|  | **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\_\_\_\_\_\_\_\_\_\_\_ «Top Video Games 1995-2021 Metacritic»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

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

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

Студент ИУ5-51Б **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_\_Цыпышев Т.А.\_\_

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

Руководитель **\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_ МаслениковК.Ю.\_

(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 «Top Video Games 1995-2021 Metacritic»\_\_

Студент группы \_\_\_\_\_\_\_ИУ5-51Б\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_Цыпышев Тимофей Александрович\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

(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 «*Top Video Games 1995-2021 Metacritic*», провести визуализацию полученных данных посредством языка программирования Python*

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

Расчетно-пояснительная записка на \_\_ листах формата А4.

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

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

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

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

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

Дата выдачи задания «10» сентября 2024 г.

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

(Подпись, дата) (И.О.Фамилия)

**Студент \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_** \_Цыпышев Т.А.\_\_

(Подпись, дата) (И.О.Фамилия)

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

**Аннотация**

По заданию требуется проанализировать базу данных, очистить ее от

ненужных данных, также нужно выдвинуть гипотезы, проанализировать их и составить графики по данным гипотезам. Для данного задания была выбрана база данных “Top Video Games 1995-2021 Metacritic”. В ходе работы будут использоваться такие библиотеки как: «Numpy», «Pandas» и «Seaborn». Анализ будет проводиться через программу «VS code» с использованием локального окружения «Python» и нативного «Jupyter 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 об оценках самых популярных игр 1995-2021 годов. В конце, будет сделан вывод о проделанной работе с данным датасетом.

**СОДЕРЖАНИЕ**

[ВВЕДЕНИЕ 5](#_gjdgxs)

[1. Определение данных для анализа 6](#_1fob9te)

[2. Описание данных 6](#_3znysh7)

[3. Формулирование гипотез 6](#_2et92p0)

[4. Подготовка данных для работы 7](#_4d34og8)

[5. Построим карту корреляций 9](#_2s8eyo1)

[6. Проанализируем распределение всех числовых признаков 9](#_sez25wvm87jr)

[7. Проанализируем распределение всех категориальных данных 10](#_pcvm325olmxv)

[8. Построим более сложные графики 11](#_kxq8llddemp)

[9. Выдвинем свои гипотезы по датасету и проверим их 12](#_o2x7qx4t82te)

[a. Гипотеза 1: Игры на разных платформах имеют разные средние оценки. 12](#_66t5g8payz7o)

[b. Гипотеза 2: Высокая оценка критиков коррелирует с высокой пользовательской оценкой. 13](#_9rau3ov4cqgy)

[c. Гипотеза 3: Год выпуска игры влияет на её популярность (по средним оценкам). 14](#_hfc7v2u5cdxd)

[d. Гипотеза 4: Чем больше оценок от пользователей - тем нижи они в среднем. 15](#_4n959bia37gd)

[e. Гипотеза 5: Дата релиза имеет корреляцию с количеством отзывов пользователей. 16](#_daqplq47bfc9)

[ЗАКЛЮЧЕНИЕ 17](#_z337ya)

[СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 19](#_3j2qqm3)

# **ВВЕДЕНИЕ**

Необходимо исследовать базу данных и выявить закономерности цен на билеты.

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 «Top Video Games 1995-2021 Metacritic».

В датасете представлены самые популярные игры 1995-2021 годов, а также отзывы и оценки критиков и пользователей.

# **Описание данных**

Для анализа были собраны данные о всех перелетах, включенных в датасет. В наборе данных содержатся:

