Московский государственный технический университет имени Н. Э. Баумана

Кафедра «Системы обработки информации и управления»

Лабораторная работа №3

по курсу

«Методы машинного обучения»

на тему:

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

**Выполнил:**

Студент ИУ5-24М

Черната Н. С.

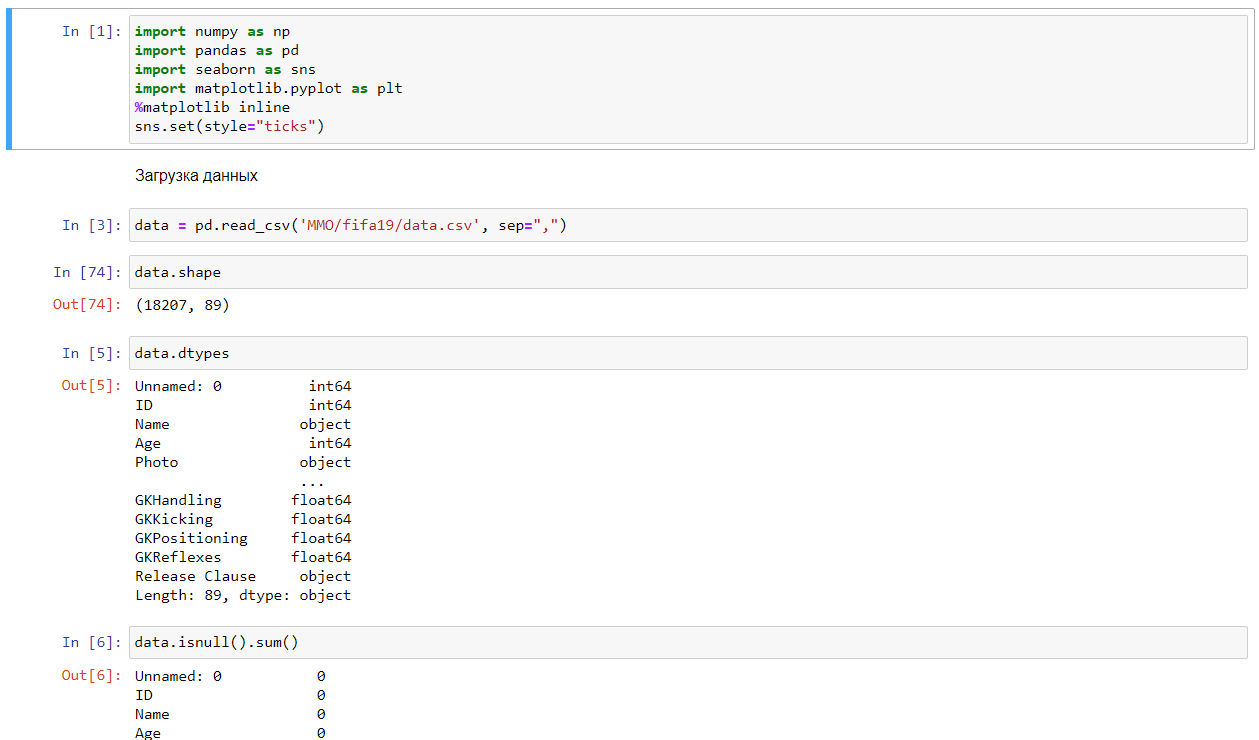
