МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное автономное образовательное учреждение высшего образования
«Омский государственный технический университет»

Факультет информационных технологий и компьютерных систем

Кафедра «Прикладная математика и фундаментальная информатика»

Домашнее задание

по дисциплине Практикум по программированию

Студента(ки) Передериной Софьи Владимировны
фамилия, имя, отчество полностью

Курс 2 Группа ФИТ-242

Направление 02.03.02. Фундаментальная информатика и информационные технологии
код, наименование

Руководитель старший преподаватель
должность, ученая степень, звание

Саматов А. П.
фамилия, инициалы, дата, подпись

Выполнил _____
дата, подпись студента(ки)

Итоговый рейтинг	

Омск 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Анализ данных - это процесс сбора, обработки, анализа и интерпретации информации, в результате которого разрозненные данные превращаются в законченную информационную продукцию – аналитический документ [1]. Рост объема данных в современном мире происходит с невероятной скоростью. Ежедневно генерируются петабайты информации: от пользовательских действий в интернете до данных IoT-устройств. По прогнозам IDC, к 2025 году объем данных достигнет 175 зеттабайт.

Цифровизация бизнеса становится необходимостью для выживания на конкурентном рынке. Все больше компаний переходят на data-driven подход в принятии решений. Согласно исследованию NewVantage Partners, 91% компаний из списка Fortune 1000 увеличивают инвестиции в аналитику данных.

Развитие технологий открывает новые возможности для анализа данных. Появление продвинутых BI-платформ и систем машинного обучения требует квалифицированных специалистов для их эффективного использования. Интеграция AI и ML в процессы анализа данных становится новым стандартом индустрии [2].

В ходе выполнения лабораторной работы были изучены библиотеки Pandas, NumPy, Scipy, Matplotlib, Seaborn языка Python.

Pandas помогает эффективно преобразовать и разделять структурированные данные. NumPy позволяет выполнить сложные научные расчёты с математическими объектами. Scipy построена на массивах и функциях NumPy и содержит инструменты для сложных математических операций. Seaborn и Matplotlib позволяют строить различные визуализации данных [3].

1 Поиск и загрузка данных

На платформе Kaggle был найден датасет PlayStation Games Info 2/15/2025. Этот набор данных содержит подробную информацию об играх для PlayStation, объединяющую официальные данные PlayStation Store с отзывами критиков и пользователей Metacritic. Он содержит в себе 10368 строк и 13 столбцов. Скриншот файла README.md, содержащего информацию о столбцах выбранного датасета, представлен на рисунке 1.

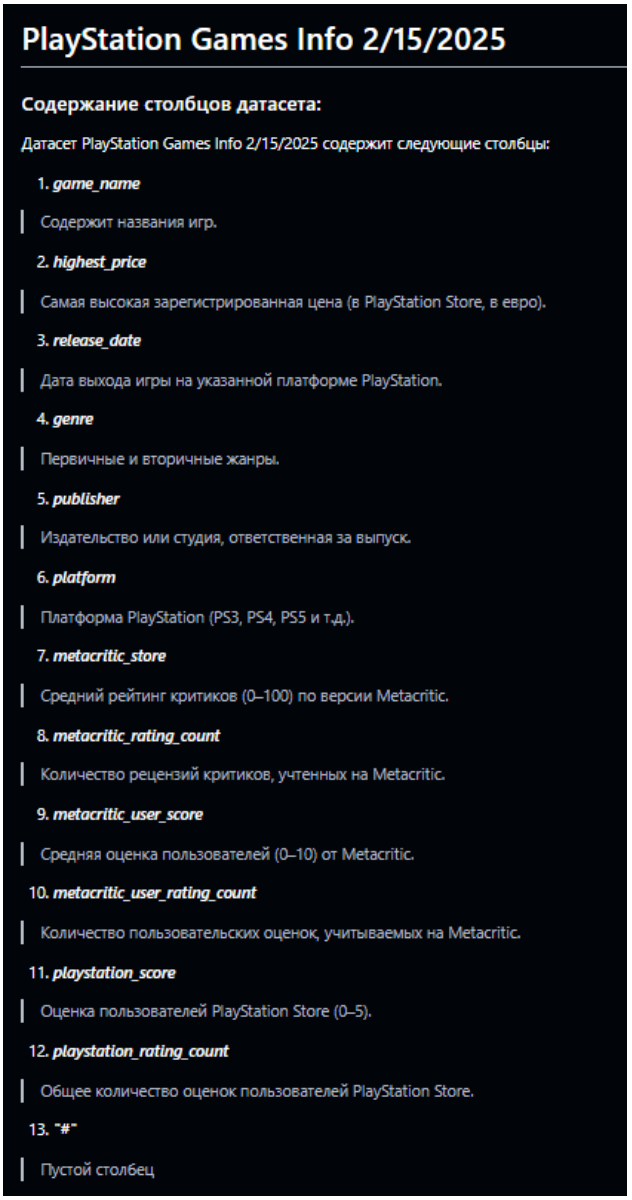

Рисунок 1 – Информация о столбцах датасета

Датасет в виде файла game_details.csv был загружен в ноутбук с помощью функции pd.read_csv("game_details.csv").

