МГТУ им. Н. Э. Баумана кафедра ИУ5

курс «Технологии машинного обучения»

Лабораторная работа №3

**«Обработка пропусков в данных, кодирование категориальных признаков, масштабирование данных»**

ВЫПОЛНИЛ:

Гаранин А. В. Группа ИУ5Ц-81Б

ПРОВЕРИЛ:

Гапанюк Ю. Е.

Москва, 2020 г.

**Цель лабораторной работы:** изучение способов предварительной обработки данных для дальнейшего формирования моделей.

# Задание:

1. Выбрать набор данных (датасет), содержащий категориальные признаки и пропуски в данных. Для выполнения следующих пунктов можно использовать несколько различных наборов данных (один для обработки пропусков, другой для категориальных признаков и т.д.)
2. Для выбранного датасета (датасетов) на основе материалов [лекции](https://github.com/ugapanyuk/ml_course_2020/blob/master/common/notebooks/missing/handling_missing_norm.ipynb) решить следующие задачи:
   1. обработку пропусков в данных;
   2. кодирование категориальных признаков;
   3. масштабирование данных.

# Выполнение работы

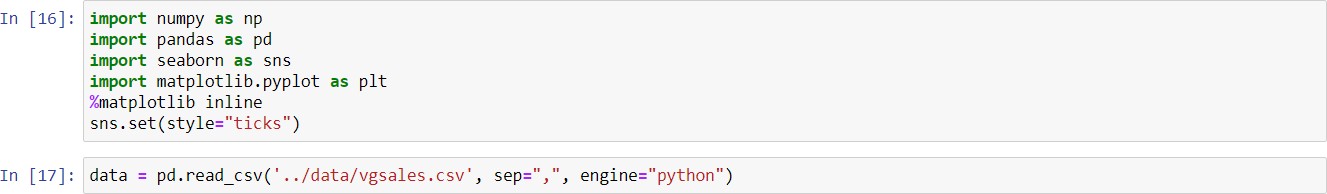
Выбранный набор данных: <https://www.kaggle.com/gregorut/videogamesales>

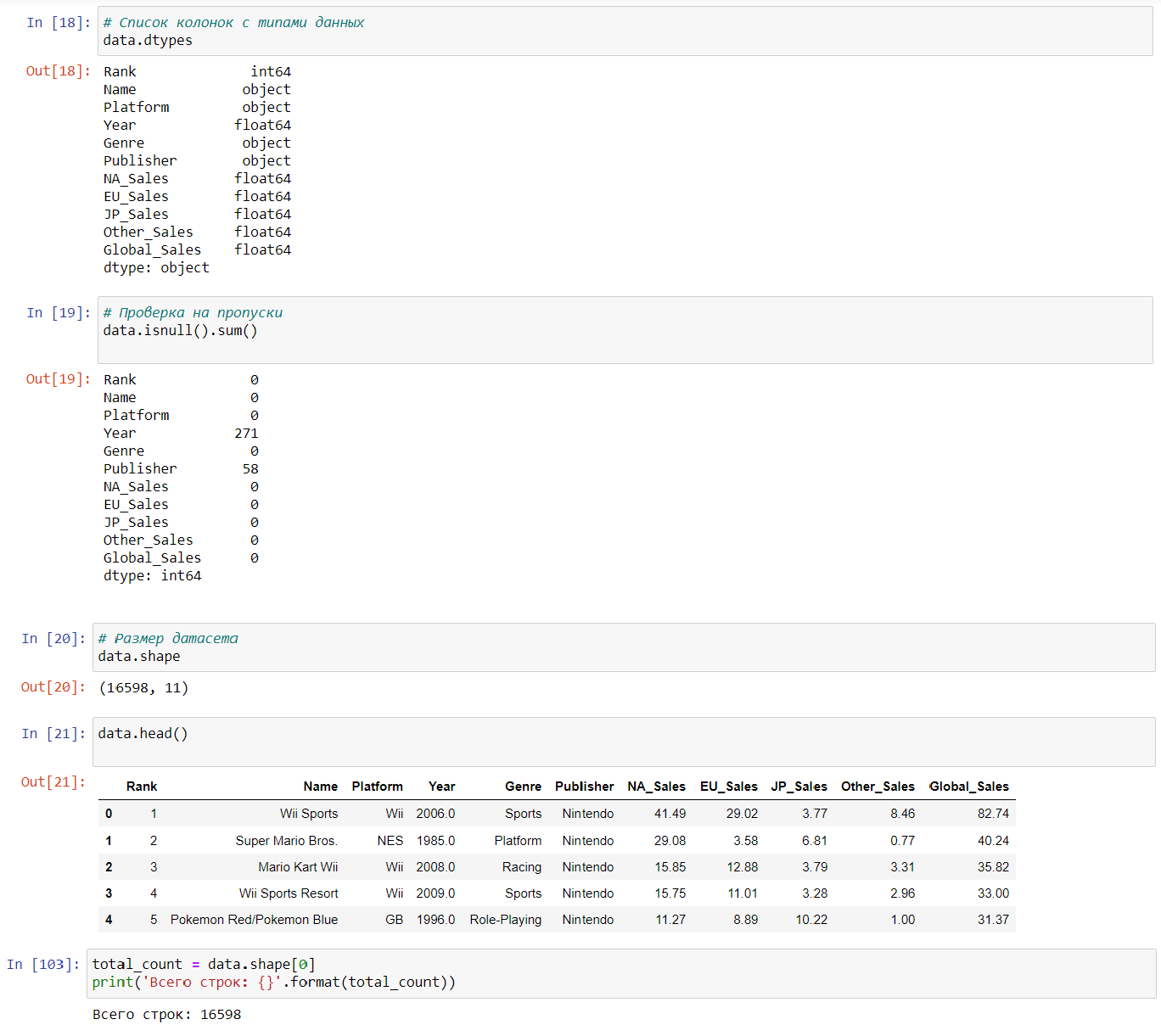
Этот набор данных содержит список видеоигр с объемом продаж более 100 000 копий.

