МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

федеральное государственное автономное образовательное учреждение высшего образования

«САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ   
АЭРОКОСМИЧЕСКОГО ПРИБОРОСТРОЕНИЯ»

КАФЕДРА №41

ЗАЩИЩЕНА С ОЦЕНКОЙ

РУКОВОДИТЕЛЬ

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| К.т.н., доц. |  |  |  | Е.Л. Турнецкая |
| должность, уч. степень, звание |  | подпись, дата |  | инициалы, фамилия |

|  |
| --- |
| ОТЧЕТ О ЛАБОРАТОРНОЙ РАБОТЕ №2 |
| Анализ связей между признаками двумерного набора данных |
| по курсу: Методология и технология проектирования информационных систем |
|  |

РАБОТУ ВЫПОЛНИЛ

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| СТУДЕНТ ГР. № | М320М |  |  |  | П. Е. Лукьянец |
|  |  |  | подпись, дата |  | инициалы, фамилия |

Санкт-Петербург 2023

*Цель работы:* изучение связи между признаками двумерного набора данных.

**Ход выполнения работы**

Для работы был выбран шестой датасет из списка под названием «6 games». В данном датасете представлена информация о видеоиграх: названии, платформе, годе выпуска, жанре, продажах в разных регионах и оценках. Чтобы начать работу, импортируем библиотеку, а затем считываем CSV файл. Выведем первые 20 строк с помощью метода head. Часть данных этого датасета представлена на рисунке 1.

Рисунок 1 – используемый датасет

Для работы с ним использовалась библиотека Pandas.

Работа была выполнена при помощи Visual Studio Code, a также Jupyter Notebook.

Ссылка на GitHub репозиторий с файлами: https://github.com/NinjaCaratist/MTPIS

Чтобы начать работу, импортируем библиотеку, а затем считываем CSV файл.

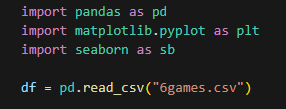


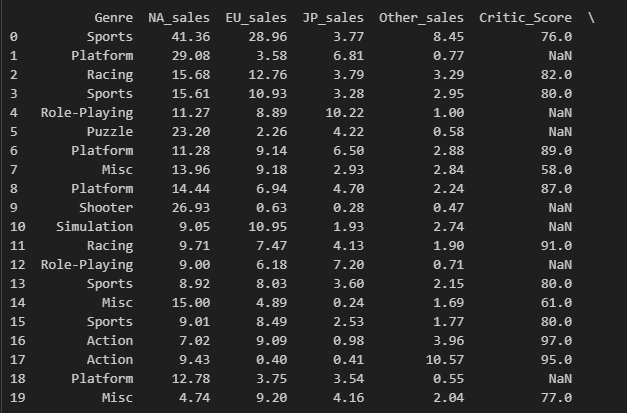
Рисунок 2 – скриншот кода

Выведем первые 20 строк с помощью метода head. На рисунке 3 показан код, а на рисунке 4 – результат его работы.



Рисунок 3 – скриншот кода





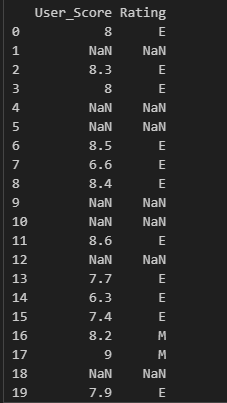


Рисунок 4 – результат вывода

Как уже было сказано, данная таблица содержит информацию о продаваемых автомобилях. Предметная область – автомобили и продажи. Опишем колонки подробнее:

Name – Название

Platform - Платформа

Year – год выпуска

Genre - Жанр

NA\_Sales, EU\_sales, JP\_Sales, Other\_SalesПродажи в разных регионах

Critic\_Store - Оценка

Теперь с помощью метода «.info» оценим данные. Этот метод возвращает название столбцов, типы данных, количество ненулевых объектов каждом столбце. Результат работы метода представлен на рисунке 5.