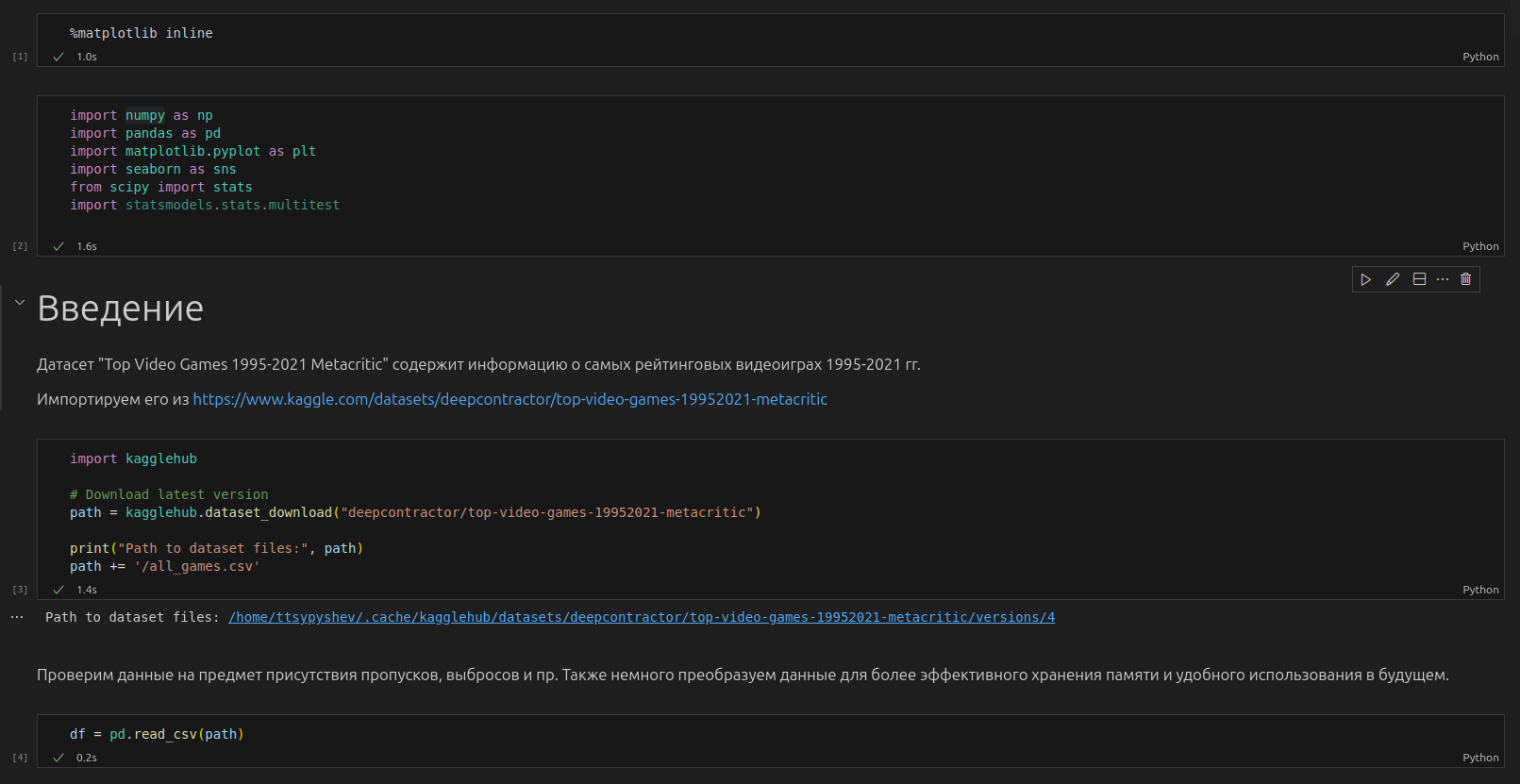
* **tbl**: идентификатор таблицы
* **name**: название игры
* **platform**: платформа, для которой выпущена игра
* **release\_date**: дата выхода игры
* **summary**: отзывы и справка об игре
* **meta\_score**: оценка критиков
* **user\_review**: оценка пользователей

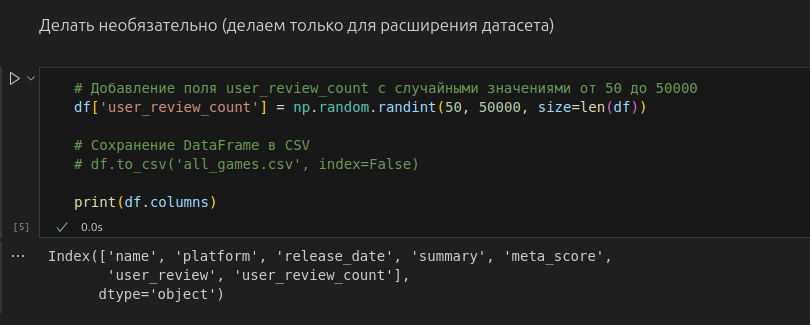
# **Формулирование гипотез**

* **Гипотеза 1**: Игры на разных платформах имеют разные средние оценки.
* **Гипотеза 2**: Высокая оценка критиков коррелирует с высокой пользовательской оценкой.
* **Гипотеза 3**: Год выпуска игры влияет на её популярность (по средним оценкам).
* **Гипотеза 4**: Чем больше оценок от пользователей — тем ниже они в среднем
* **Гипотеза 5**: Дата релиза имеет корреляцию с количеством отзывов пользователей