Показатели:

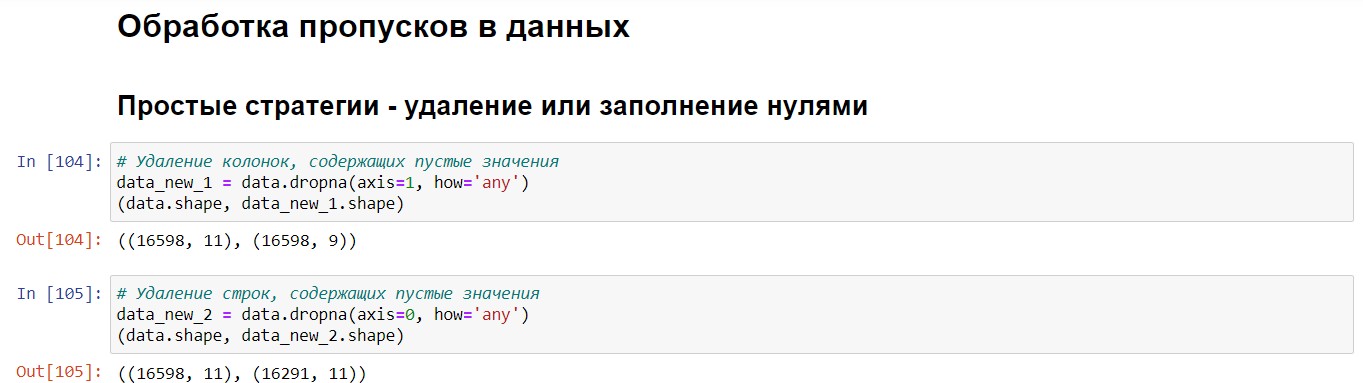
* Rank - Рейтинг общих продаж
* Name - Название игры
* Platform - платформа выпуска игр (т.е. ПК, PS4 и т. Д.)
* Year- год выпуска игры
* Genre - Жанр игры
* Publisher - издатель игры
* NA\_Sales - Продажи в Северной Америке (в миллионах)
* EU\_Sales - Продажи в Европе (в миллионах)
* JP\_Sales - Продажи в Японии (в миллионах)
* Other\_Sales - Продажи в остальном мире (в миллионах)
* Global\_Sales - общий объем продаж по всему миру.

1. Загрузка и первичный анализ:



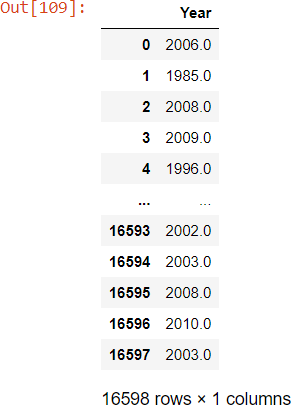
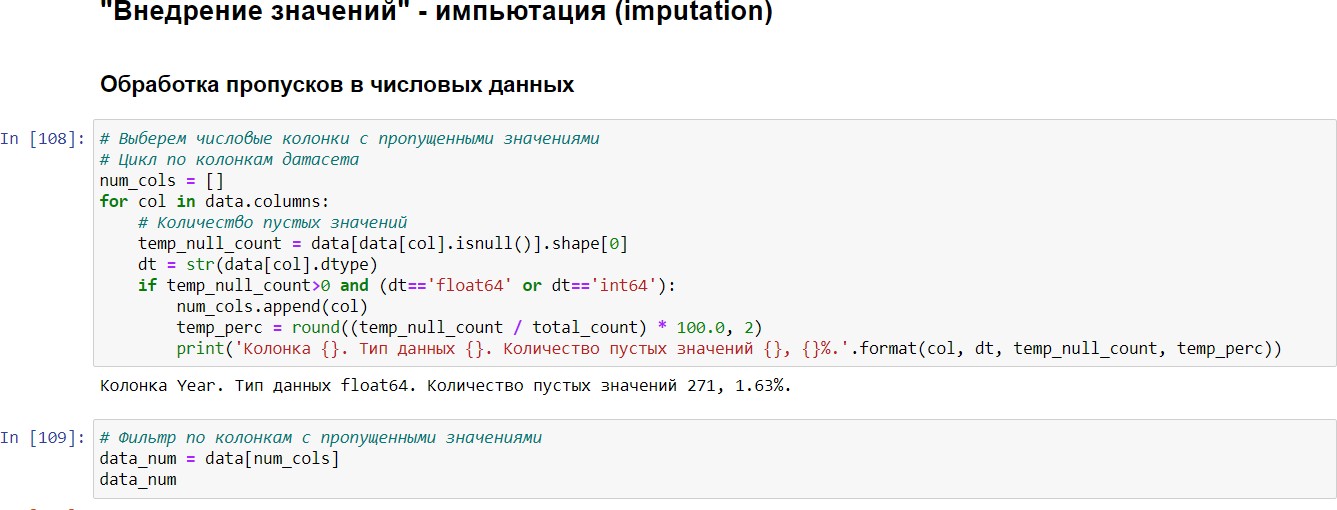


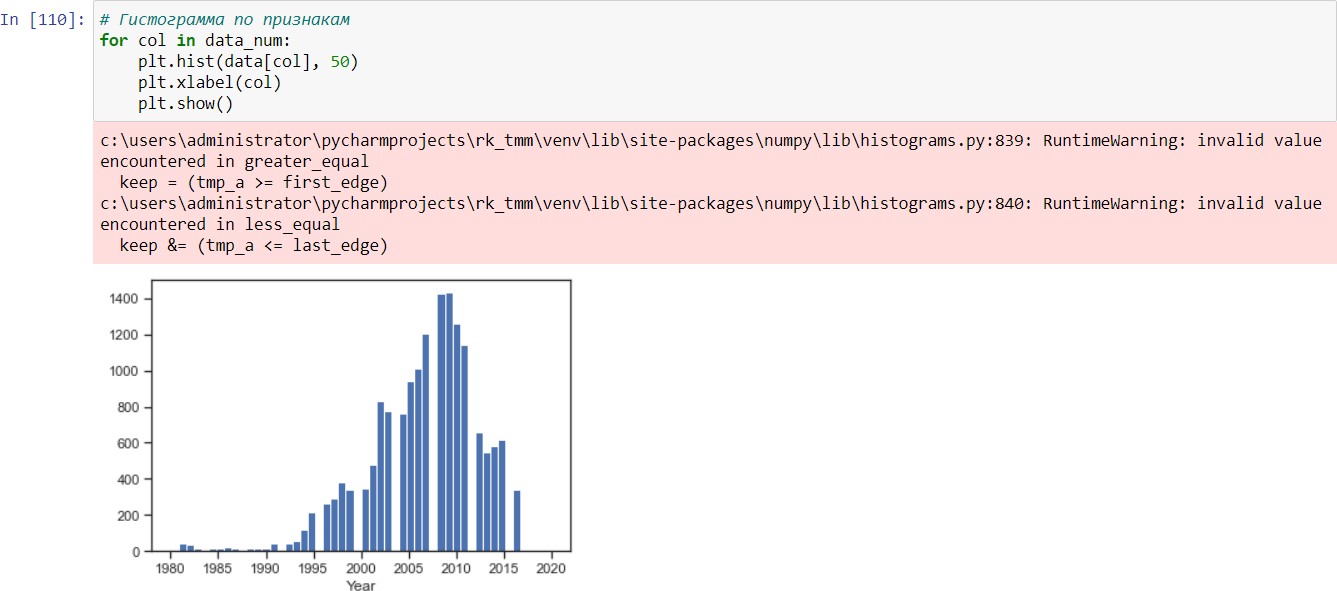
1. Обработка пропусков
   1. Простая стратегия – удаление или заполнение нулями

