Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

Изображение выглядит как текст, керамические изделия, фарфор

Автоматически созданное описание

**Курсовая работа по дисциплине**

**«Методы машинного обучения»**

**на тему:**

**«Исследование продаж на рынке видеоигр»**

**ИСПОЛНИТЕЛЬ:**

Мокренко Никита Валерьевич

ИУ5-34М

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

"\_\_"\_\_\_\_\_\_\_\_\_\_\_2021 г.

Москва 2021

# Оглавление

[Оглавление 2](#_Toc91616257)

[Задание 3](#_Toc91616258)

[Подготовка данных 4](#_Toc91616259)

[Загрузка датасета 4](#_Toc91616260)

[Устранение пропусков данных 4](#_Toc91616261)

[Обработка категориальных признаков 4](#_Toc91616262)

[Нормализация числовых признаков 5](#_Toc91616263)

[Масштабирование признаков 5](#_Toc91616264)

[Отбор признаков 5](#_Toc91616265)

[Результат работы моделей 7](#_Toc91616266)

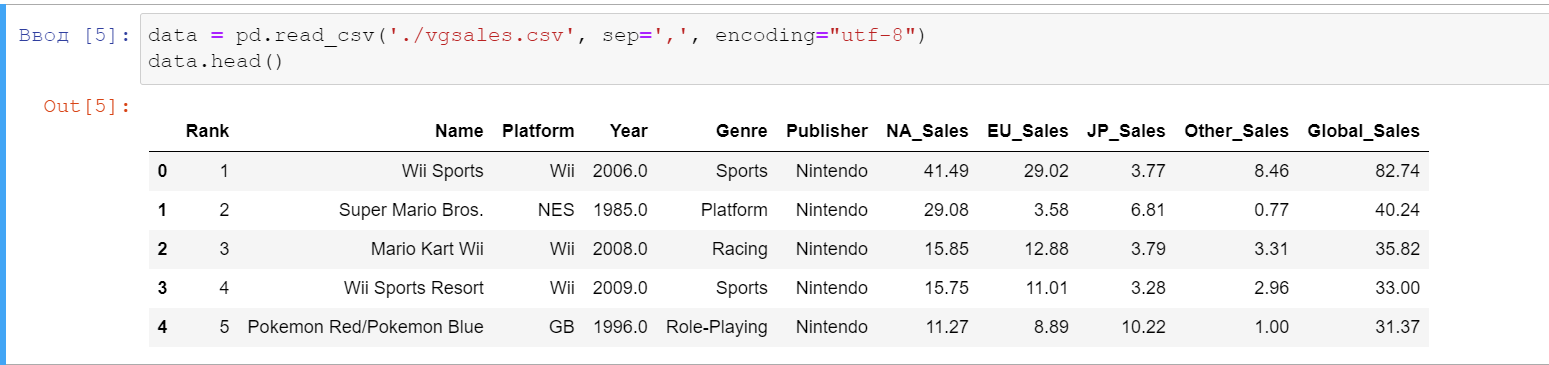
[AutoML 8](#_Toc91616267)

# Задание

1. Поиск и выбор набора данных для построения модели машинного обучения. На основе выбранного набора данных строится модель для задачи классификации.
2. Для выбранного датасета решить следующие задачи:
   1. устранение пропусков в данных;
   2. кодирование категориальных признаков;
   3. нормализацию числовых признаков;
   4. масштабирование признаков;
   5. обработку выбросов для числовых признаков;
   6. обработку нестандартных признаков (которые не является числовым или категориальным);
   7. отбор признаков, наиболее подходящих для построения модели;
3. Обучить модель и оценить метрики качества для двух выборок:
   1. исходная выборка, которая содержит только минимальную предобработку данных, необходимую для построения модели (например, кодирование категориальных признаков).
   2. улучшенная выборка, полученная в результате полной предобработки данных в пункте 2.
4. Построить модель с использованием произвольной библиотеки AutoML.
5. Сравнить метрики для трех полученных моделей.