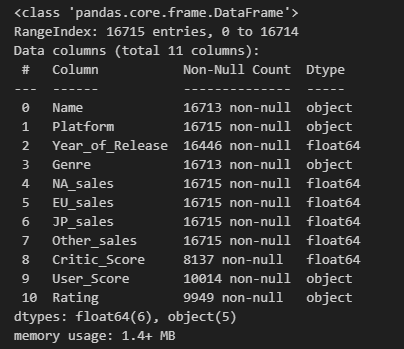


Рисунок 5 – результат вывода

Теперь выведем на экран названия столбцов с помощью df.columns. Названия всех колонк приемлимы.



Рисунок 6 – скриншот кода

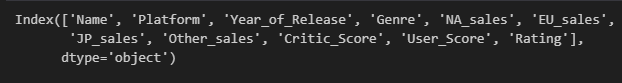


Рисунок 7 – результат вывода

Найдём пропуски и устраним их. При помощи метода «isna» найдём все пропуски в таблице, а также при помощи sum выведем количество пропусков в каждом столбце. Как мы можем увидеть присутствуют пропуски в столбцах названия, жанра, года выпуска, оценок и рейтинга. Так как важным столбцом является лишь название, строки без информации в нём не имеют смысла, поэтому их стоит удалить при помощи метода «dropna». Однако можно не удалять строки без данных в дргугих столбцах. Поэтому при помощи метода fillna заполним пустые значения оценок нулями, жанра и рейтинга строкой 'none', а года выпуска 1950 годом, в котором ещё не выпускались игры. Проверяем пропуски еще раз и их больше нет. Код представлен на рисунке 8. Проверяем пропуски еще раз и их больше нет.

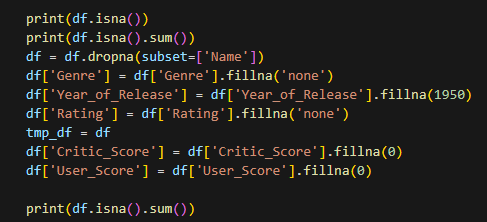


Рисунок 8 – скриншот кода

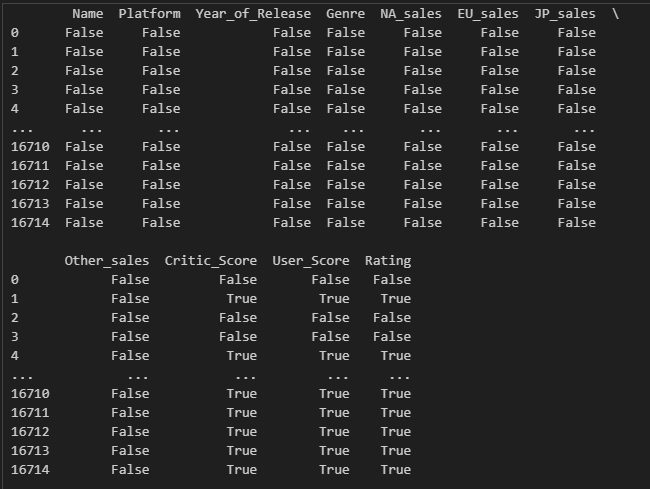
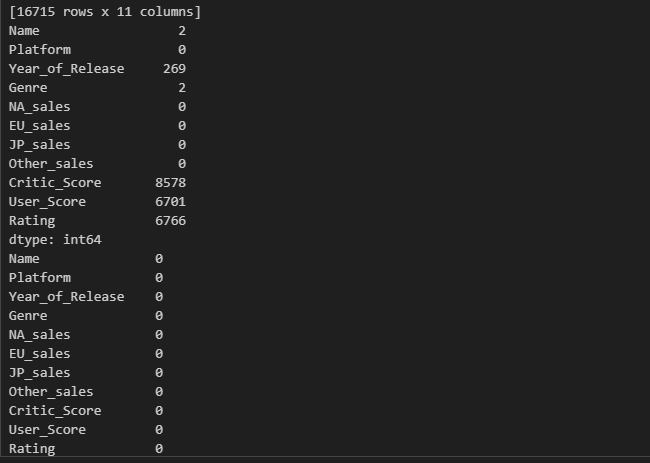
 

Рисунок 9 – результат вывода

Проверим данные на наличие дубликатов – рисунок 10. Полностью повторяющихся строк нет, результат представлен на рисунке 11. Уникальным значением в данном наборе является название, а точнее название, платформа и год выпуска, только в таком сочетании строка уникальна. Получается составной перивчный ключ. Уникальных значений по имени 11559.