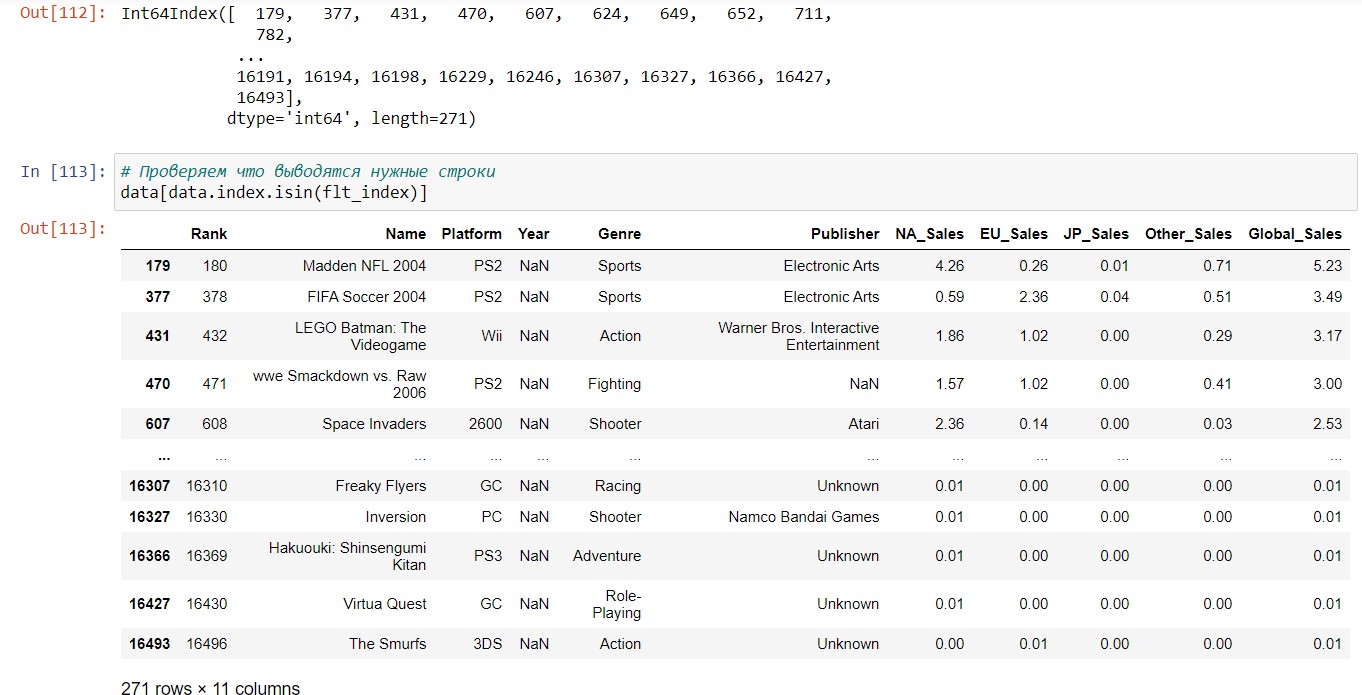
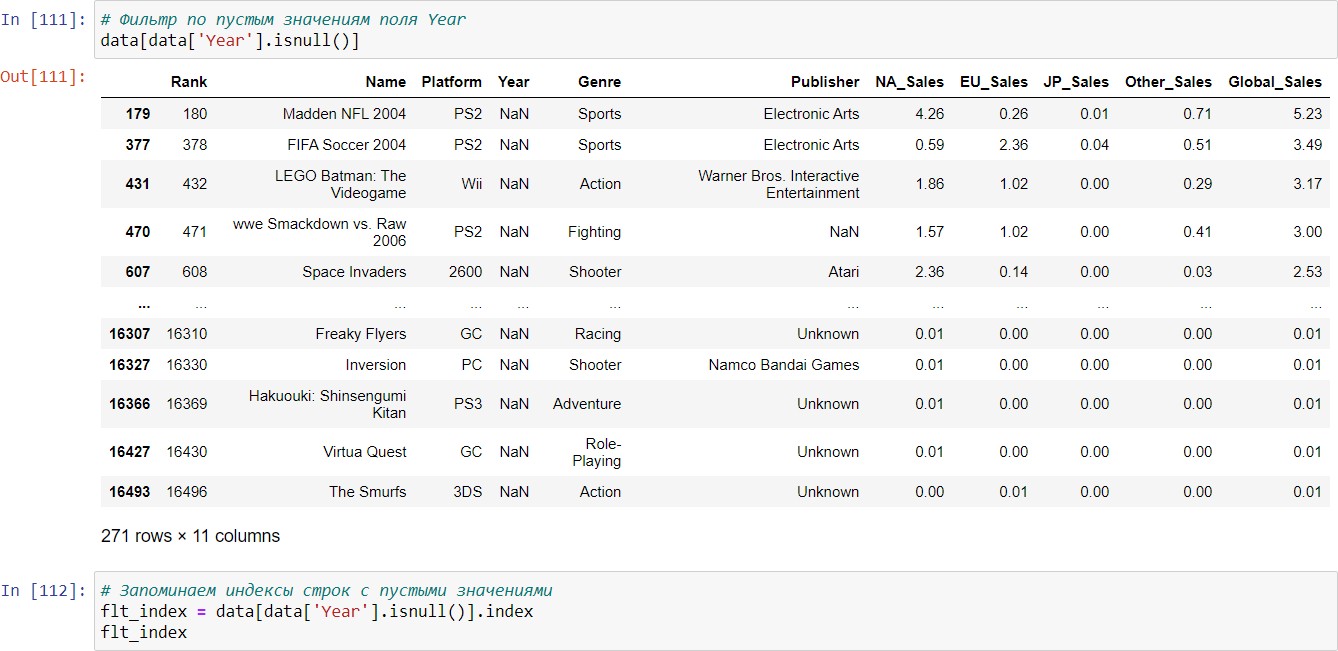