Скриншоты с содержанием датасета представлены на рисунках 2 – 5.

	game_name	highest_price	release_date	genre	publisher	platform	metacritic_score	metacritic_rating_count
0	Grand Theft Auto IV	€24.99	Feb 15, 2012	Action / Shooter / Racing	Rockstar	PS3	98.0	86.0
1	Red Dead Redemption 2	€59.99	Oct 26, 2018	Action / Adventure / Unique	Rockstar Games	PS4	97.0	99.0
2	Red Dead Online	€69.99	Oct 29, 2018	Action / Adventure	Rockstar Games	PS4	97.0	99.0
3	Grand Theft Auto 3	€9.99	Oct 4, 2012	--	Rockstar Games	PS3	97.0	56.0
4	Grand Theft Auto V	€69.99	Sep 17, 2013	Action / Adventure	Rockstar Games	PS3	97.0	66.0
...	...	...	...	...	...	...	...	...
12018	Handball 16	NaN	Nov 27, 2015	Sports	Bigben Interactive	PS Vita	NaN	NaN
12019	Wired Arcade Bundle	NaN	Nov 20, 2015	Arcade / Simulation	WIRED PRODUCTIONS LIMITED	PS4 / PS3 / PS Vita	NaN	NaN
12020	Arcade Archives MAGMAX	NaN	Dec 2, 2015	Shooter / Arcade	HAMSTER CORPORATION	PS4	NaN	NaN
12021	AvengeXX	NaN	Dec 4, 2015	Action	SSPROJECTX	PS3 / PSP	NaN	NaN
12022	MotoGP™15 Compact	NaN	Dec 10, 2015	Racing	MILESTONE SRL	PS3	NaN	NaN

Рисунок 2 – Содержимое датасета

metacritic_user_score	metacritic_user_rating_count	playstation_score	playstation_rating_count	Unnamed: 12
8.3	5541.0	4.32	48904.0	NaN
8.9	31932.0	4.74	379257.0	NaN
8.9	31932.0	4.74	379346.0	NaN
8.0	2079.0	4.59	1437.0	NaN
8.5	14322.0	4.65	40895.0	NaN
...	...	...	...	...
NaN	NaN	3.04	74.0	NaN
NaN	NaN	3.29	28.0	NaN
NaN	NaN	4.17	23.0	NaN
NaN	NaN	4	2.0	NaN
NaN	NaN	3.7	47.0	NaN

Рисунок 3 – Содержимое датасета

```
1  df.info()
   [5]
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12023 entries, 0 to 12022
Data columns (total 13 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   game_name               12023 non-null  object
 1   highest_price           3463 non-null   object
 2   release_date            11959 non-null  object
 3   genre                   11959 non-null  object
 4   publisher               11742 non-null  object
 5   platform                12014 non-null  object
 6   metacritic_score        950 non-null    float64
 7   metacritic_rating_count 950 non-null    float64
```

Рисунок 4 – Вывод df.info()

```
1  df.describe()
   [6]
```

8 rows × 8 rows × 6 cols

	metacritic_score	metacritic_rating_count	metacritic_user_score	metacritic_user_rating_count	playstation_rating_count
count	950.000000	950.000000	951.000000	951.000000	7.158000e+03
mean	76.305263	32.181053	6.508517	980.539432	3.821006e+03
std	9.408842	28.258793	2.474323	6026.066475	3.637133e+04
min	64.000000	4.000000	0.000000	0.000000	1.000000e+00
25%	68.000000	10.000000	6.200000	14.000000	7.000000e+00
50%	74.000000	22.000000	7.300000	64.000000	6.700000e+01
75%	84.000000	47.000000	8.000000	313.000000	6.327500e+02
max	98.000000	145.000000	9.500000	165959.000000	1.819326e+06

Рисунок 5 – Вывод df.decribe()

2 Разведывательный анализ данных

В ходе выполнения работы были построены следующие визуализации:

2.1 Гистограмма

На гистограмме (рисунок 6) представлено распределение игр по их максимальной цене на 10 интервалах. Большинство высоких столбцов сосредоточено в левой половине гистограммы. Это позволяет понять, что большая часть игр, для которых в датасете указана цена, продаются по стоимости, не превышающей 40 евро. С ростом цены количество соответствующих игр уменьшается.