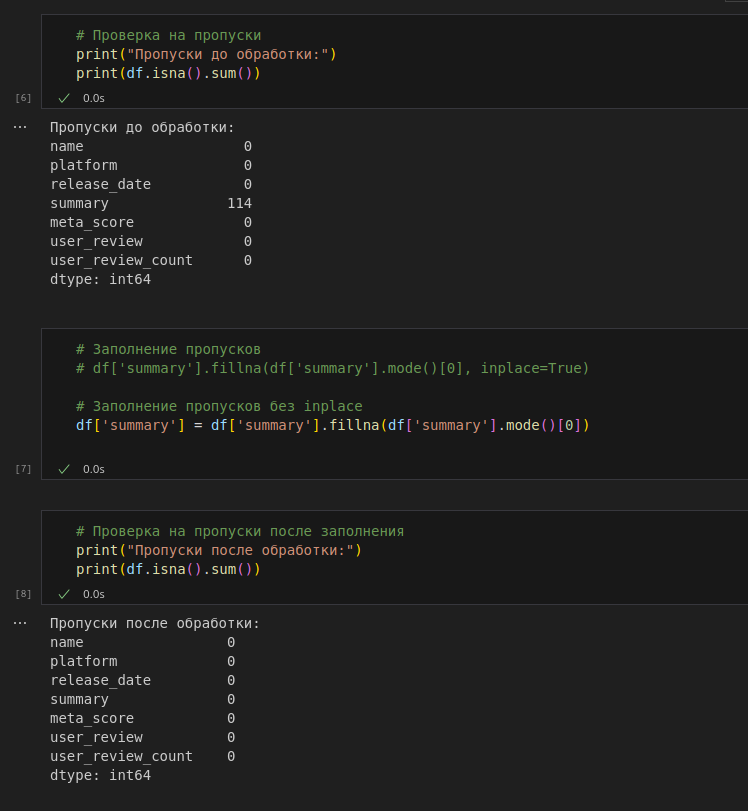
# **Подготовка данных для работы**

Загружаем датасет и подключаем необходимые библиотеки

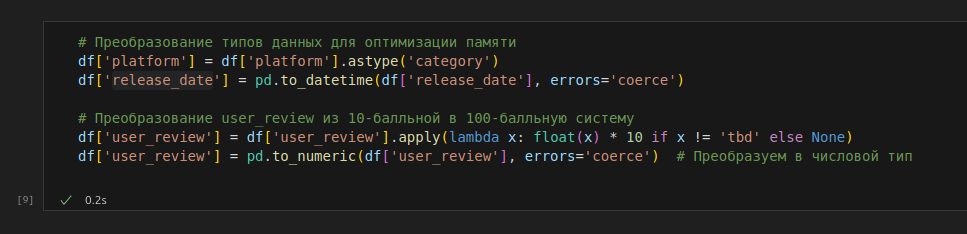
  
Добавляем дополнительные данные в таблицу



Проводим очистку данных



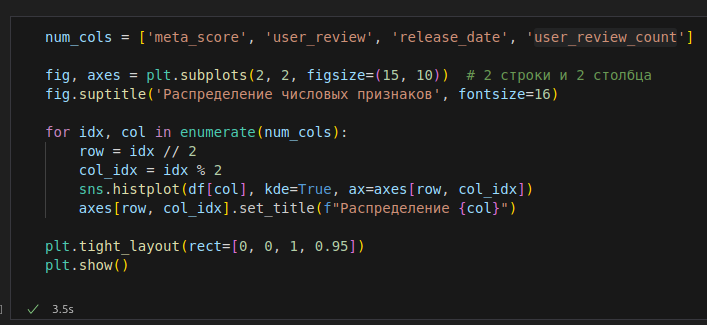
Выбираем из датасета поля, необходимые для анализа