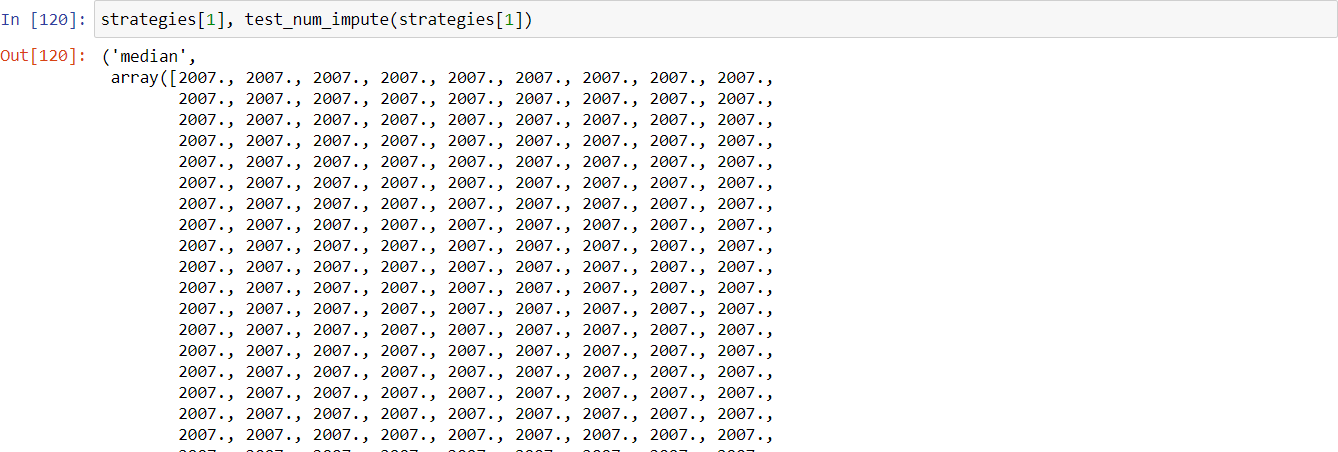
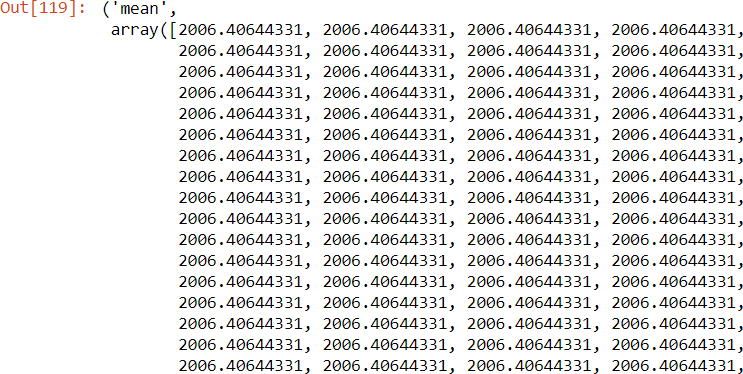
* 1. Импьютация
     1. Обработка пропусков в числовых данных