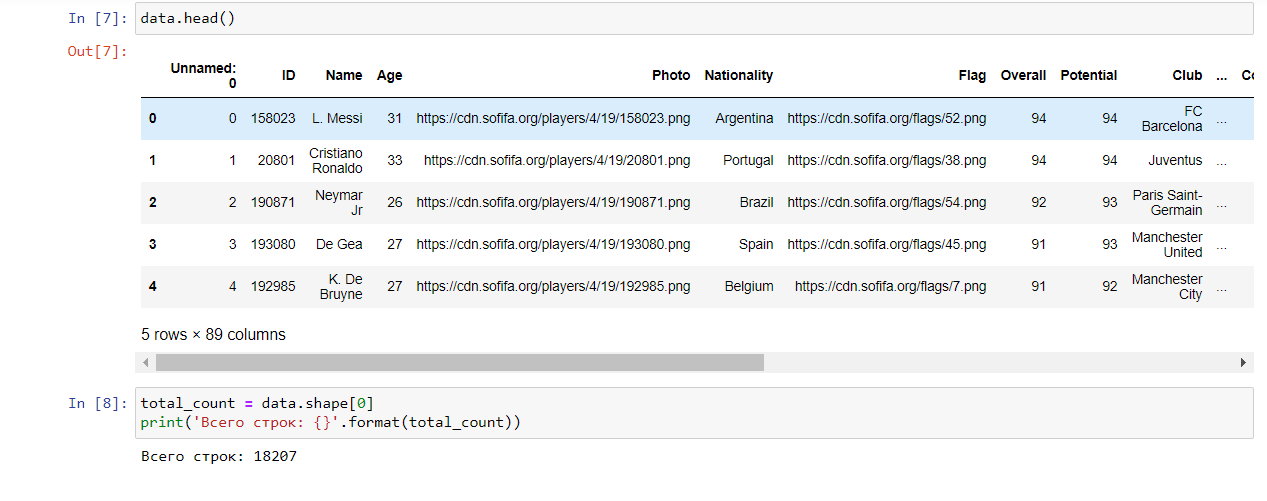
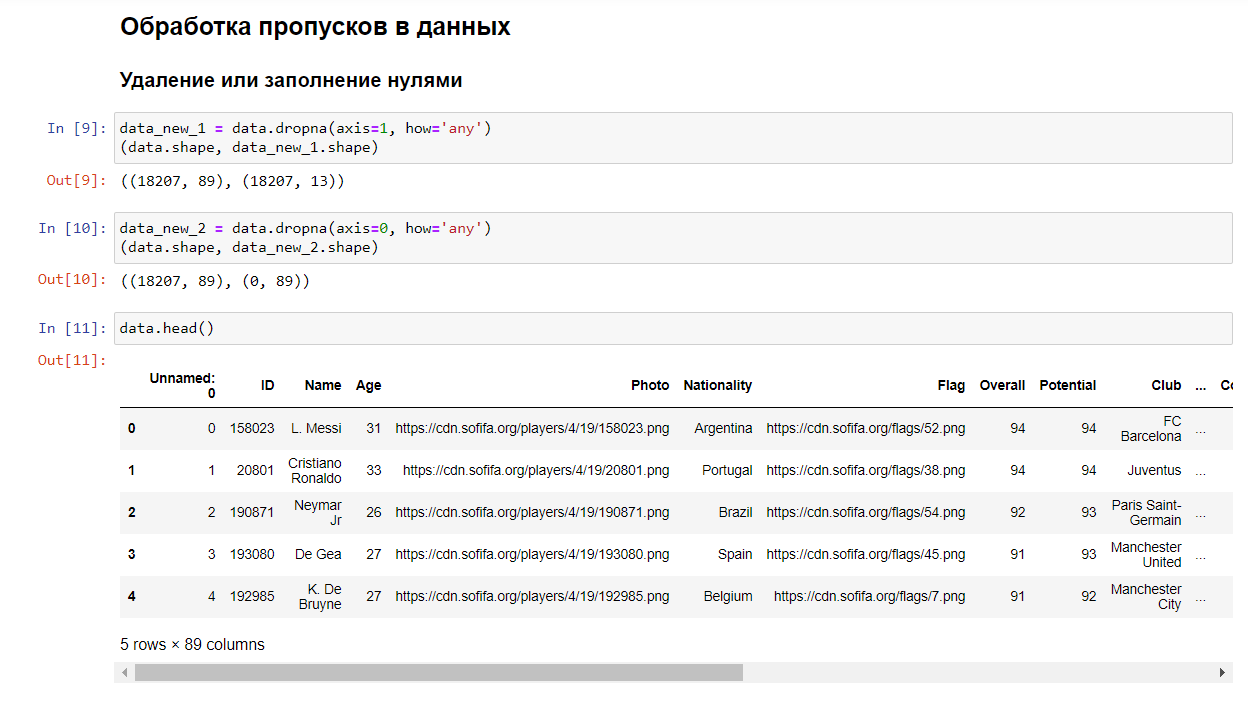
Москва, 2020

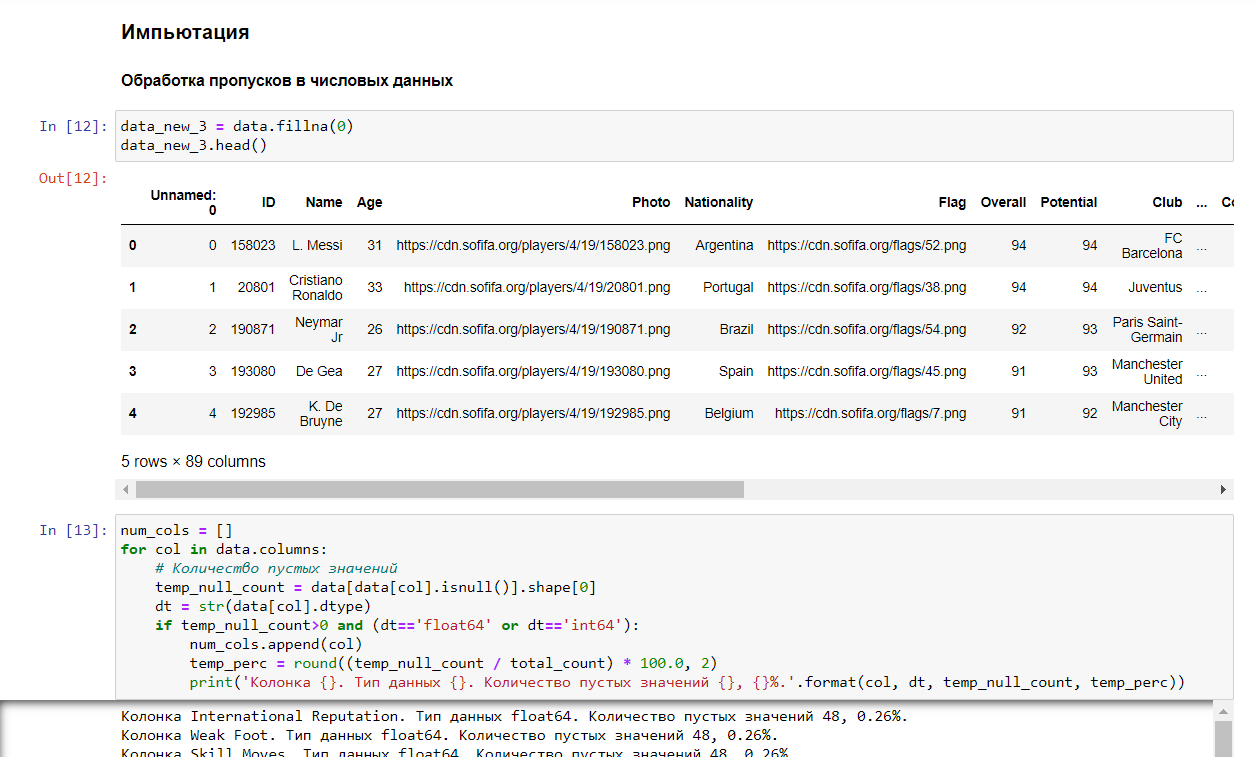