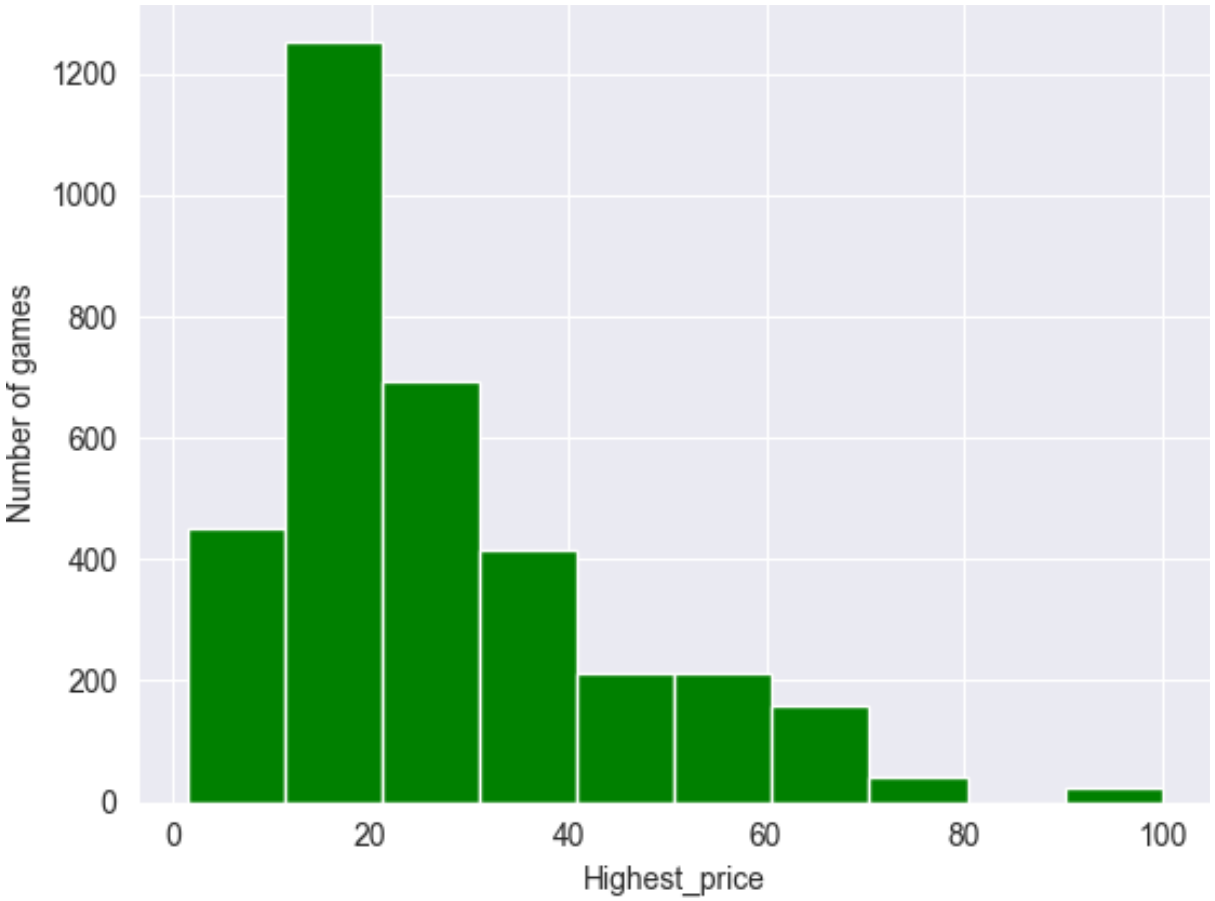

Рисунок 6 – Гистограмма

2.2 "Ящик с усами"

"Ящик с усами" для переменной highest_price представлен на рисунке 5. Он показывает распределение значений цены [4]. Первый квартиль (левая сторона ящика) лежит близко к началу координат. Третий квартиль (правая

сторона ящика) лежит около 40. Медиана (линия внутри ящика) для данного распределения лежит около значения 20. Значит, большая часть из представленных игр имеет низкую цену. Сильный дисбаланс между длинами "усов" говорит о том, что большие цены варьируются в большей степени, чем низкие. Несколько выбросов лежат по правую сторону от ящика, что говорит о небольшом количестве статистически незначимых игр из высокого ценового сегмента.