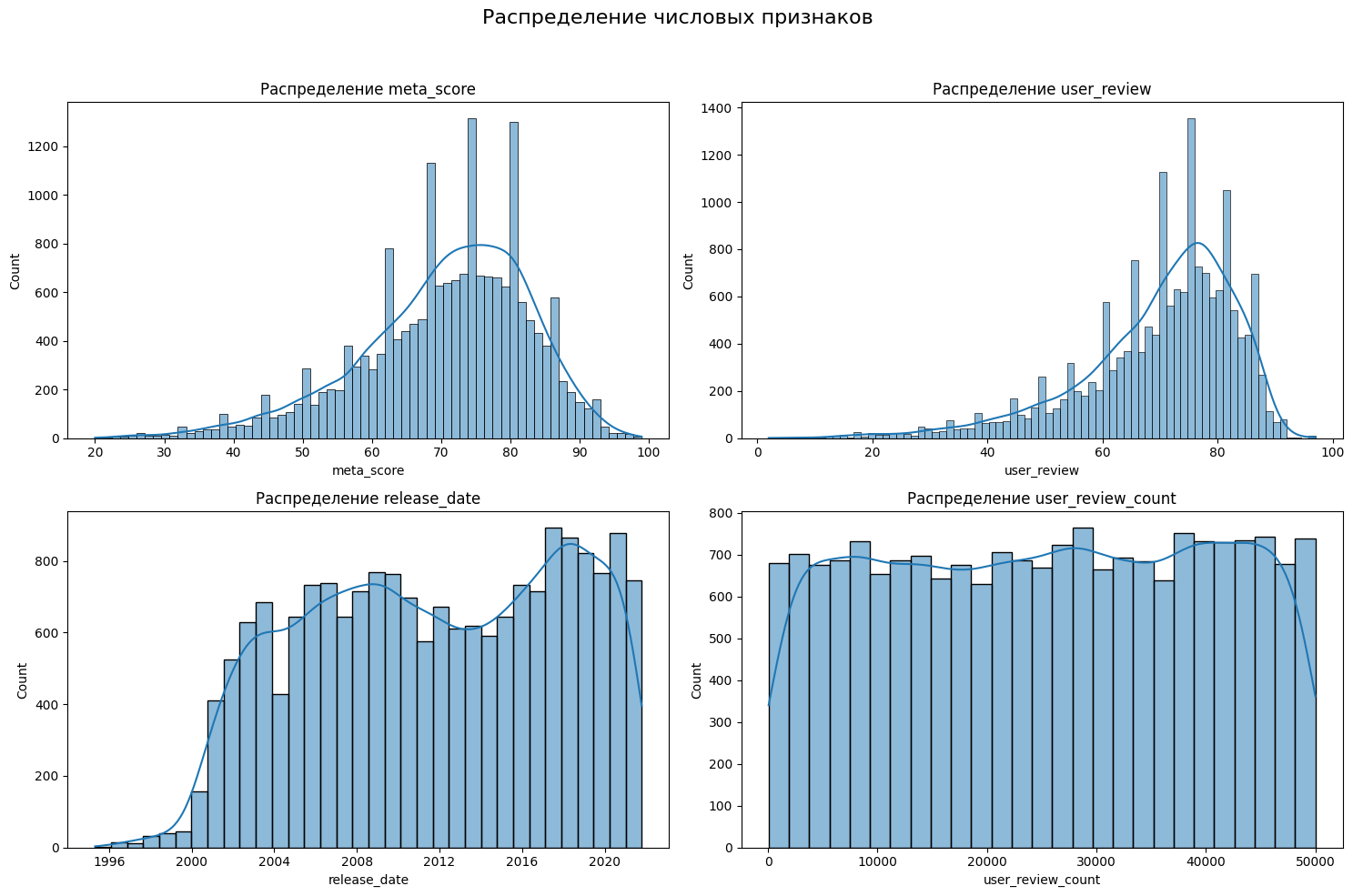


# **Построим карту корреляций**

## 

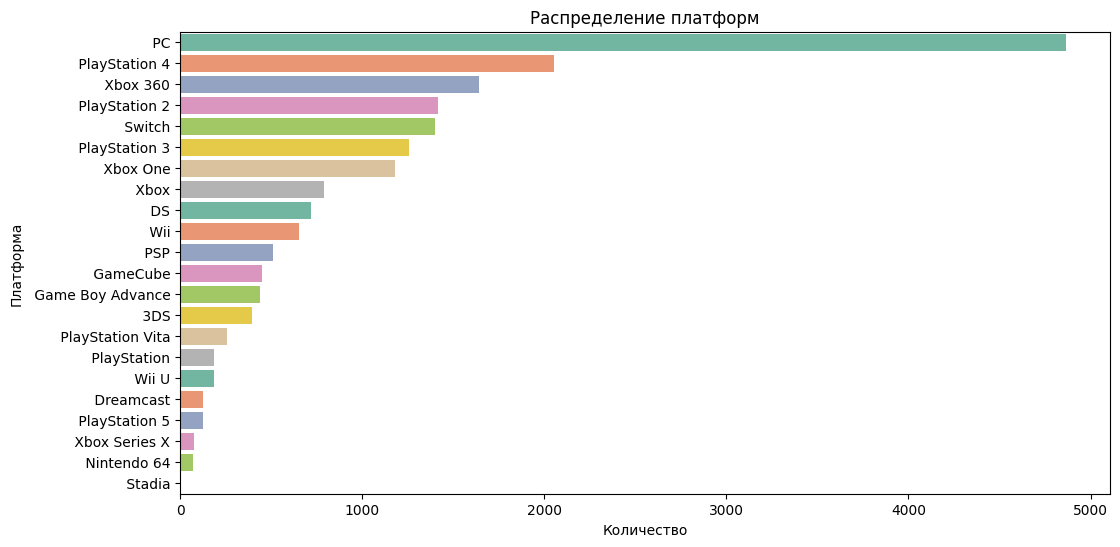
# **Проанализируем распределение всех числовых признаков**