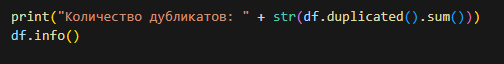
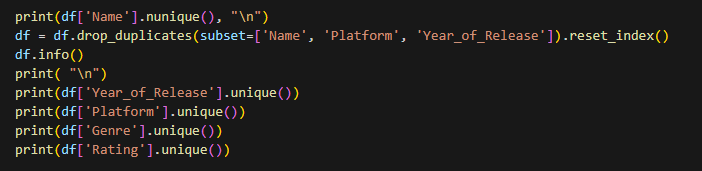


Рисунок 10 – скриншот кода



Рисунок 11 – результат вывода

При помощи метода drop\_duplicates удаляем дубликаты по столбцам Названия, платформы и года – рисунок 12. После переиндексации и вывода информации видим, что на одну запись стало меньше - рисунок 13. В записях остальных колонок ошибок нет.

Рисунок 12 – скриншот кода

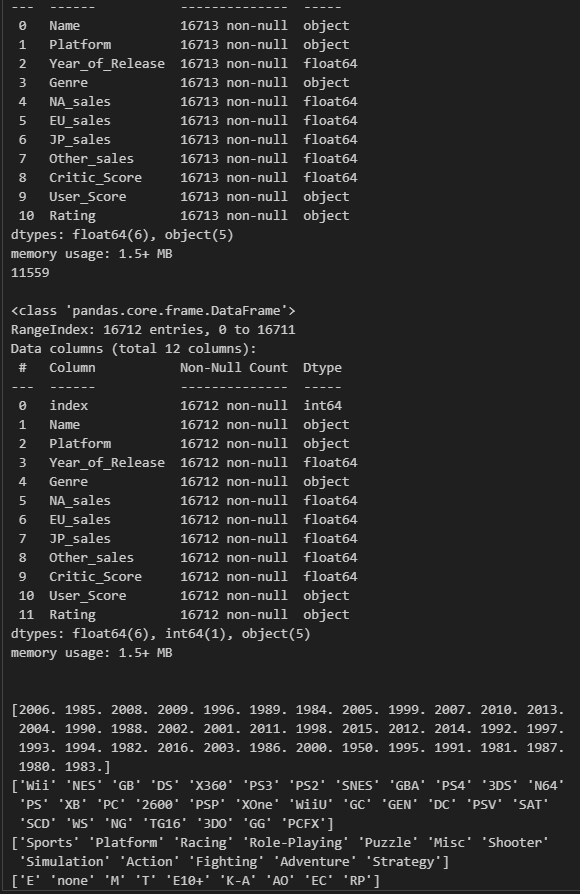


Рисунок 13 – результат вывода

Проверим все ли типы данных соответствуют действительности. Все столбцы, кроме года выпуска и оценки пользователей соответствуют своему типу. Поэтому при помощи метода «to\_datetime» изменяем тип на временной. А при помощи метода "to numeric" приводим оценки к формату с плавающей точкой, при этом прописывая условия о не превращающихся значениях в NaN (см. рис. 14). Затем эти значения приравниваем к нулю и выводим информацию (см. рис. 15).