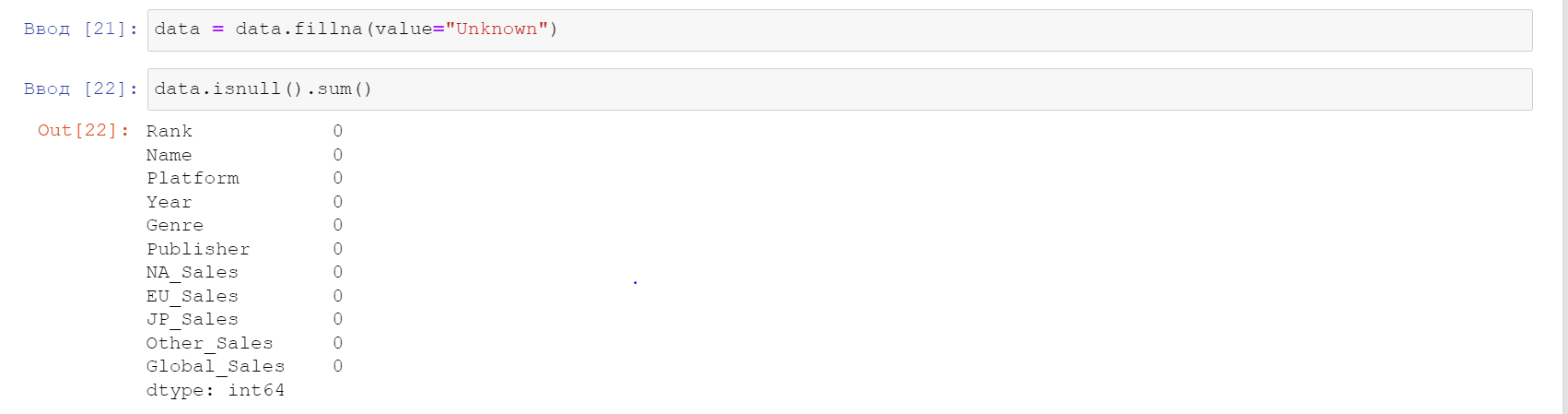
# Подготовка данных

## Загрузка датасета

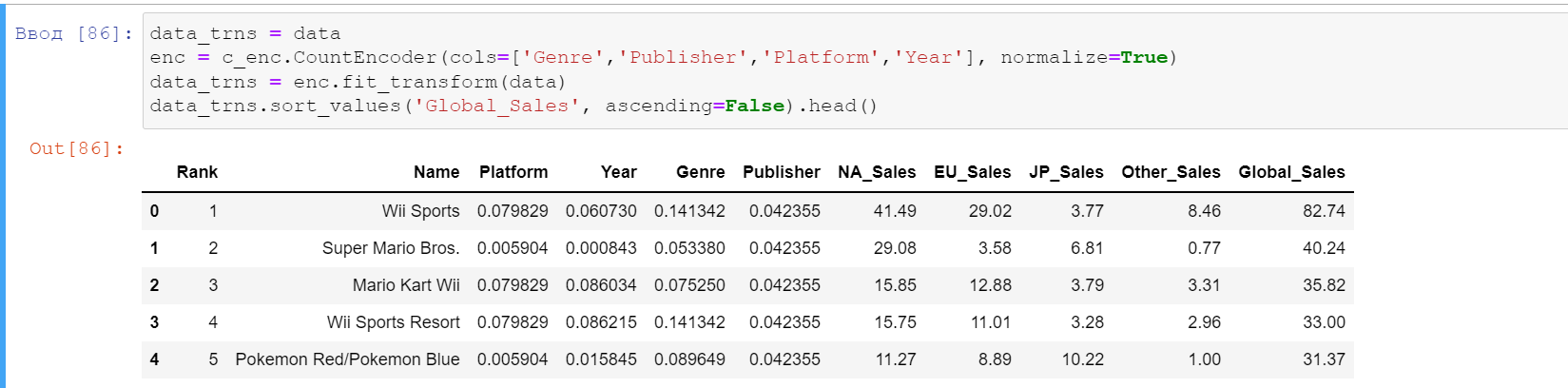


## Устранение пропусков данных

Пропуски присутствуют в двух столбцах. Так как оба признака являются категориальными и количество пропусков незначительно, можно выполнить замену пропусков одним значением («Unknown»), либо вообще исключить строки с пропусками.



## Обработка категориальных признаков

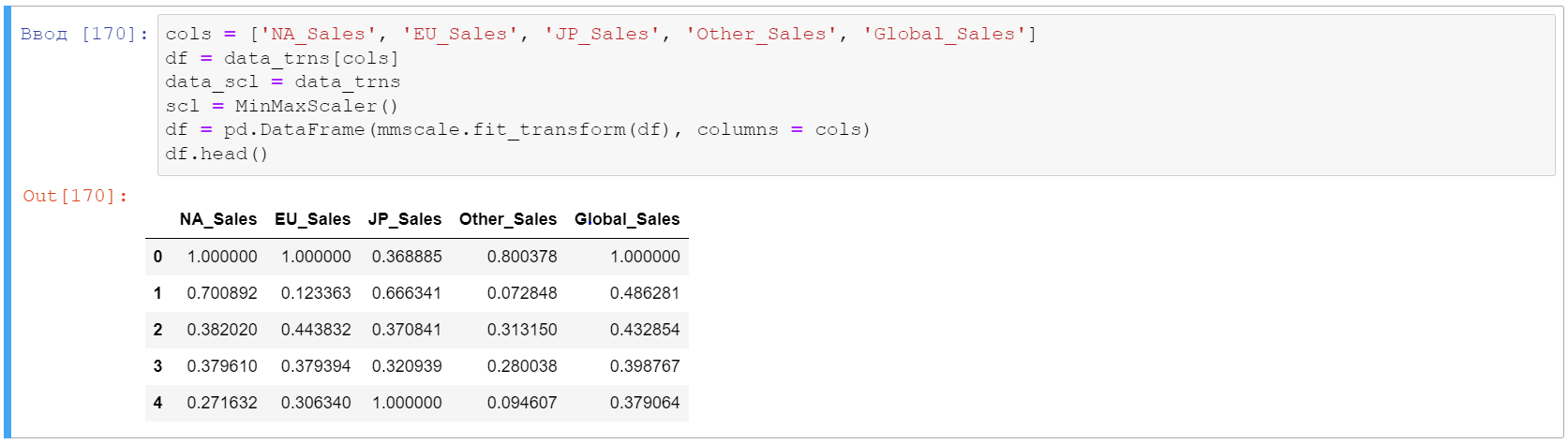
Можно использовать Count Encoder для признаков Genre, Platform, Year и Publisher.