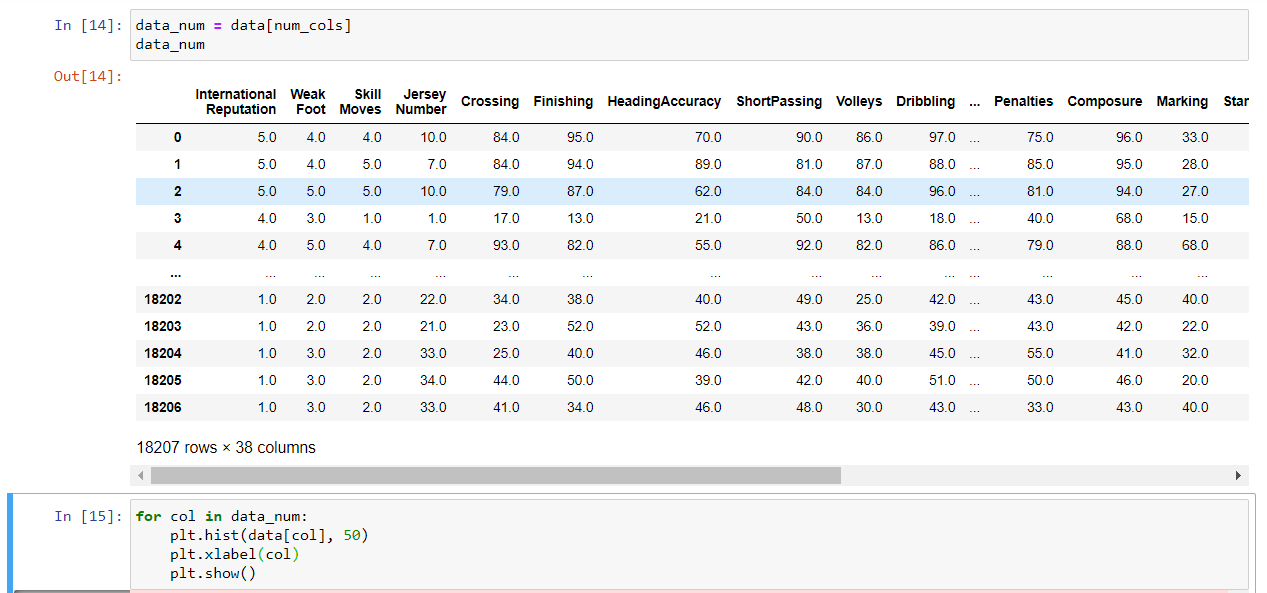
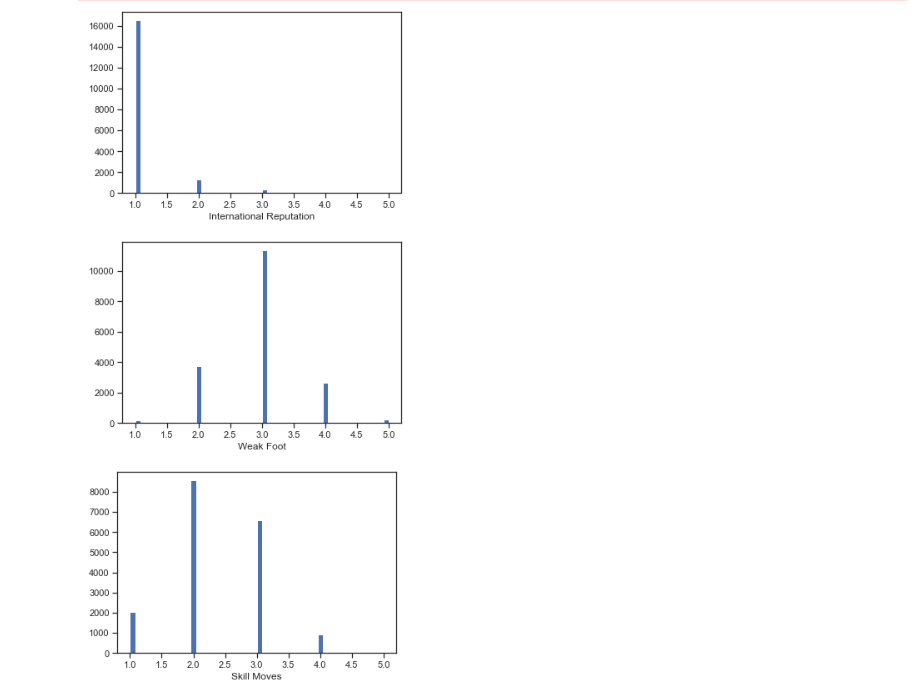