****

****

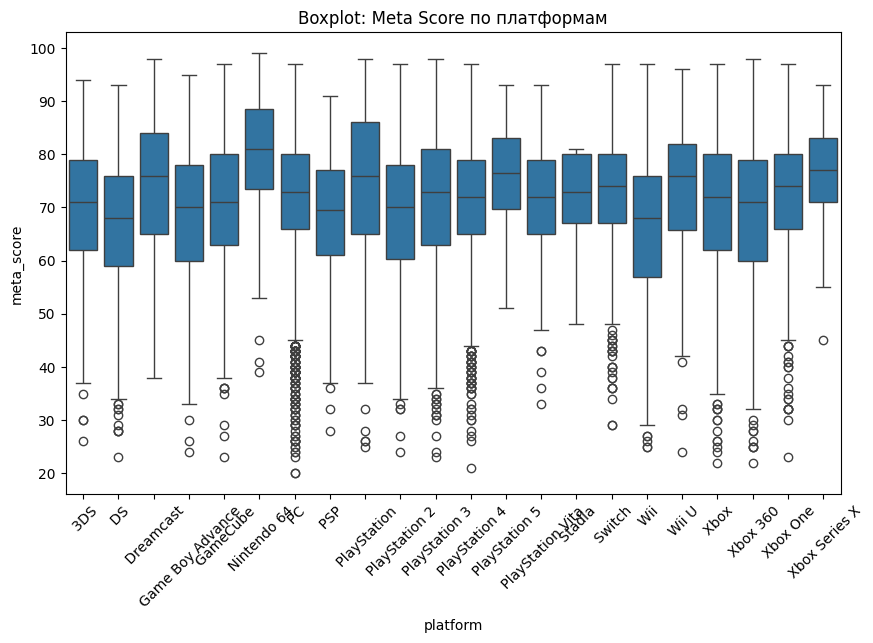
# **Проанализируем распределение всех категориальных данных**