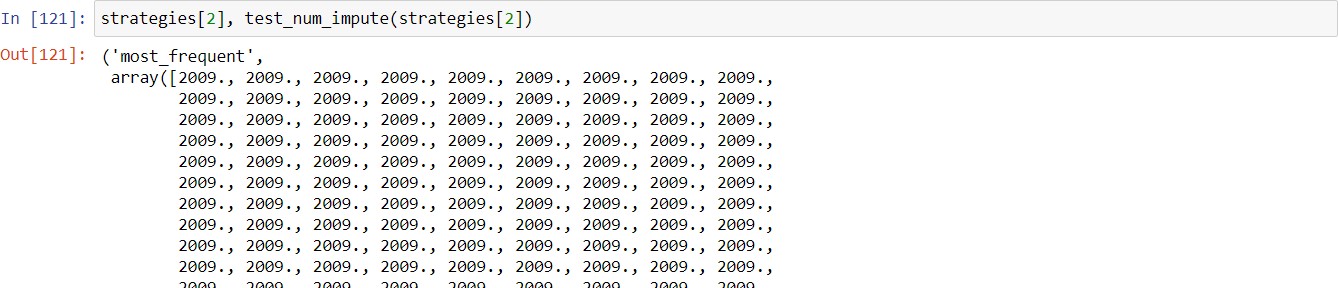






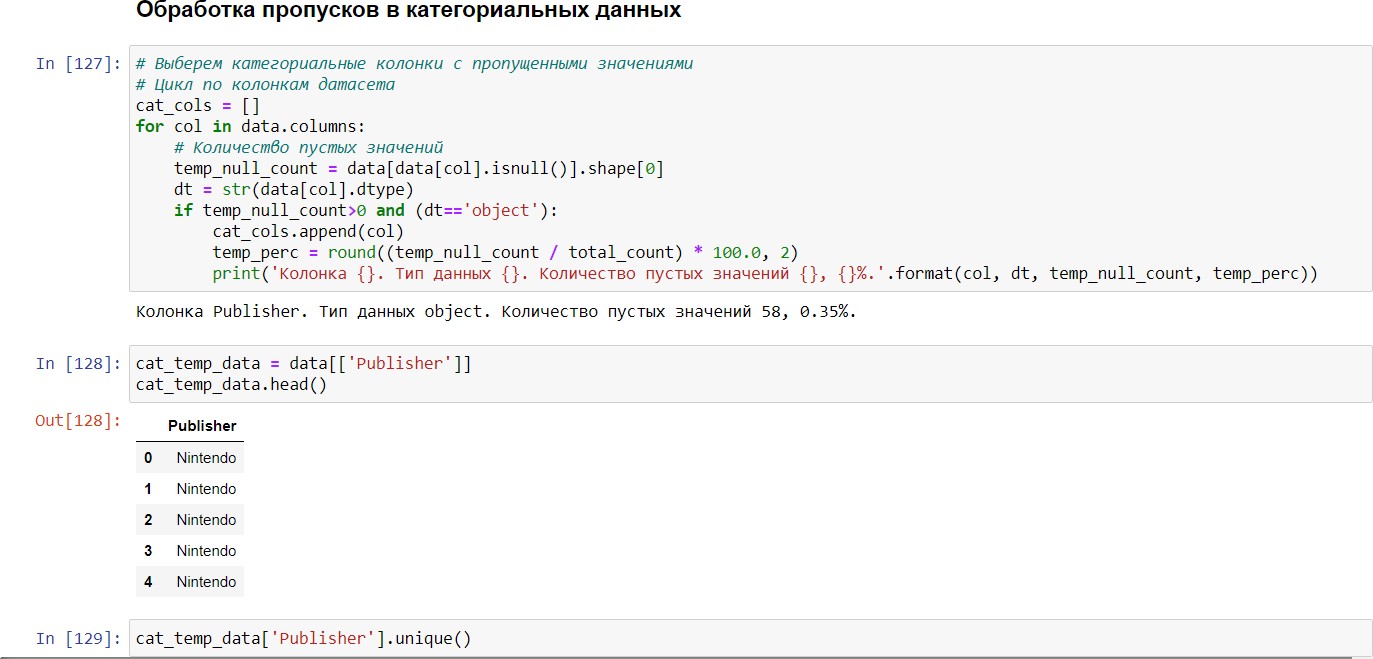