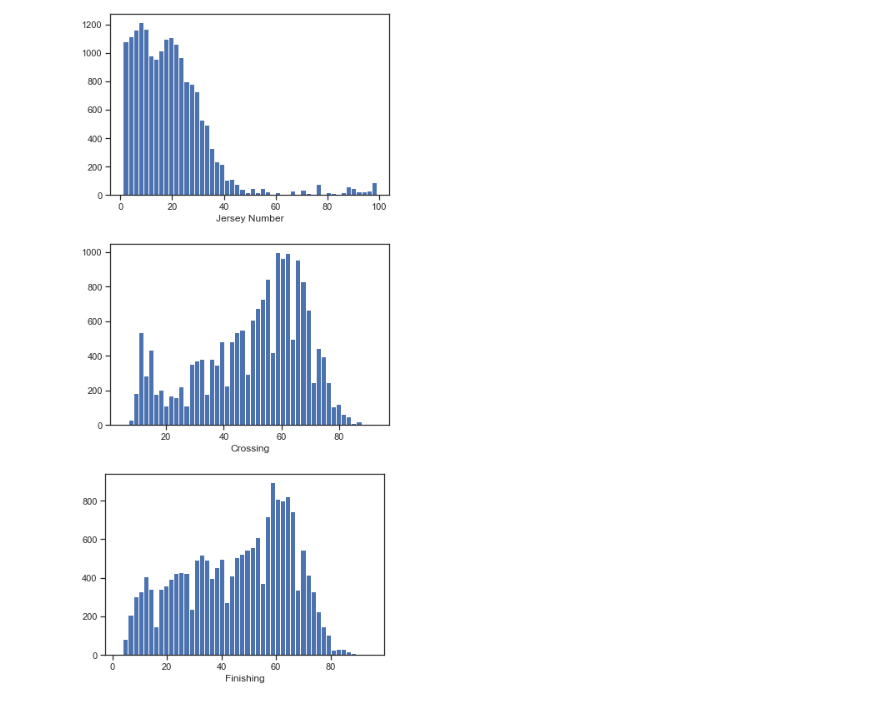
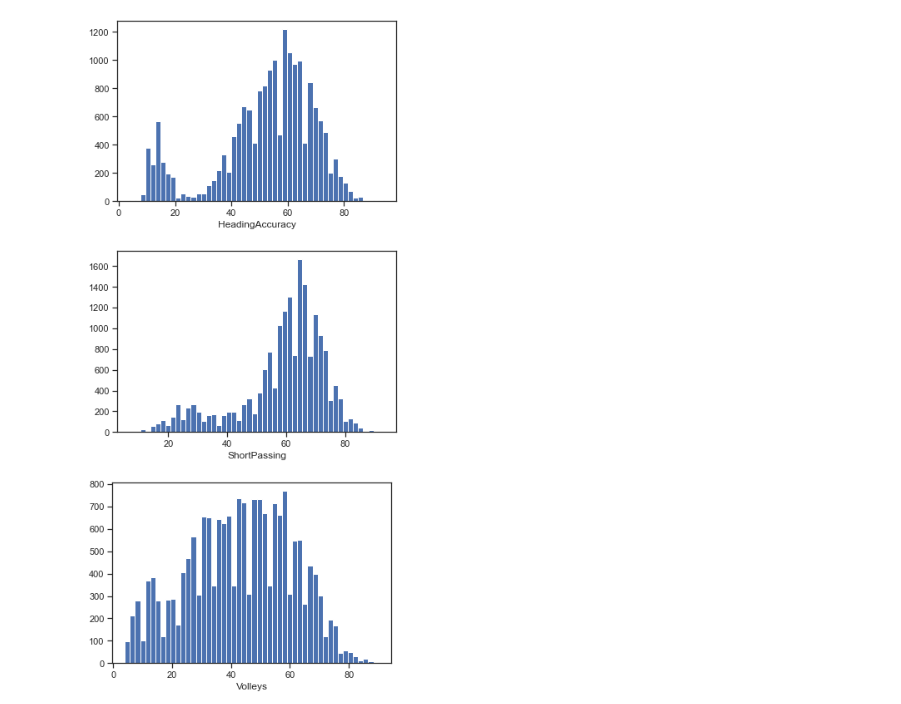
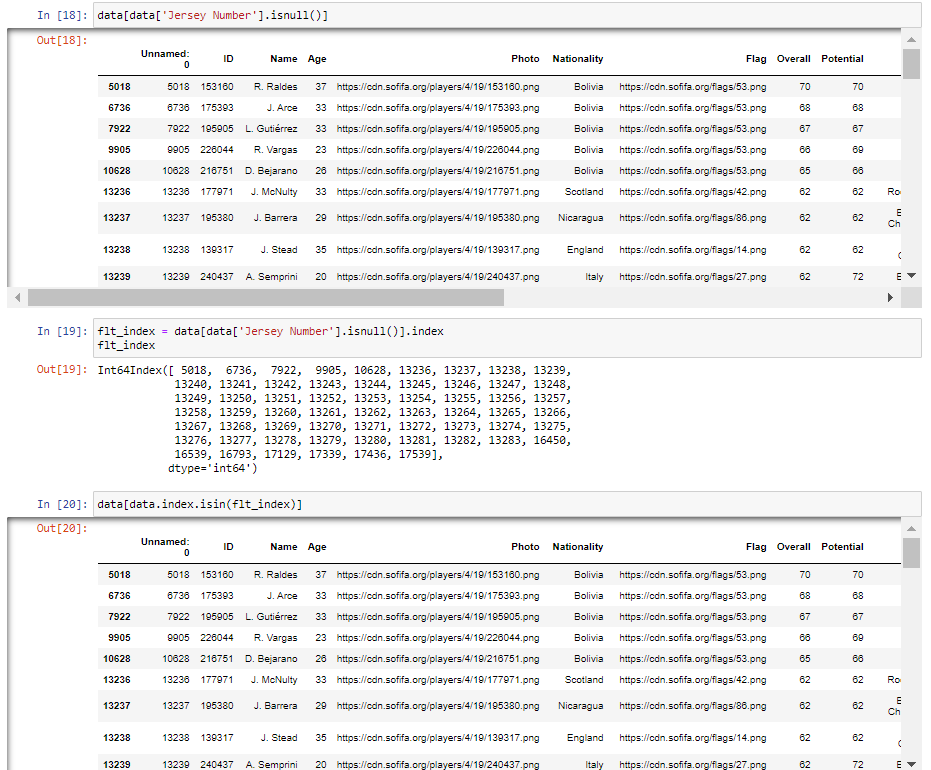