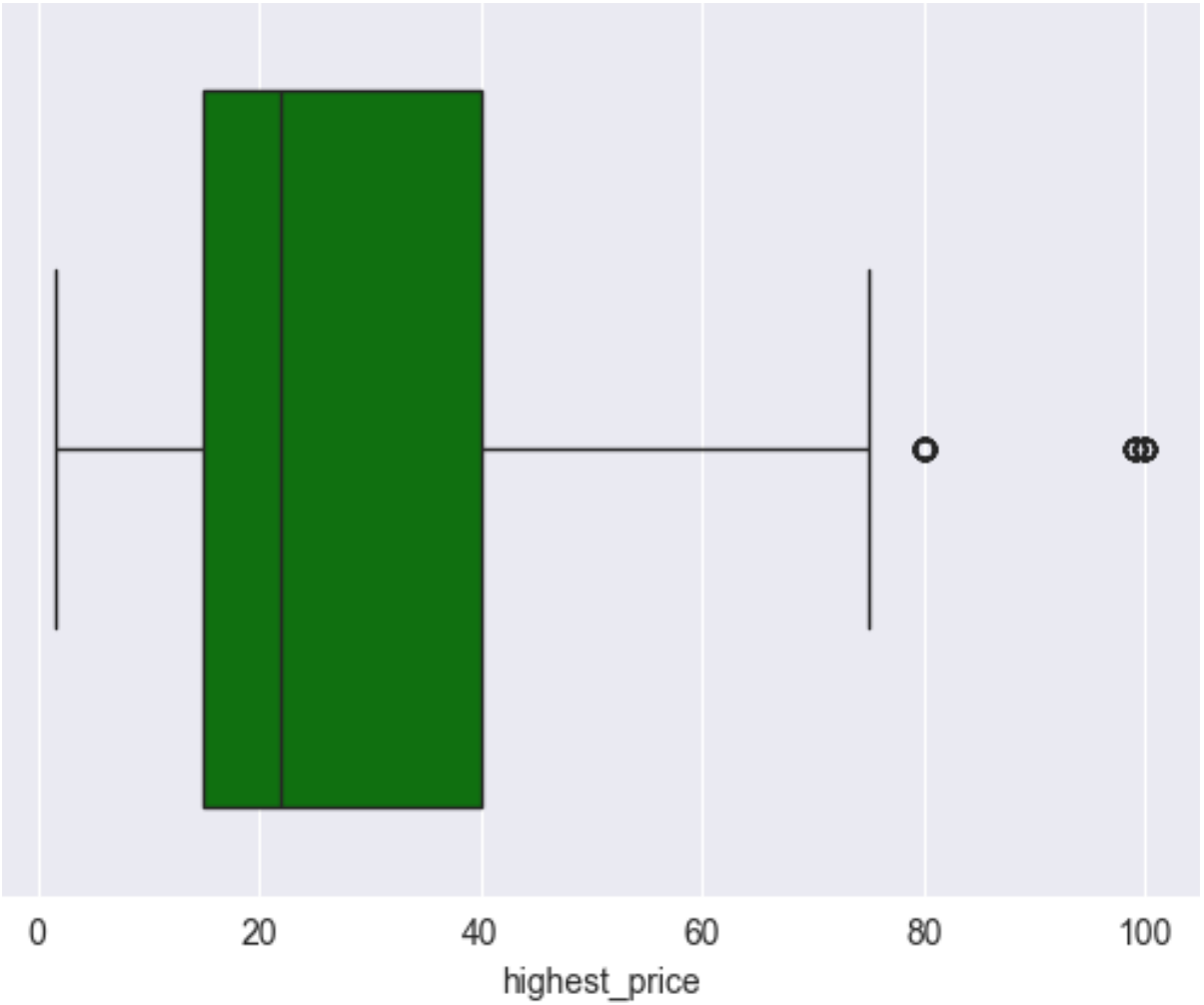

Рисунок 7 – "Ящик с усами"

2.3 Круговая диаграмма

На рисунке 8 представлена круговая диаграмма для платформ. По убыванию занимаемой площади диаграммы платформы расположены в следущем порядке: PS4, PS5, PS3, PS Vita, PSP. Это говорит о том, что большая

часть представленных игр была доступна на PS4 и PS5.