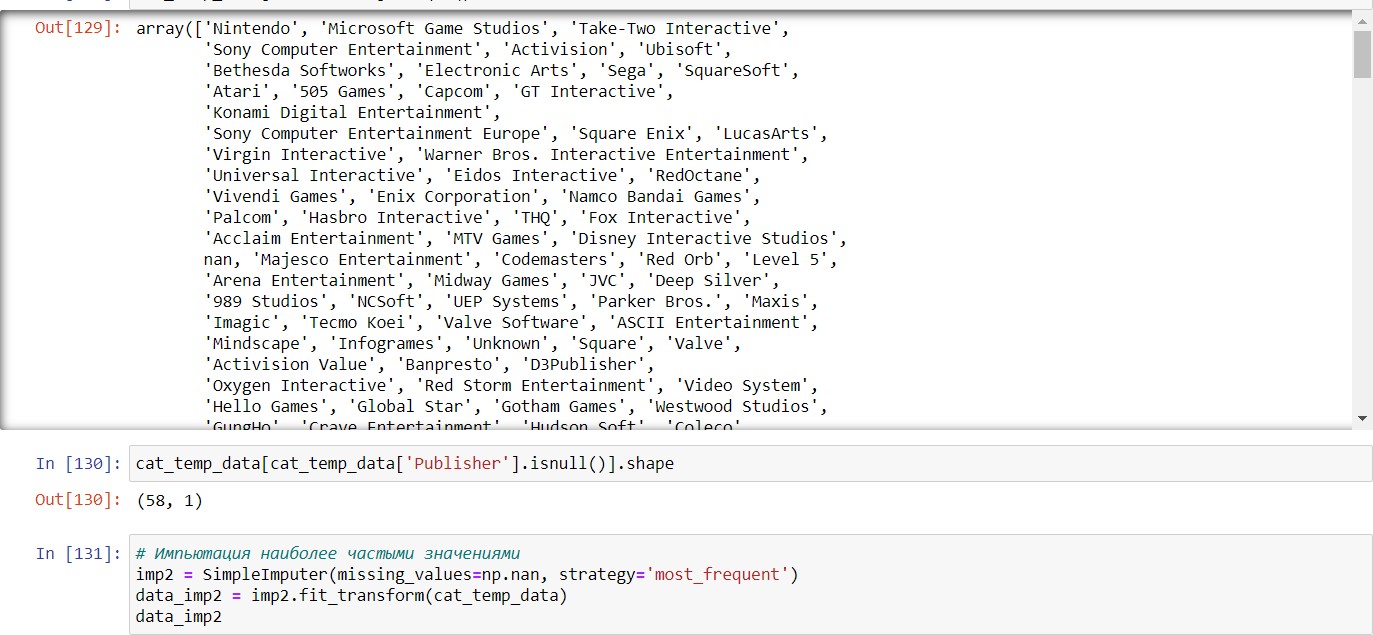


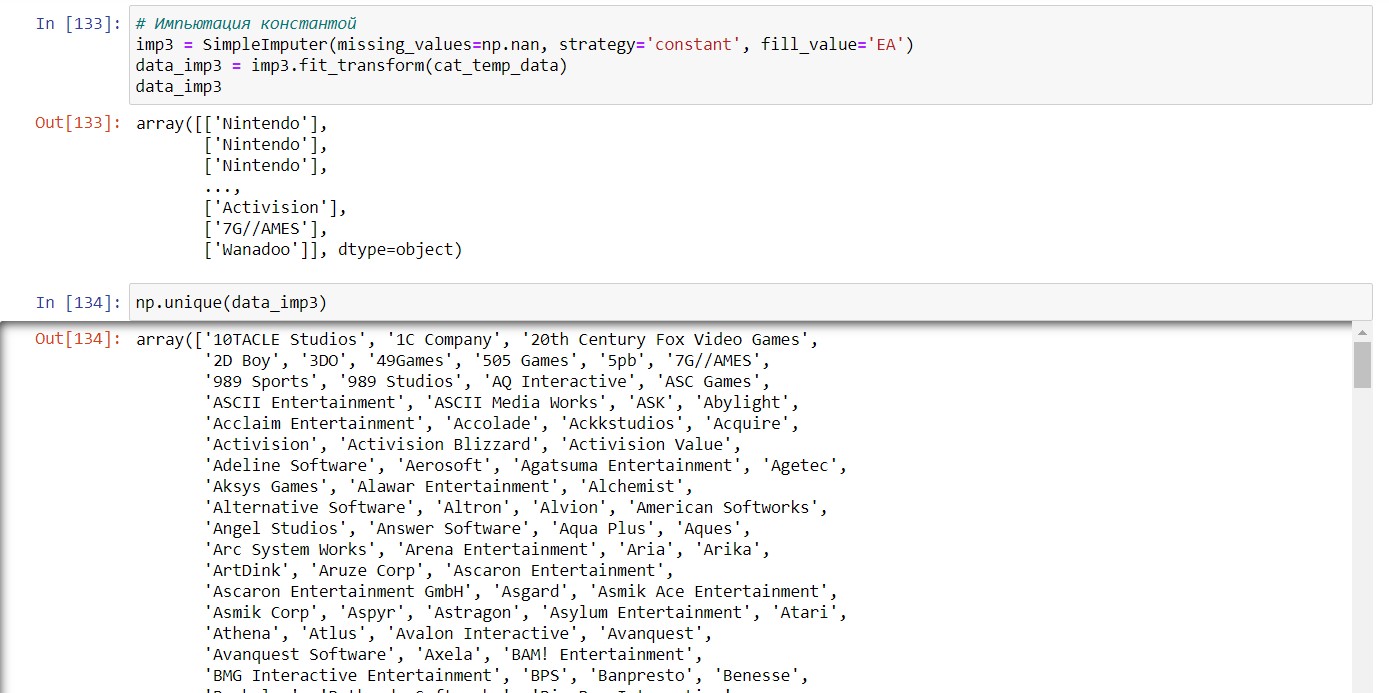
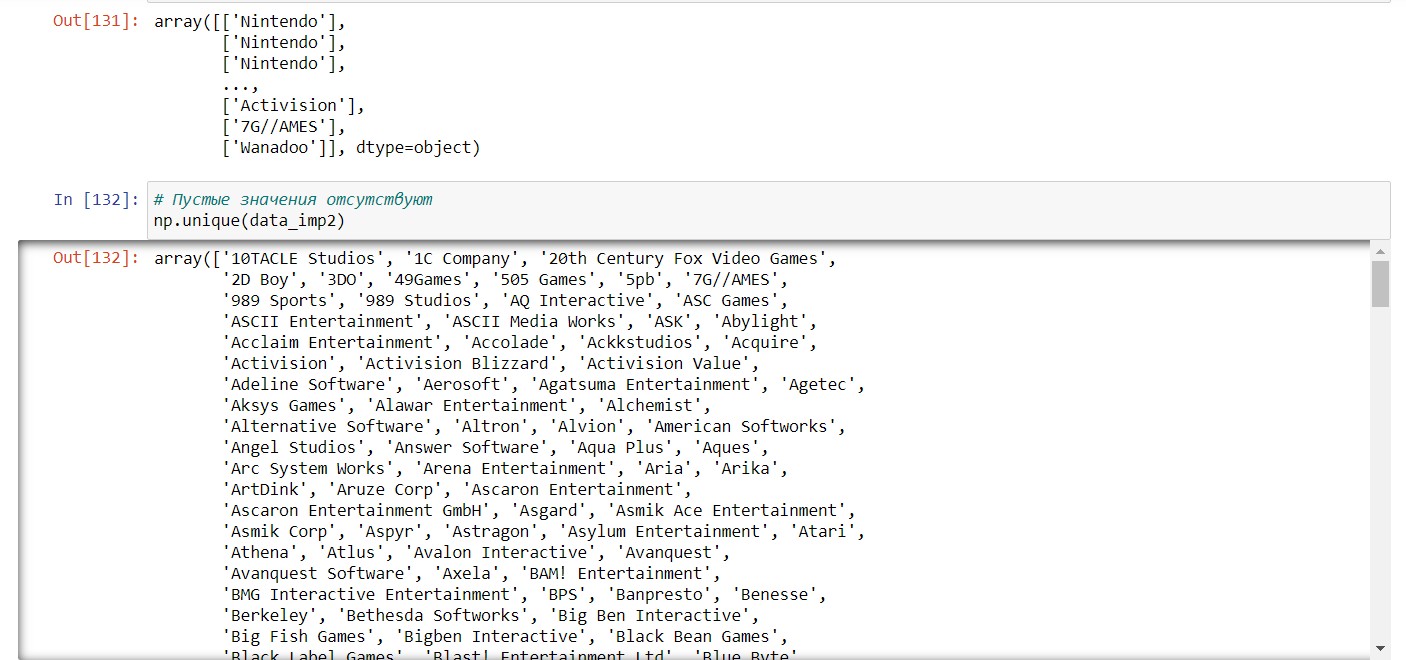