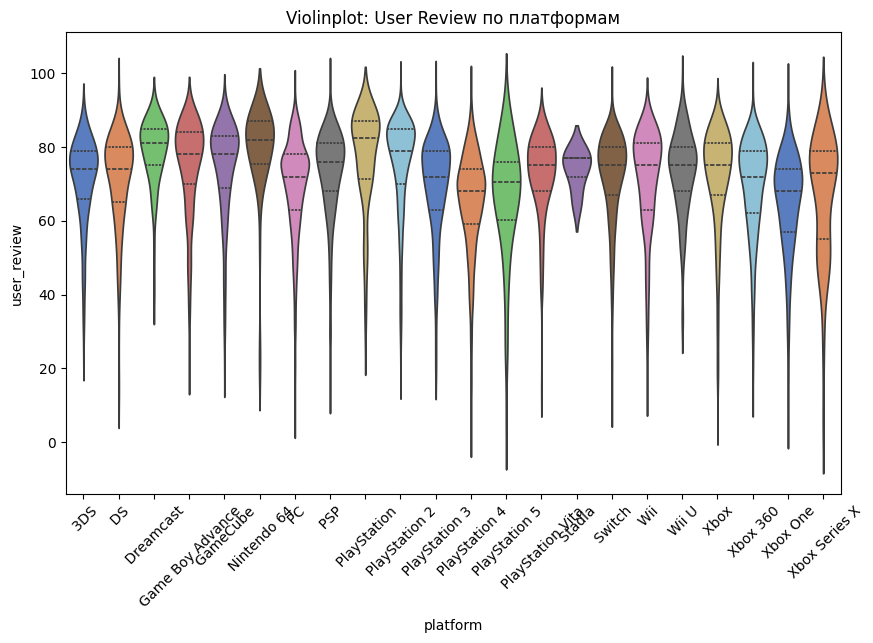
****

****

# **Построим более сложные графики**

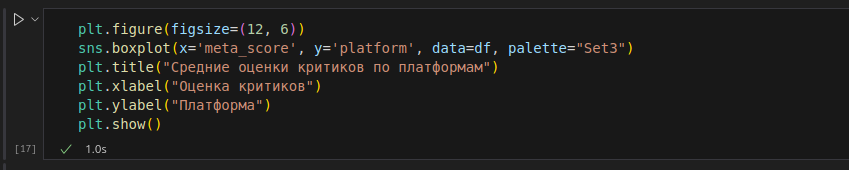
****

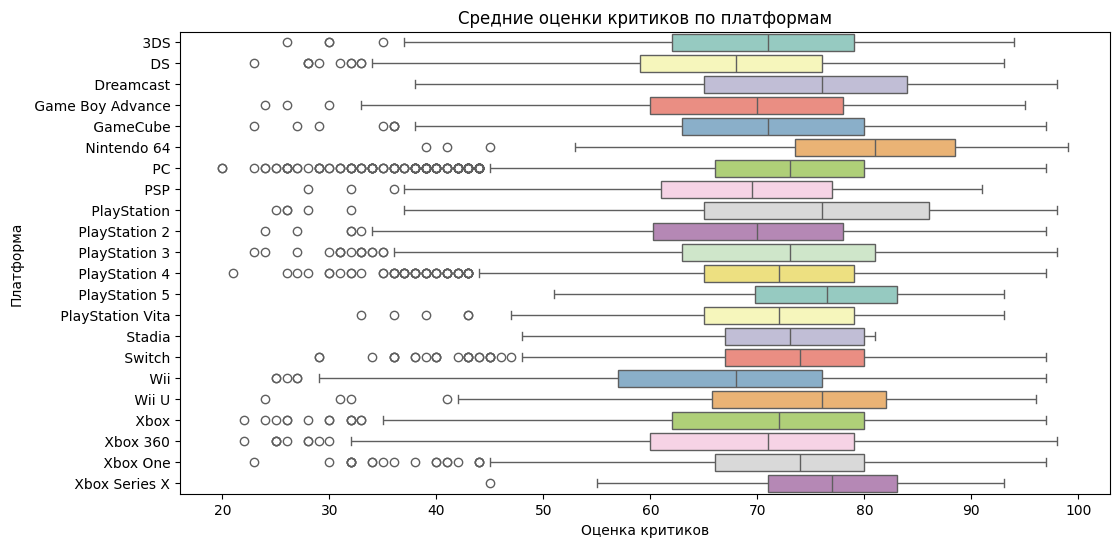
****

****

# **Выдвинем свои гипотезы по датасету и проверим их**

## **Гипотеза 1:** Игры на разных платформах имеют разные средние оценки.



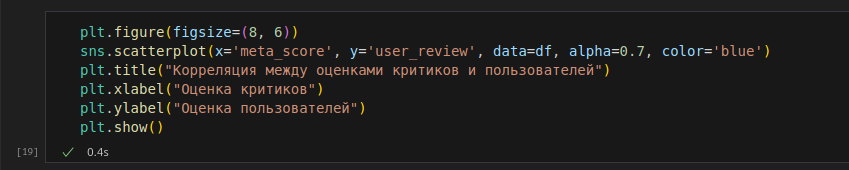


На основании проведенного ANOVA-теста (анализ дисперсии) с полученными значениями F-статистики (22.47) и P-значения (1.65e-85), можно сделать вывод, что средние оценки критиков различаются для разных платформ. P-значение значительно меньше порогового уровня 0.05, что позволяет отклонить нулевую гипотезу о равенстве средних оценок для всех платформ.

Визуализация, представленная с помощью коробчатой диаграммы (boxplot), также подтверждает это утверждение, показывая, что оценки критиков на разных платформах варьируются, с заметными различиями между ними.

Таким образом, гипотеза о различиях средних оценок по платформам подтверждается.

## **Гипотеза 2:** Высокая оценка критиков коррелирует с высокой пользовательской оценкой.





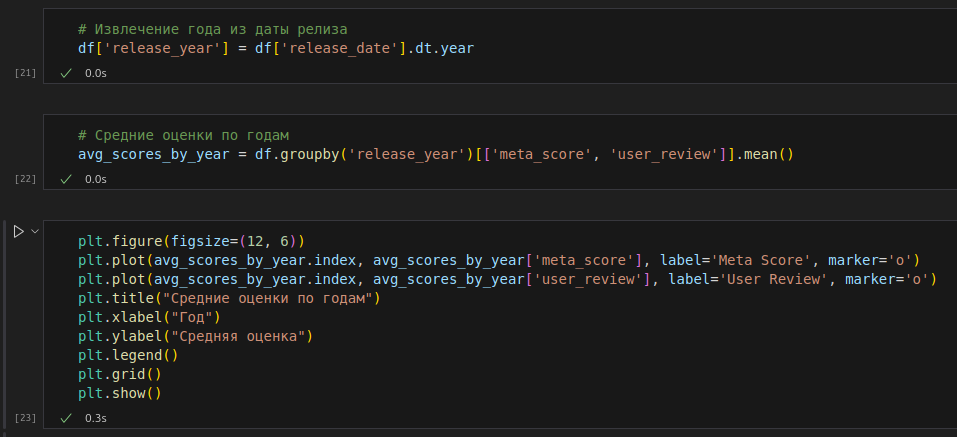
Результаты расчета корреляции Пирсона дали значения NaN как для коэффициента корреляции, так и для P-значения. Это может указывать на проблемы с данными, например:

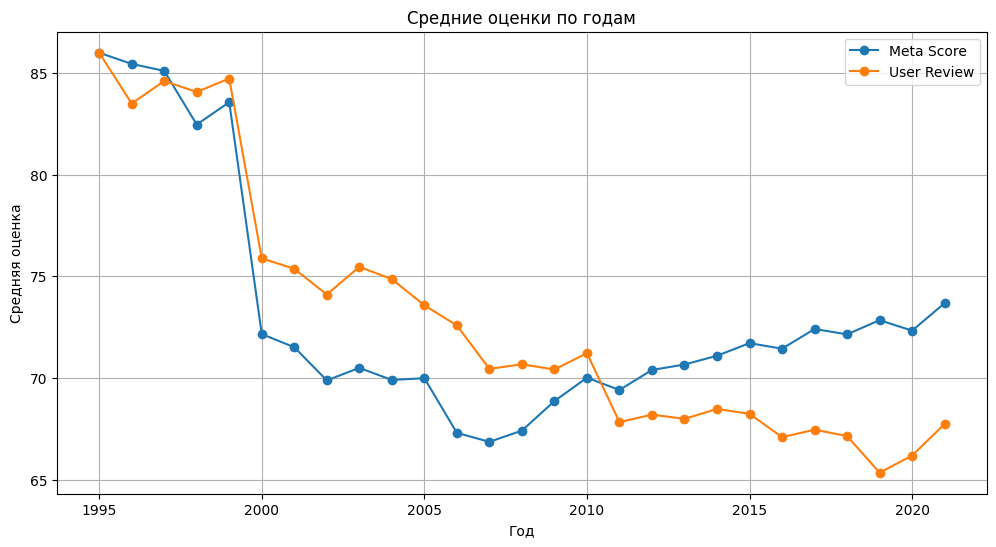
- Пропущенные значения (NaN): В колонках meta\_score или user\_review могут быть пропуски, что приводит к некорректным расчетам.

- Недостаточная вариативность данных: Если значения одной или обеих переменных содержат слишком много повторяющихся значений или вообще не изменяются, это также может вызвать проблемы при вычислении корреляции.

- Проблемы с типами данных: Убедитесь, что данные в колонках представлены в числовом формате, а не как строки.

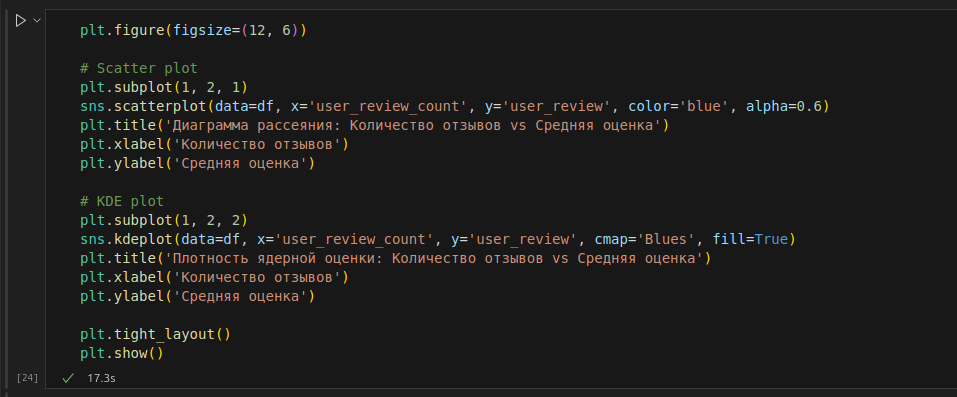
## **Гипотеза 3:** Год выпуска игры влияет на её популярность (по средним оценкам).