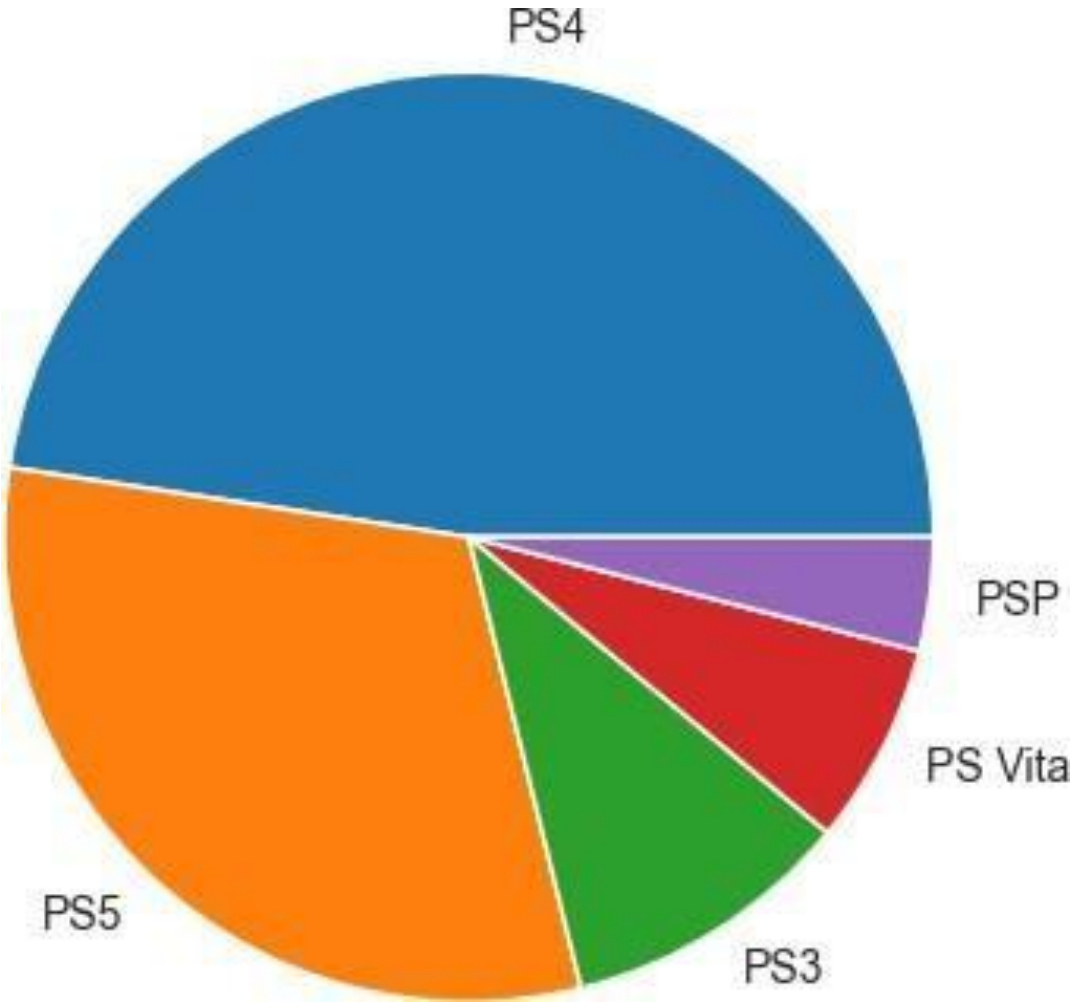

Рисунок 8 – Круговая диаграмма

2.4 Тепловая карта

Полученная тепловая карта представлена на рисунке 9. Она отображает значения взаимной корреляции для числовых характеристик представленных игр (цена, рейтинг критиков, количество рецензий критиков, средняя оценка пользователей, количество оценок пользователей на Metacritic и PS Store).

Наибольшая корреляция цены связана с количеством оценок критиков на Metacritic (0,39), при этом цена слабо коррелирует с количеством пользовательских оценок (0.095). Значение цены коррелирует с количеством

оценок в PS Store (0.24). Цена практически не коррелирует с оценкой в PS Store (-0.0049). Таким образом, более дорогие игры имеют большую популярность среди критиков.

Величины оценок критиков коррелируют с количеством их оценок (0.5), а также с пользовательским рейтингом (0.45) и оценкой в PS Store (0.46). В меньшей мере оценки критиков коррелируют с количеством оценок на Metacritic и в PS Store (0.23 и 0.21 соответственно). Так, оценки критиков согласуются с оценками пользователей.

Количество оценок критиков коррелирует с пользовательским рейтингом (0.38) и количеством оценок пользователей на Metacritic (0.32). В меньшей мере они коррелируют с количеством оценок в PS Store (0.21). Таким образом, оценки критиков хорошо коррелируют с пользовательскими.

