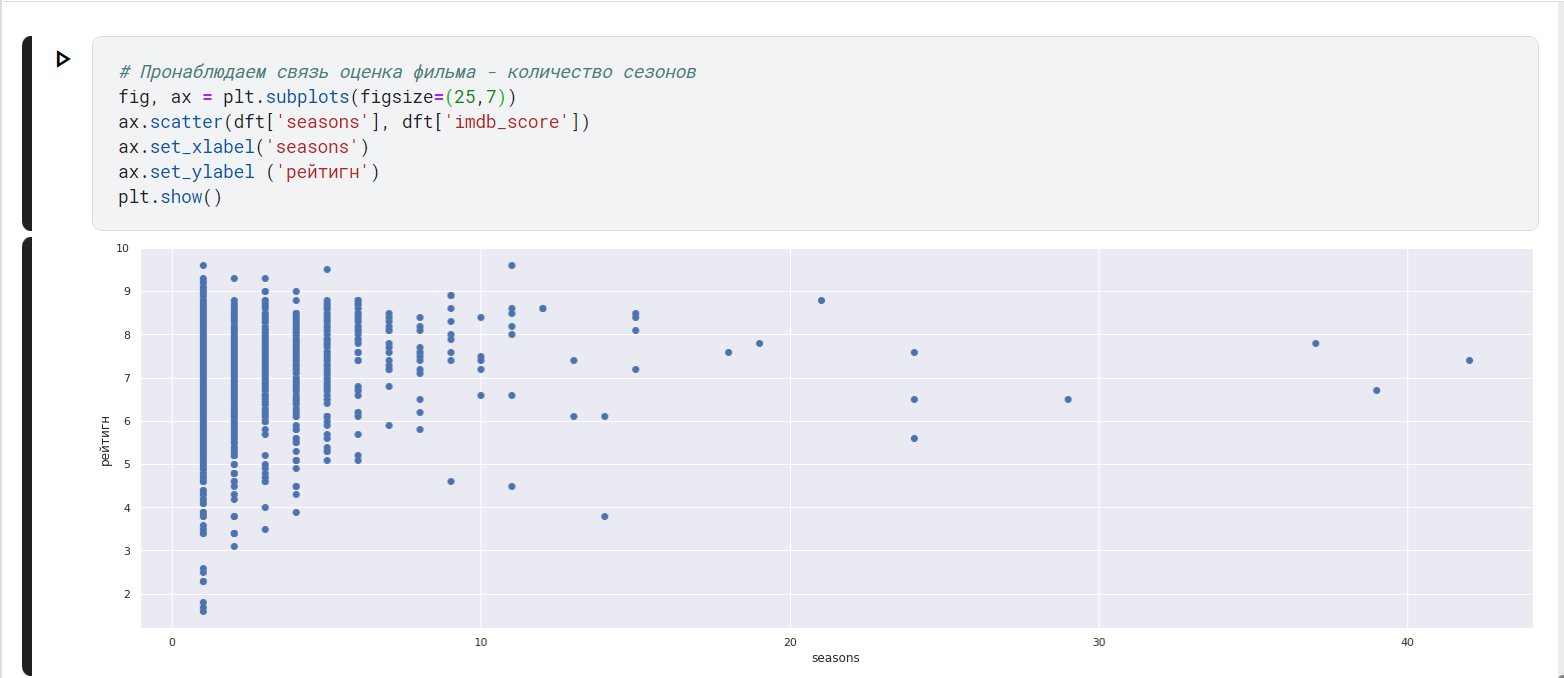
На диаграмме видно, что в момент зарождения кинематографии фильмы длились не более часа. Также мы можем проследить количество вышедших за год фильмов, и вплоть до 1990х мы можем говорить о единичных выпусках. Однако в процессе развития кинематографии мы видим как увеличении количества продукции, так их продолжительности. Уже после 2010 года можно найти фильмы, которые длятся буквально пару минут так и те, что придется смотреть несколько часов.

Проанализируем связь оценка - длительность фильма:



Однозначного ответа на вопрос какая взаимосвязь между длительностью и баллом фильма дать нельзя. Однако хорошо видно, что фильмы продолжительностью более 100 минут имеют достаточно высокую оценку. Это может свидетельствовать о том, что при работе над более большим проектом создатели прикладывают больше усилий и стараются сделать свою работу хорошо.

Проанализируем связь рейтинг – количество сезонов:



Мы видим несколько проектов, у которых достаточно высокий рейтинг, но при этом они имеют не более 10 сезонов. Также есть несколько фильмов с рейтингом ниже 5, но количеством сезонов 10-20. Что повлияло на такие результаты? Думаю однозначно ответить на этот вопрос нельзя, так как кроме данных представленных в таблице есть еще ряд факторов влияющих на выпуск фильмов (например, уход актера из проекта, нехватка финансов и т.д.).

Но, несмотря на «исключения» правила, чем выше рейтинг, тем больше сезонов мы видим 3 фильма с рейтингом ~7 количество сезонов, которых близко к 40!

После проведенного анализа данных, к сожалению, нельзя точно выразить, отчего зависит длительность фильма, его рейтинг, количество сезонов, год показа и т.д. Однако поисследовав эти данные можно установить некоторые закономерности и аналогии.