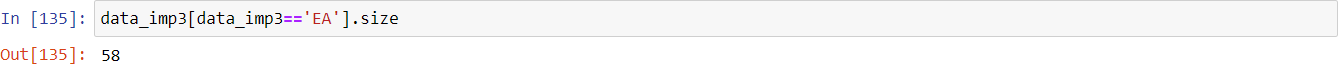


* + 1. Обработка пропусков в категориальных данных

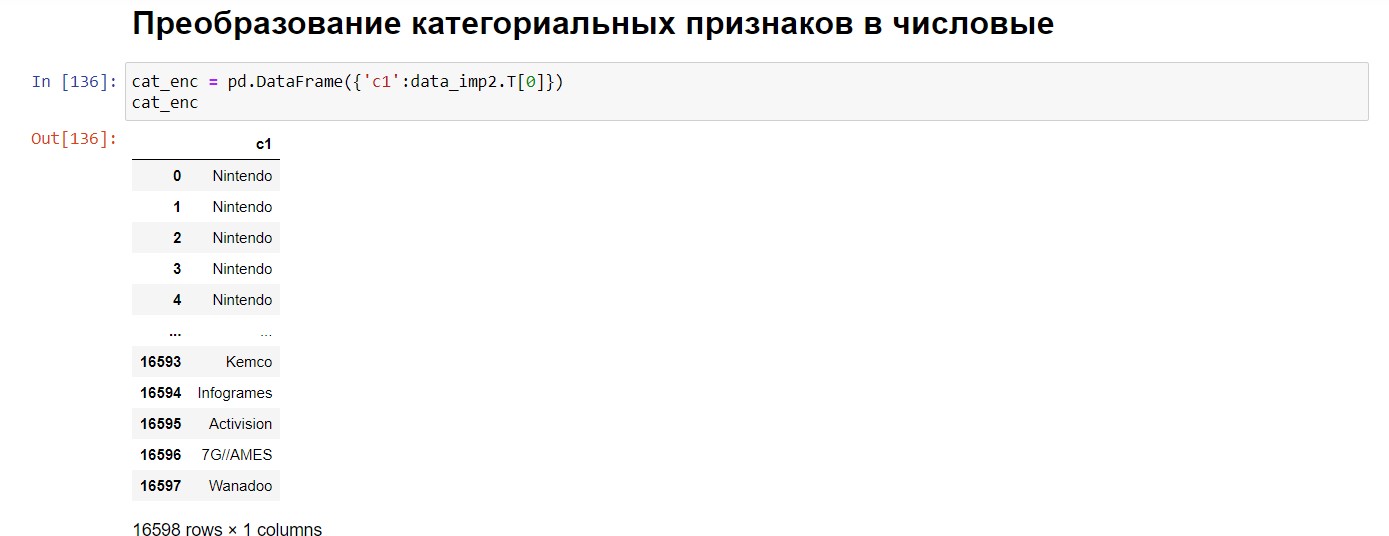




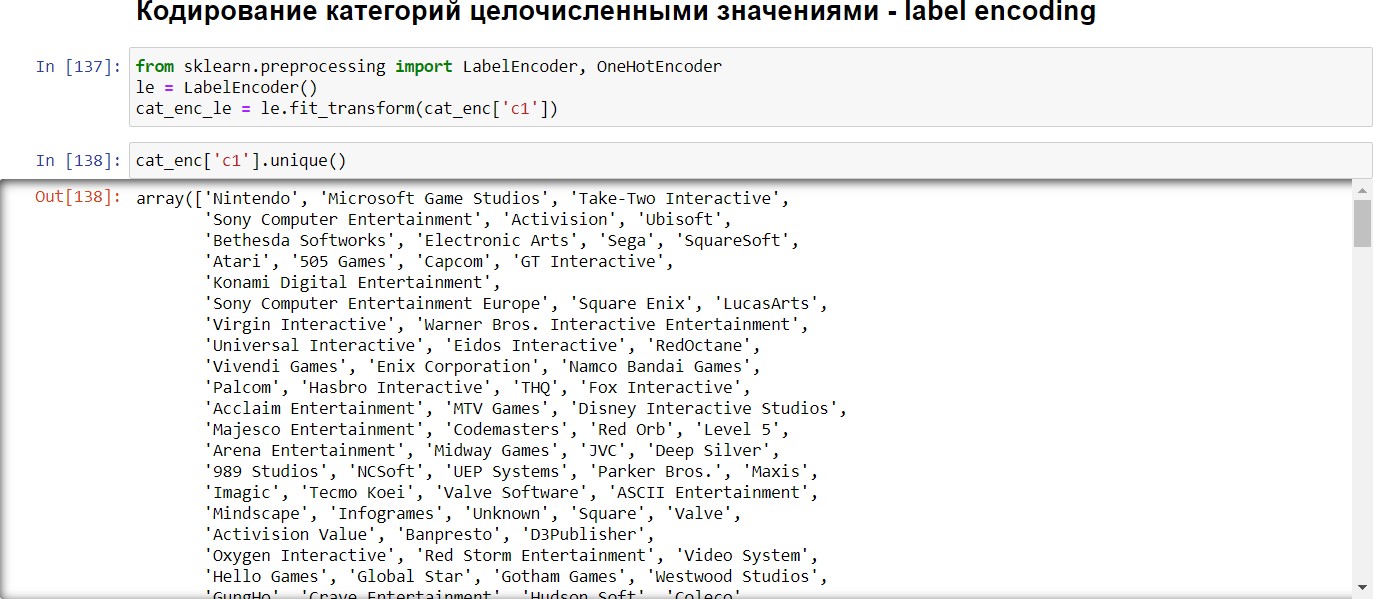


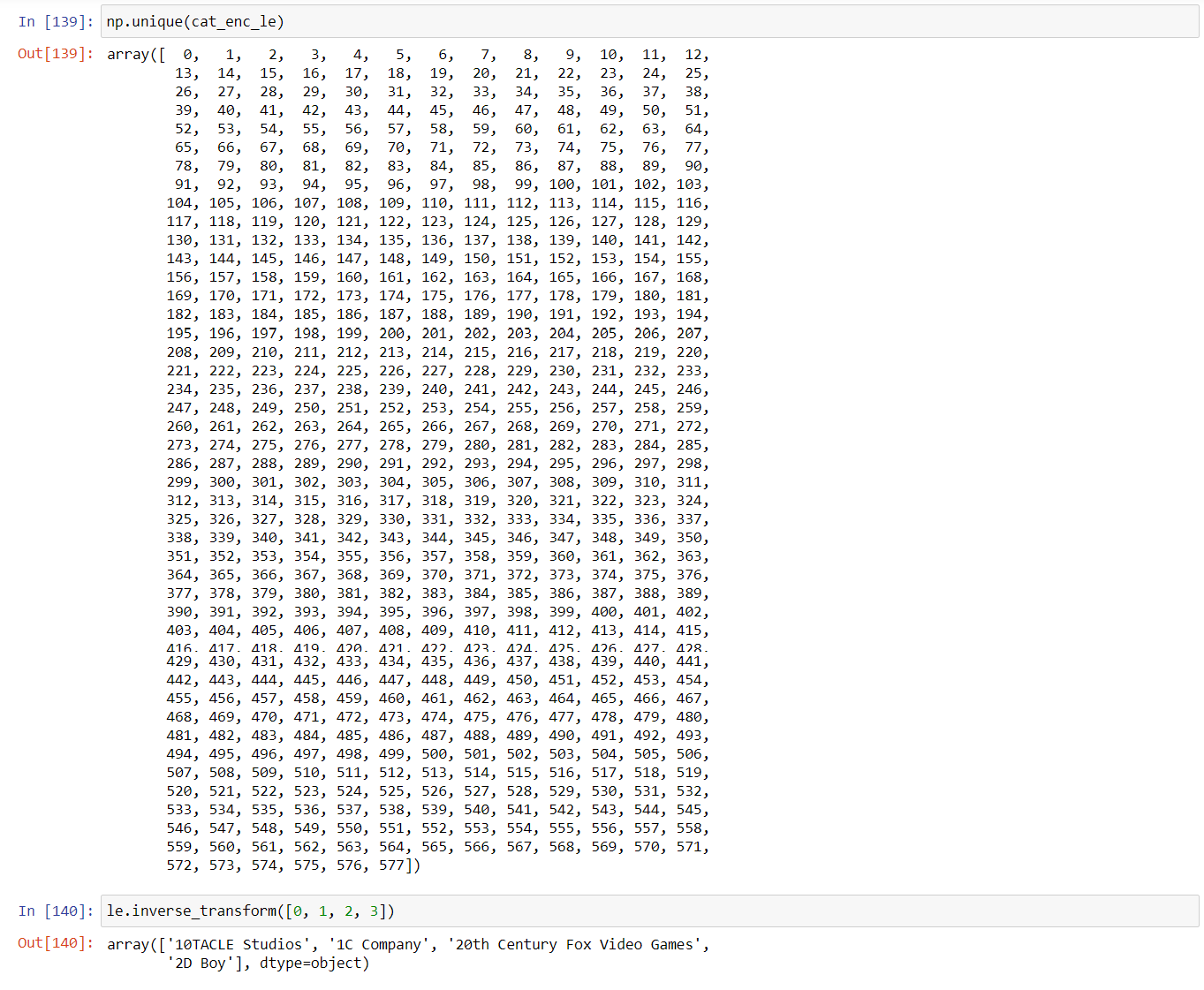


1. Преобразование категориальных признаков в числовые

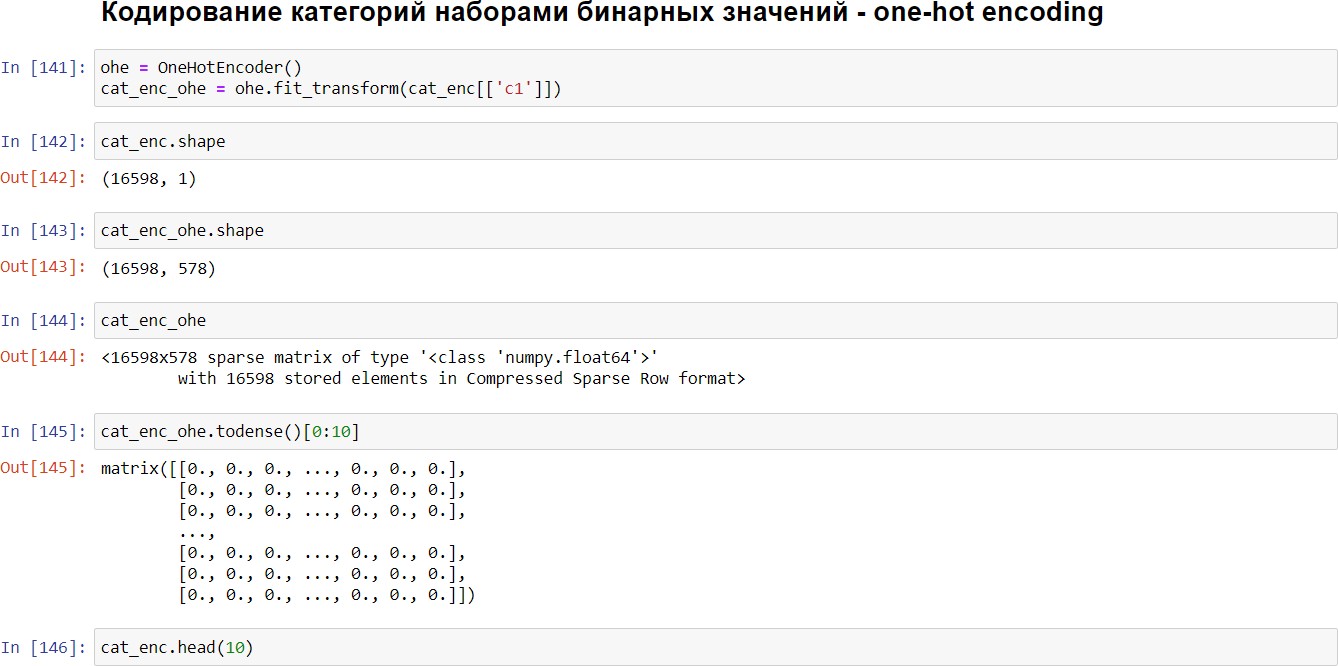


* 1. Кодирование категорий целочисленными значениями – label encoding