Величина пользовательских оценок на Metacritic слабо коррелирует с их количеством на Metacritic (0.079) и PS Store (0.097). При этом оценки пользователей на Metacritic хорошо коррелируют с оценками пользователей в PS Store (0.33).

Величина оценки в PS Store почти не коррелирует с количеством оценок на этой платформе (0.025).

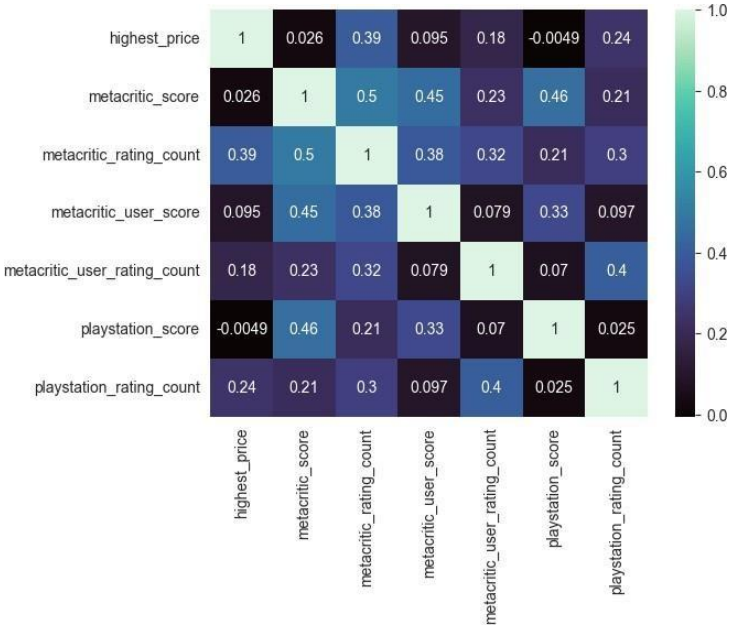

Рисунок 9 – Тепловая карта

2.5 Диаграмма countplot

На рисунке 10 представлена группированная столбчатая диаграмма, показывающая распределение игр по платформам и жанрам. С помощью высоты столбцов диаграммы можно сделать вывод о количестве игр, доступных на платформе. В порядке убывания это PS4, PS5, PS3, PS Vita, PSP. Самый популярный жанр среди всех платформ - Action. Самое большое многообразие жанров на платформе PS4, наименьшее количество жанров доступно на PSP.