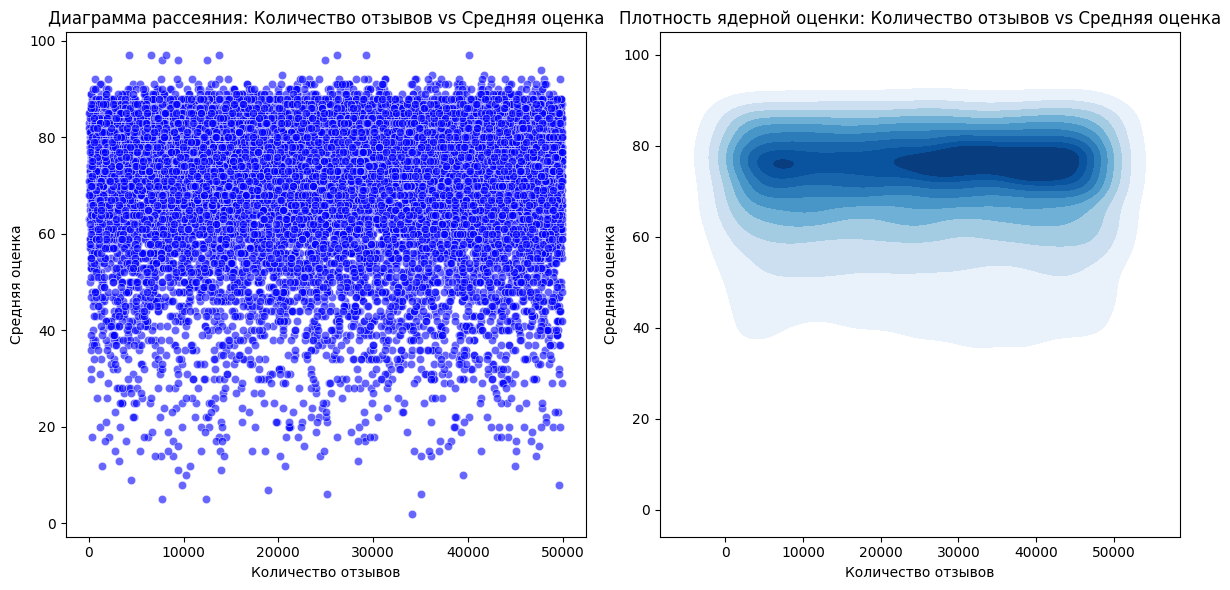




По графику мы видим, что в 2000 году был резкий спад оценок. Это связано с резким увиличением популярности и вследствии количеством выпуска игр.

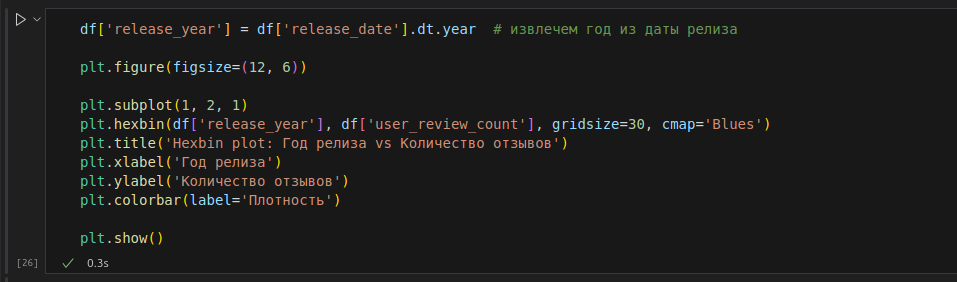
## **Гипотеза 4:** Чем больше оценок от пользователей - тем нижи они в среднем.

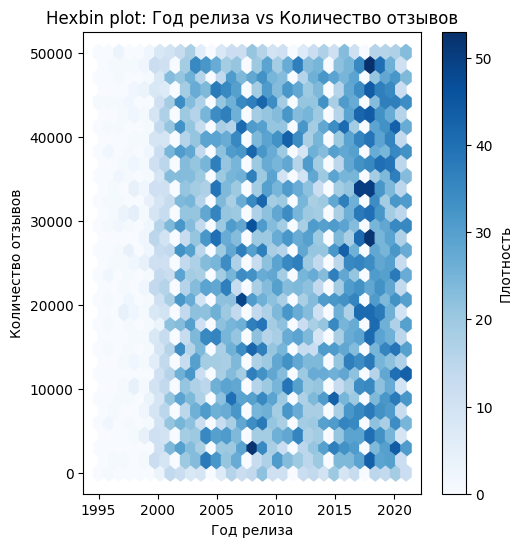




Данные и анализ не подтверждают гипотезу о том, что с увеличением количества отзывов оценки пользователей становятся ниже. Вместо этого наблюдается почти полное отсутствие взаимосвязи между этими переменными.

## **Гипотеза 5:** Дата релиза имеет корреляцию с количеством отзывов пользователей.





На основе проведенного анализа данных и вычисленного коэффициента корреляции (-0.000232) можно сделать вывод, что между годом релиза и количеством отзывов пользователей отсутствует значимая корреляция. Графическая визуализация (hexbin plot) также не выявила четкой зависимости между этими переменными. Таким образом, гипотеза о наличии корреляции между датой релиза и количеством отзывов пользователей не подтверждается.

# **ЗАКЛЮЧЕНИЕ**

В ходе научно-исследовательской работы был проведен анализ базы данных «Top Video Games 1995-2021 Metacritic», выявлены важные статистические данные, которые могут помочь при выборе рейса для полета.

Для выполнения данной работы было изучено новое программное обеспечение: «Pandas», «Seaborn» и «Plotly» на базе языка программирования Python. Также мы пользовались «Jupyter Python»: в нём мы писали код и анализировали диаграммы. В ходе работы я получил новые навыки в области анализа данных и познакомился с вышеперечисленными 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).