## Нормализация числовых признаков

Не провожу нормализацию, т.к. в данных много нулевых значений.

## Масштабирование признаков

Использую MinMaxScaler для масштабирования признаков, относящихся к продажам.



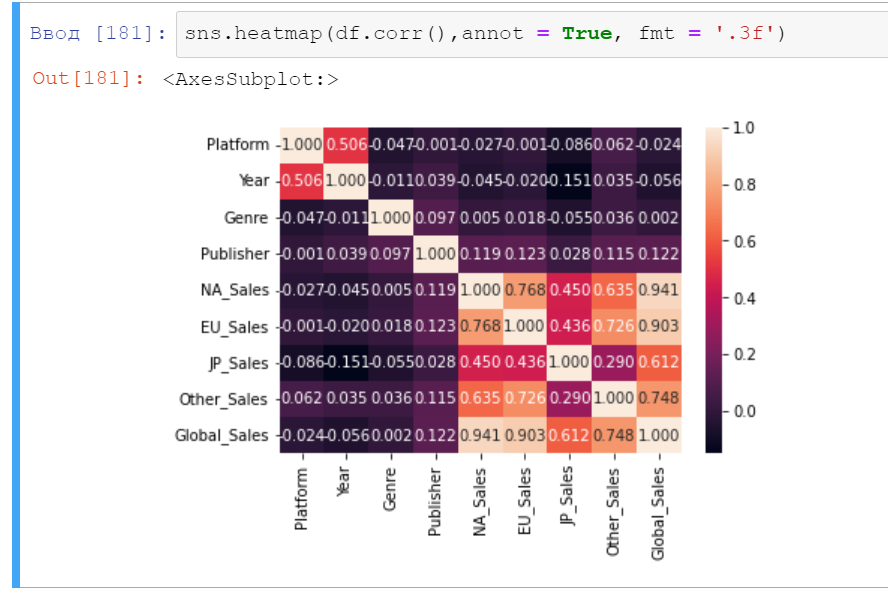
## Отбор признаков

Можно пока исключить признак Rank, т.к. это уникальное значение, обозначающее место в рейтинге. Гораздо разумнее предсказывать продажи.

Проверим датасет на наличие константных признаков:

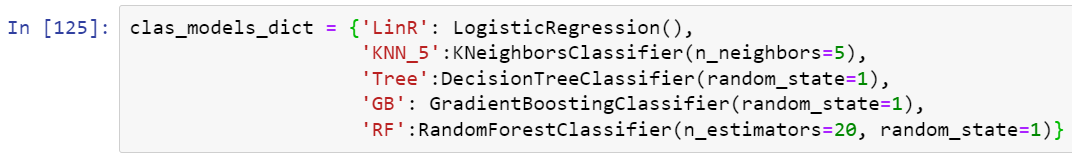


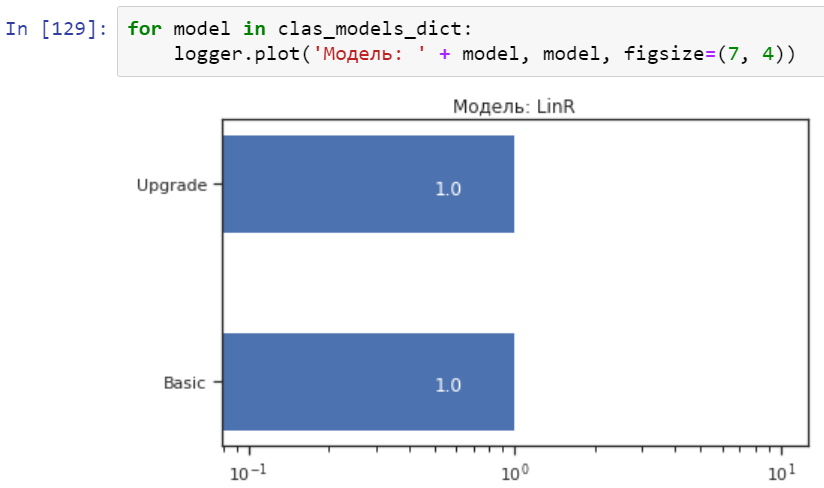
Теперь посмотрим корреляционную матрицу:



Видно, что признак продаж по миру сильно коррелирует с продажами в Европе и Северной Америке (что логично), так что при построении модели машинного обучения его можно будет убрать.

# Результат работы моделей





# AutoML

Обучение при помощи технологии AutoML