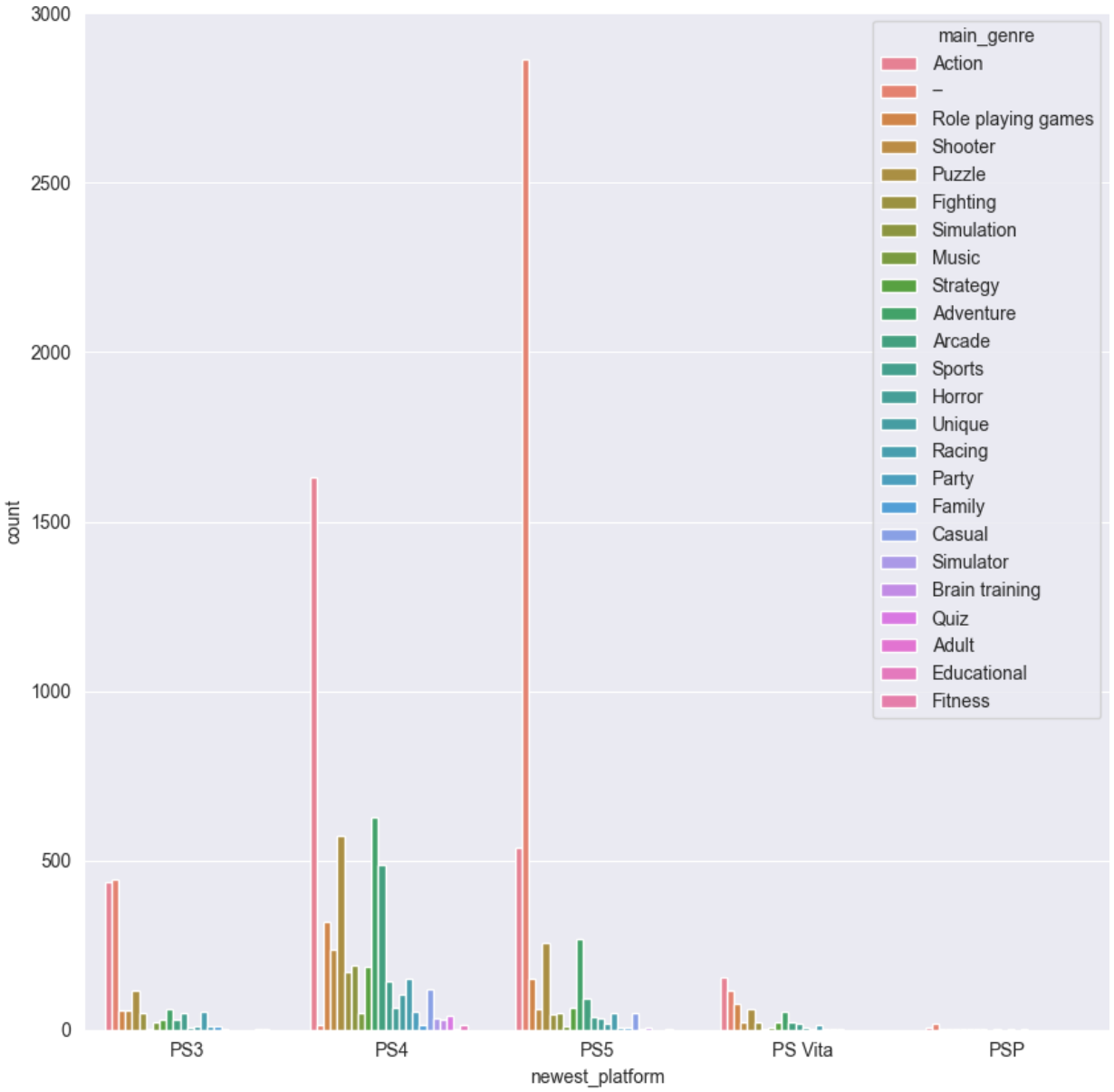

Рисунок 10 – Диаграмма countplot

3 Предварительная обработка данных

В рамках предварительной обработки данных было выполнено несколько преобразований. Были удалены дубликаты строк и пустой столбец. Столбец с именем game_name был переименован в game. Значения столбца highest_price и playstation_score были конвертированы из типа string в тип float32 с помощью функции to_numeric (рисунок 11).

```
df.drop_duplicates(inplace=True)      df
df.drop(columns="Unnamed: 12", inplace=True)
df.rename(columns={"game_name":"game"}, inplace=True)

df["highest_price"] = (df["highest_price"]
                       .str.replace("€", "")
                       .str.strip())
df["highest_price"] = pd.to_numeric(df["highest_price"], downcast="float")

df["playstation_score"] = df["playstation_score"].replace("--", np.nan)
df["playstation_score"] = pd.to_numeric(df["playstation_score"], downcast="float")
df
```

Рисунок 11 – Преобразование типов

С помощью метода isna была получена информация о количестве пропусков в столбцах (рисунок 12).

```
missing_data = pd.DataFrame({"missing_data": df.isna().sum()})
missing_data
✓ [106] < 10 ms
```

1-10 12 rows × 1 cols

	missing_data
game	0
highest_price	8553
release_date	62
genre	62
publisher	279
platform	7
metacritic_score	11066
metacritic_rating_count	11066
metacritic_user_score	11065
metacritic_user_rating_count	11065

Рисунок 12 – Количество пропусков

Пустые значения числовых столбцов были заполнены средними значениями по столбцу, остальные столбцы были заполнены модой по столбцу с помощью методов fillna, mean и mode (рисунок 13).

```
for float_column in ["highest_price", "metacritic_score",
                     "metacritic_rating_count", "metacritic_user_score", "metacritic_user_rating_count",
                     "playstation_rating_count"]:
    df[float_column] = df[float_column].fillna(df[float_column].mean())

for object_column in ["release_date", "genre", "publisher", "platform"]:
    df[object_column] = df[object_column].fillna(df[object_column].mode()[0])

df
```

Рисунок 13 – Заполнение пропусков

Результат заполнения пропущенных данных представлен на рисунке 14.