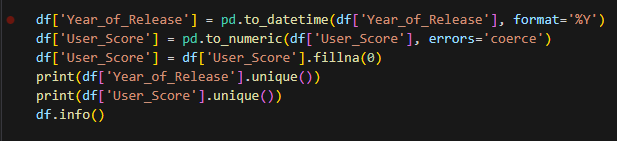


Рисунок 14 – скриншот кода

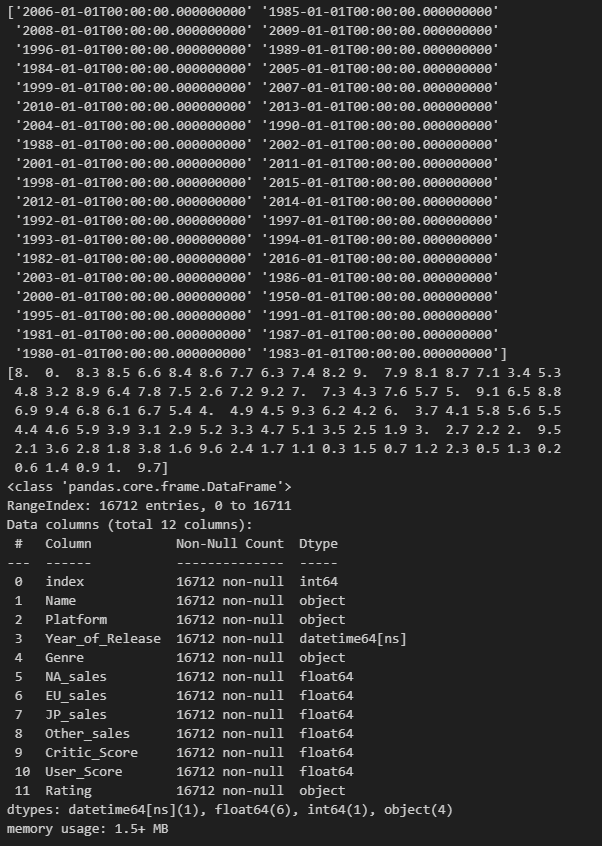


Рисунок 15 – результат вывода

Создадим сводную таблицу при помощи метода «data\_pivot», код представлен на рисунке 16. Индексацию возьмём по платформам, а колонки по жанрам игр. Подсчёт будет по сумме продаж в Северной Америке. Таким образом, получится таблица, показывающая, на какой жанр и с какой консолью приходится больше всего сумма продаж. Это шутеры на xbox360, что ожидаемо. Это можно увидеть на рисунке 17.

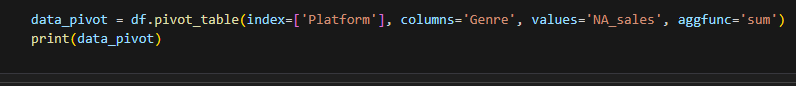


Рисунок 16 – скриншот кода

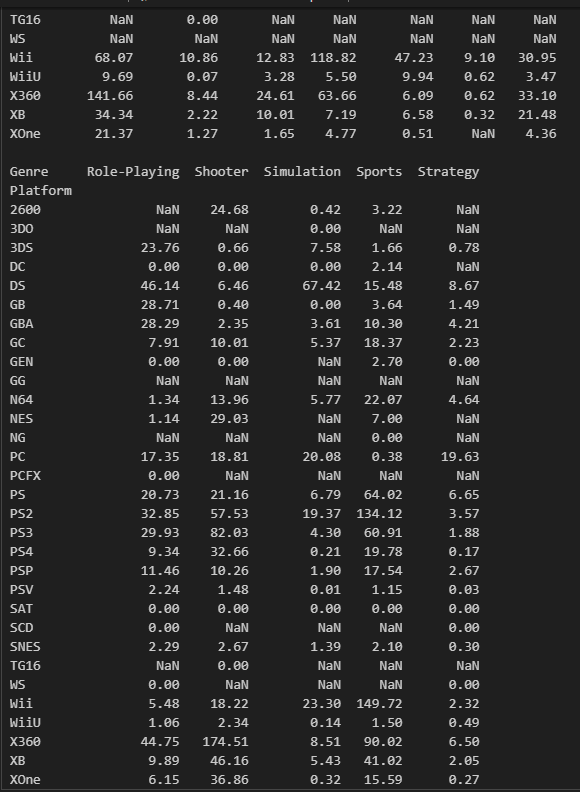


Рисунок 17 – результат вывода

Теперь при помощи метода describe выведем описание статистики по всем атрибутам. После этого построим несколько графиков: по количеству значений различных атрибутов, по количеству продаж в Америке в различные года, а также по зависимости переменных друг от друга.