* 1. Кодирование категорий наборами бинарных значений – one-hot encoding

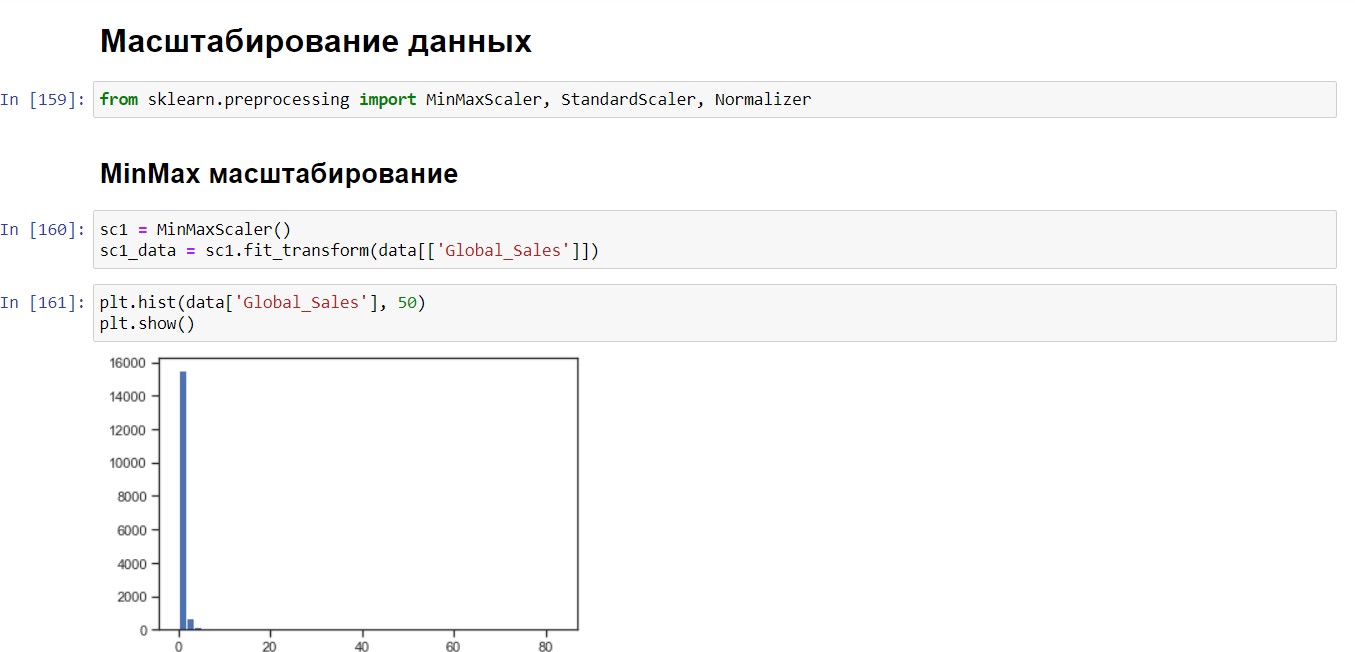


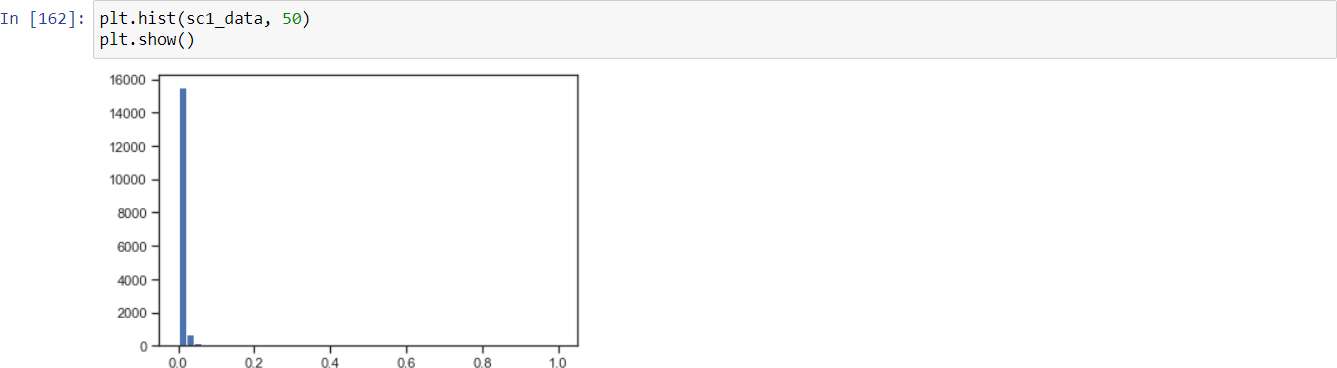


* 1. Pandas\_get\_dummies – быстрый вариант one-hot кодирования



1. Масштабирование данных
   1. MinMax масштабирование





* 1. Масштабирование данных на основе Z-оценки – StandardScaler



* 1. Нормализация данных