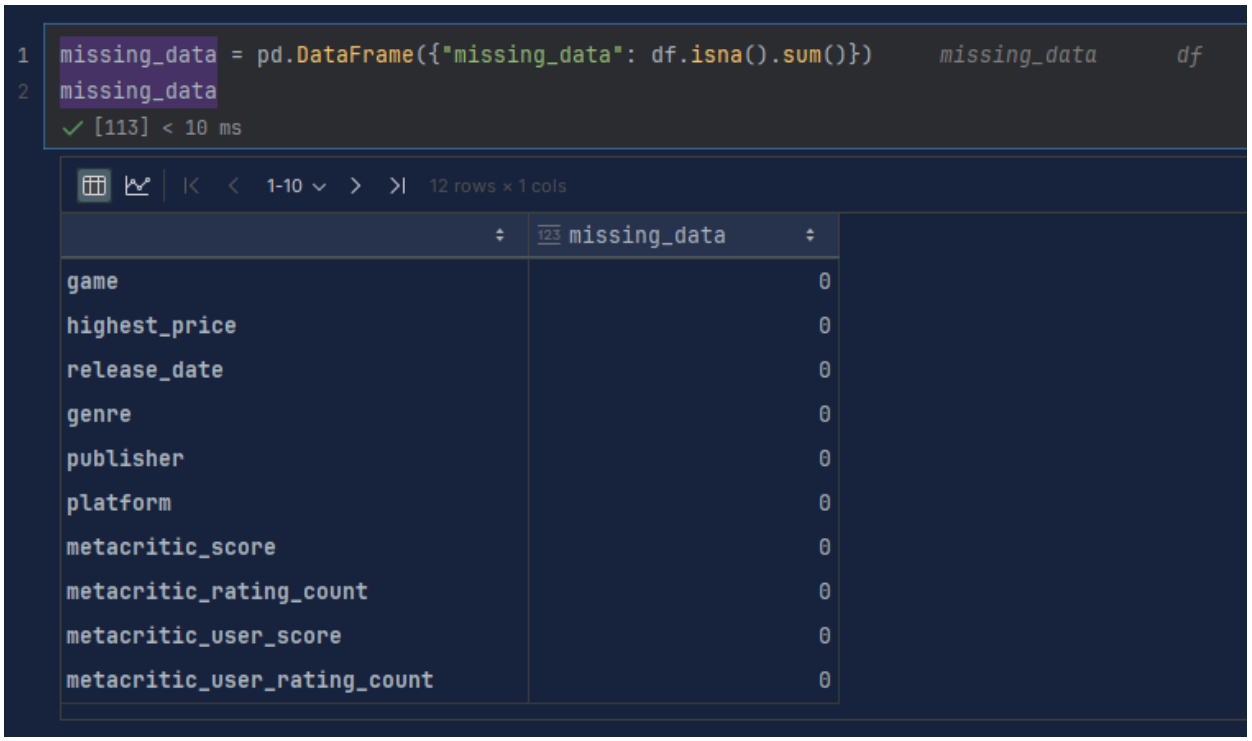

Рисунок 14 – Результат заполнения пропусков

Предобработанные данные были сохранены в формате csv (рисунок 15).

```
df.to_csv("preprocessed_df.csv")
```

Рисунок 15 – Сохранение предобработанных данных

ЗАКЛЮЧЕНИЕ

В ходе выполнения лабораторной работы был выбран датасет PlayStation Games Info 2/15/2025, содержащий подробную информацию об играх для PlayStation, объединяющую официальные данные PlayStation Store с отзывами критиков и пользователей Metacritic.

Для данного датасета был создан файл README.md с информацией о содержании столбцов. В рамках разведывательного анализа данных был построен ряд диаграмм: гистограмма, отображающая распределение количества игр по их максимальной цене, "ящик с усами" для распределения значений цены, круговая диаграмма о платформах, тепловая карта, отображающая значения взаимной корреляции для числовых характеристик игр (цена, рейтинг критиков, количество рецензий критиков, средняя оценка пользователей, количество оценок пользователей на Metacritic и PS Store), группированная столбчатая диаграмма с распределением игр по платформам и жанрам.

В рамках предобработки данных пропуски в датасете были заполнены. Для числовых столбцов было использовано среднее значение, а для нечисловых столбцов - мода.

Датасет с предобработанными данными был сохранен в файл preprocessed_df.csv.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Википедия. Анализ данных. URL: https://ru.wikipedia.org/wiki/Анализ_данных (дата обращения: 16.11.2025).

2 Почему аналитик данных? Актуальность и перспективность профессии. URL: https://blog.sf.education/pochemu-analitik-dannyh-aktualnost-i-perspektivnost-professii/ (дата обращения: 16.11.2025).

3 Топ-9 библиотек в Python для профессионального анализа данных. URL: https://practicum.yandex.ru/blog/biblioteki-python-dlya-data-science/ (дата обращения: 16.11.2025).

4 Википедия. Диаграмма. URL: https://ru.wikipedia.org/wiki/Диаграмма (дата обращения: 16.11.2025).