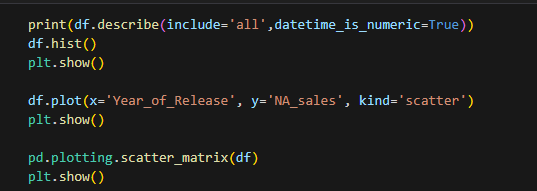


Рисунок 18 – скриншот кода

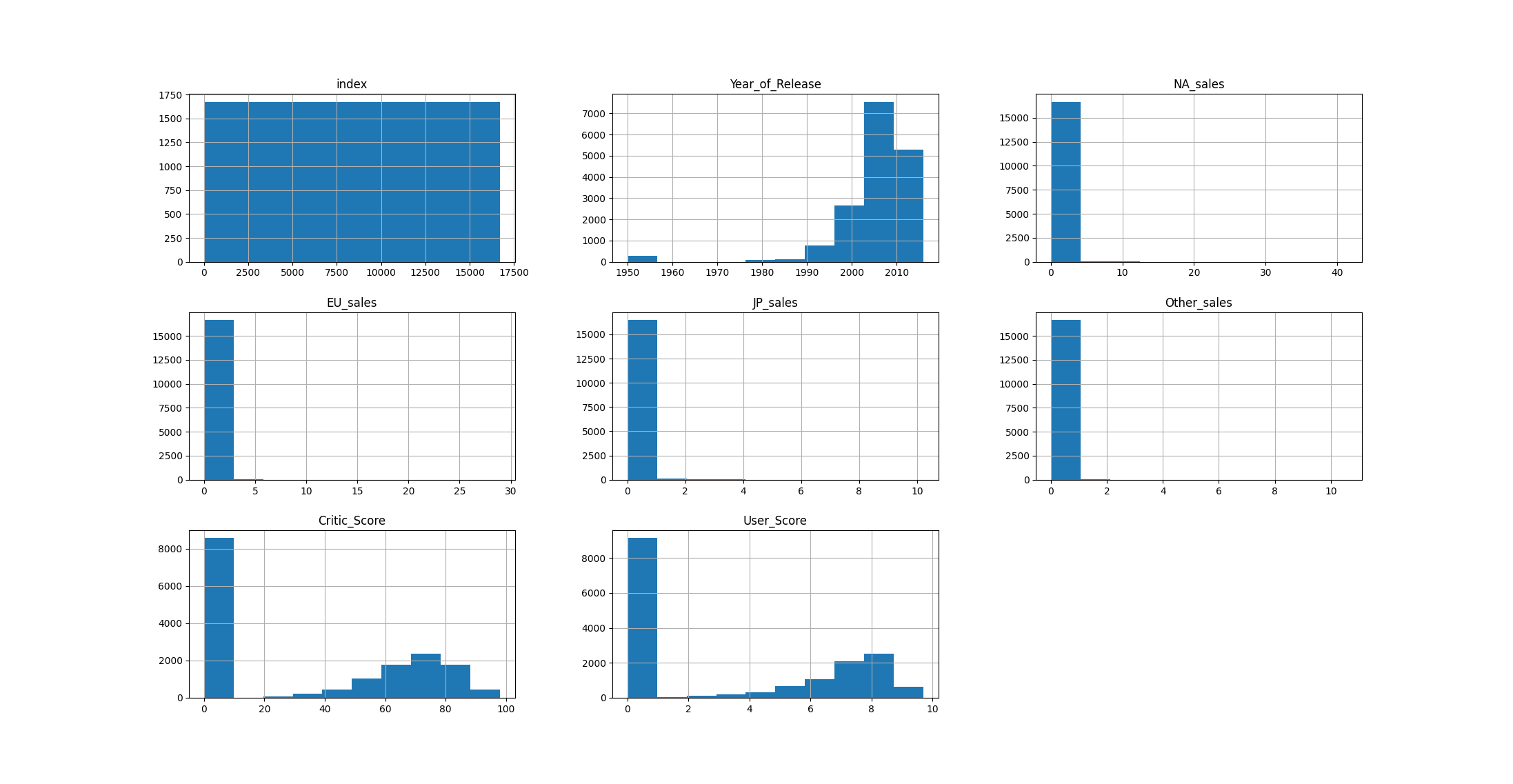


Рисунок 19 – построенные гистограммы

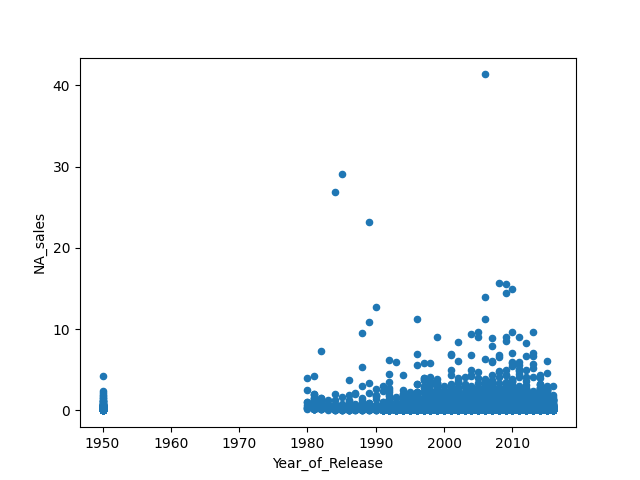


Рисунок 20 – построенный график

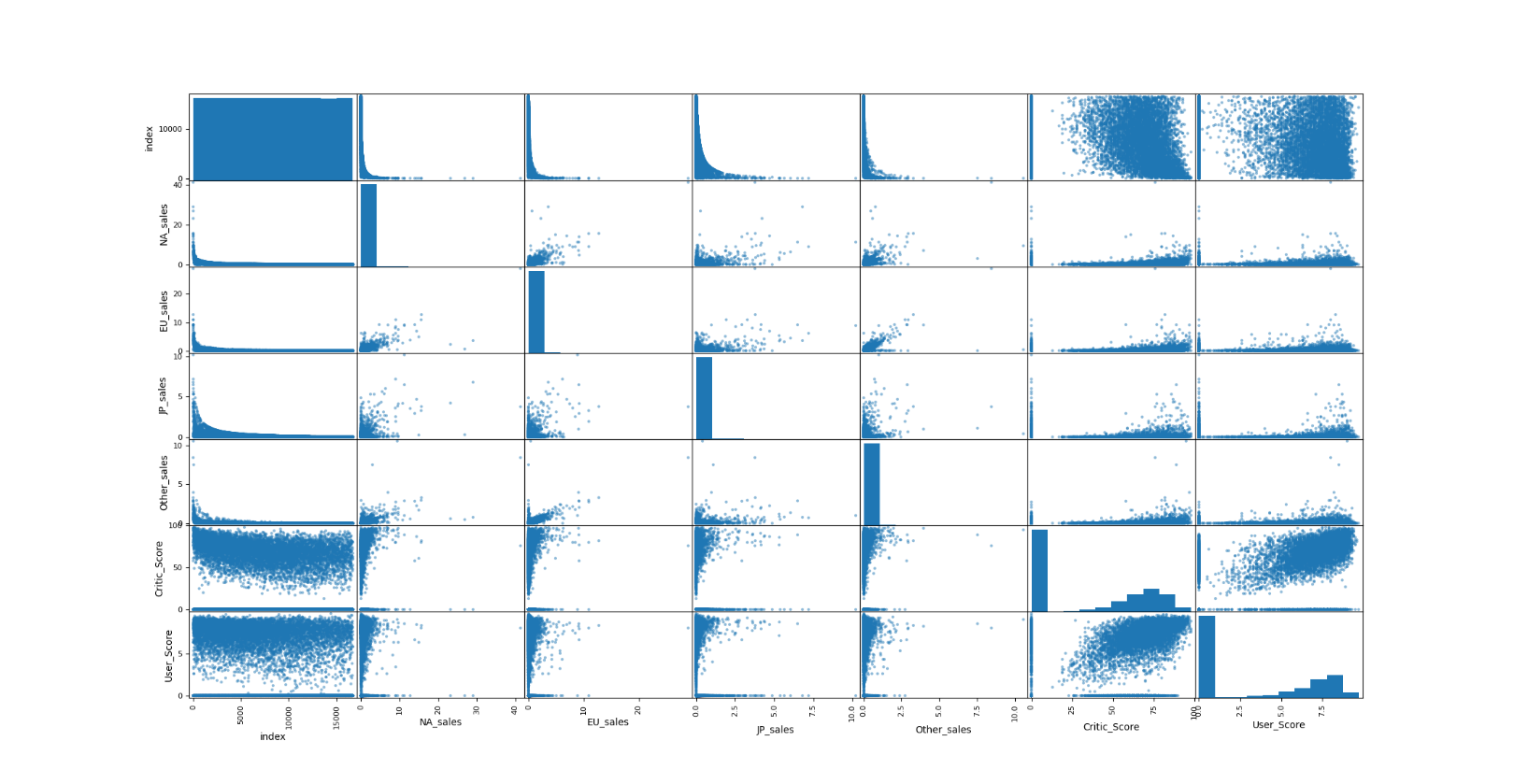


Рисунок 21 – построенная матрица зависимомтей

Теперь исследуем взаимосвязь между переменными с помощью оценки

коэффициента корреляции и ковариации, а также построим тепловую карту по этим значениям. В созданном заранее временном дата фрейме, удалим строки с отсутсвующими значениями оценок для лучшей оценки корелляции и ковариации. Интерпретируем результаты. По результатам ковариации можно сказать, что большинство числовых параметров изменяются в одном направлении, кроме продаж в Японии от оценок критиков и пользователей. По результатам корреляции, что самая близкая к линейной является зависимость оценок пользователей от оценок критиков. Также немалый коэффициент зависимости между собой имеют продажи в СА, Европе и остальном мире. У Японцев какой-то свой отдельный мир.

Один столбец с целевым признаком выделить сложно, скорее можно выделить группу атрибутов, связанных с продажами, так как главным для любого продукта является его окупаемость. Продажи, как показывает корреляция, не сильно зависит от оценок. Поэтому следует предположить, что влияют другие показатели, такие как: маркетинг, популярность серии или выпускающей корпорации, выпуск на разных платформах и т.д.

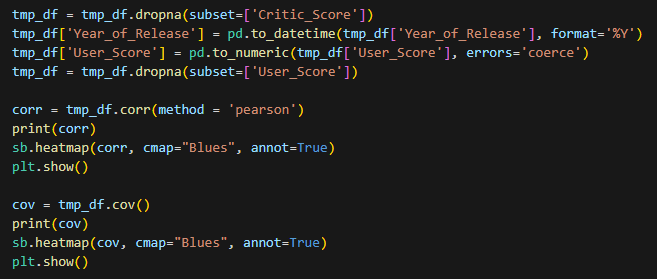


Рисунок 22 – скриншот кода

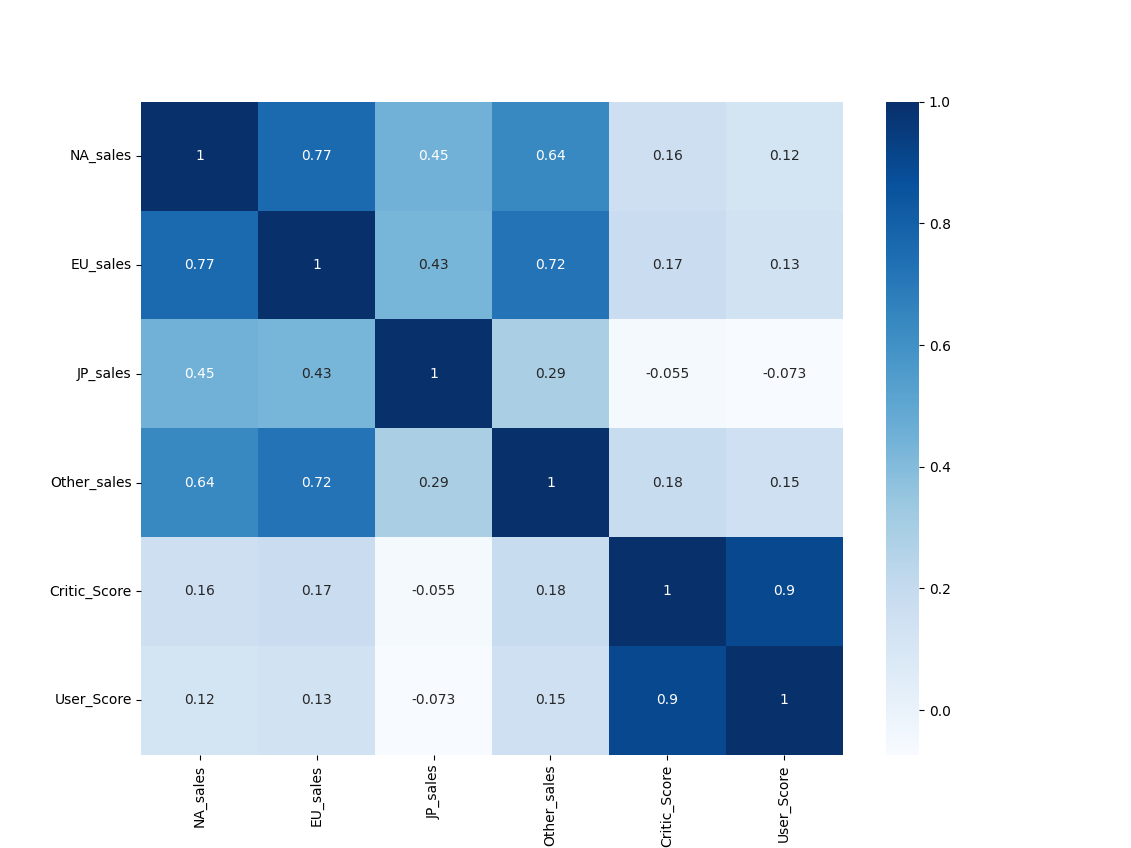


Рисунок 23 – построенная тепловая карта корреляции

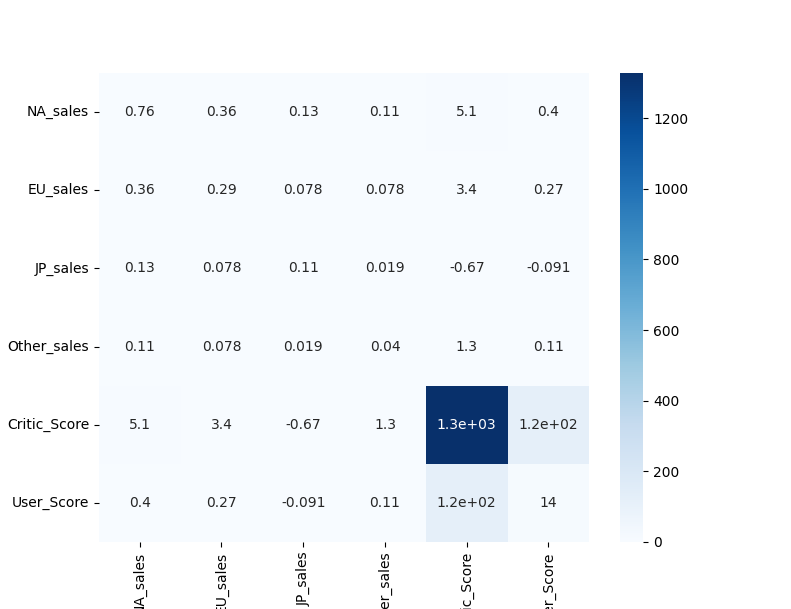


Рисунок 24 – построенная тепловая карта ковариации

*Вывод:* Вывод: Таким образом, в ходе выполнения лабораторной работы был выбран и описан выбранный датасет про продажи и оценки консольных игр, изучен интерфейс и возможности Jupyter Notebook, изучены базовые функции библиотеки Pandas и разработана программа, которая считывает данные, выводит о них информацию, удаляет дубликаты, пропуски, изменяет тип данных и создаёт сводную таблицу и графики, а также высчитывает коэффициенты корелляции и ковариации и строет тепловые карты по ним.