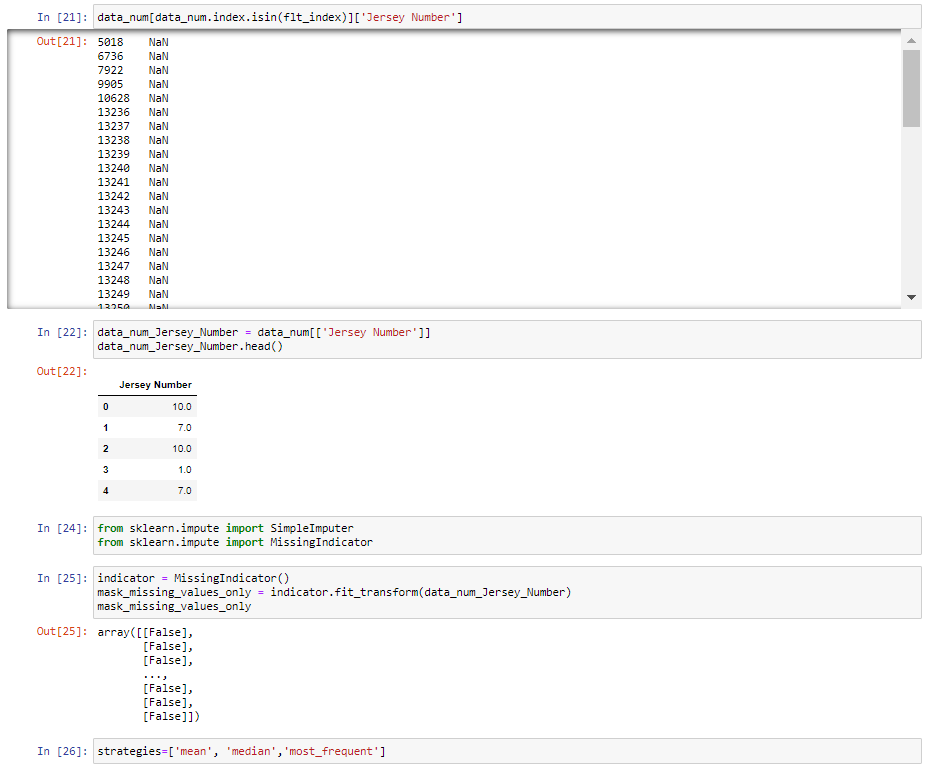
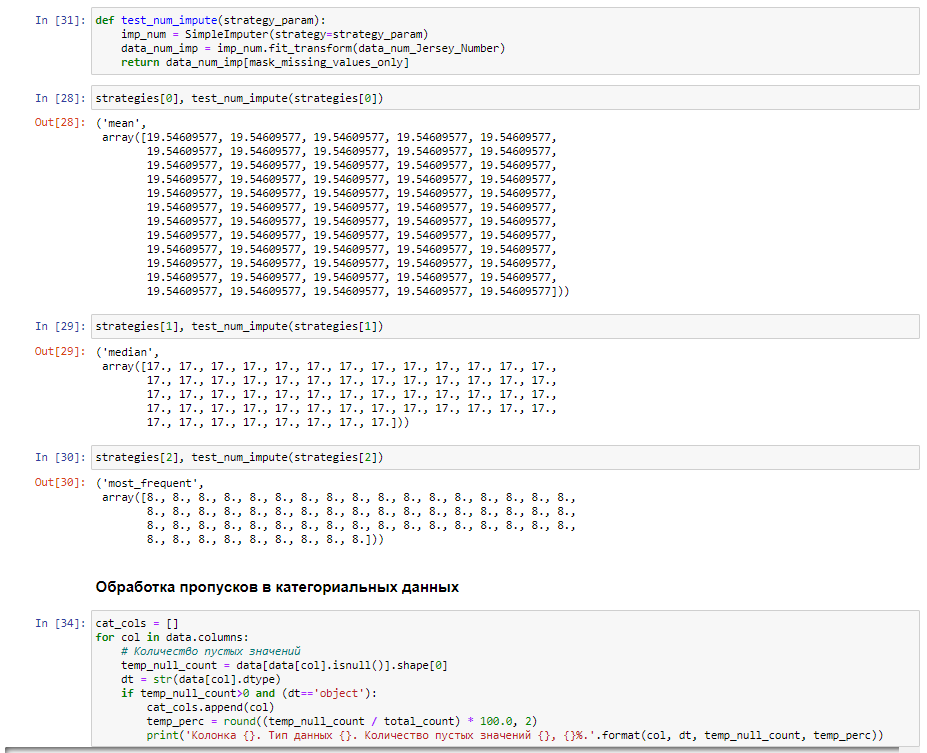
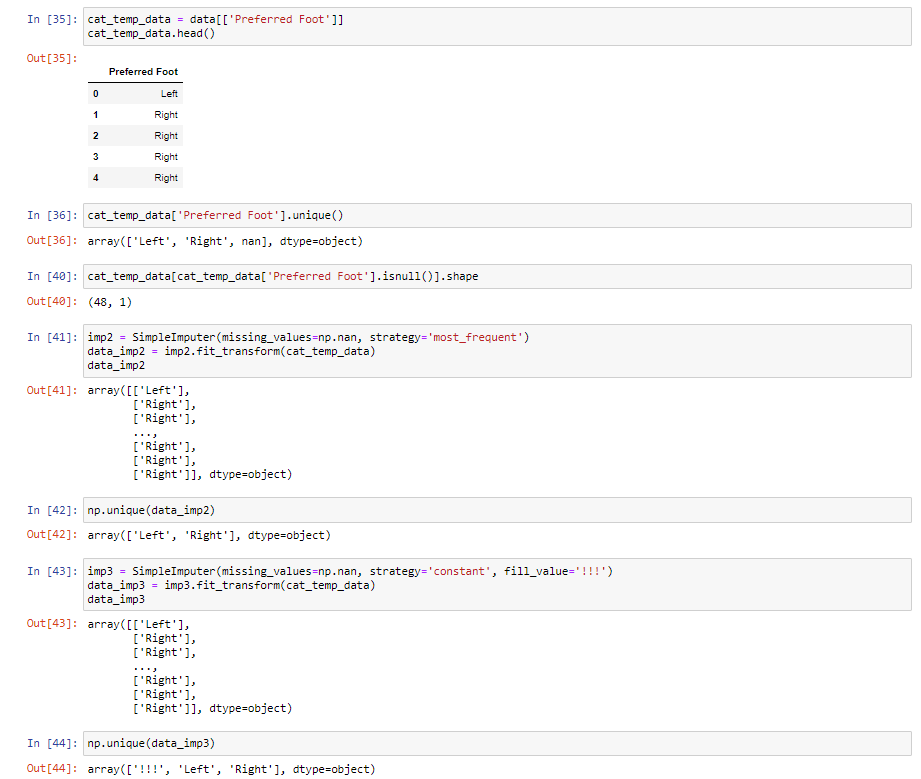
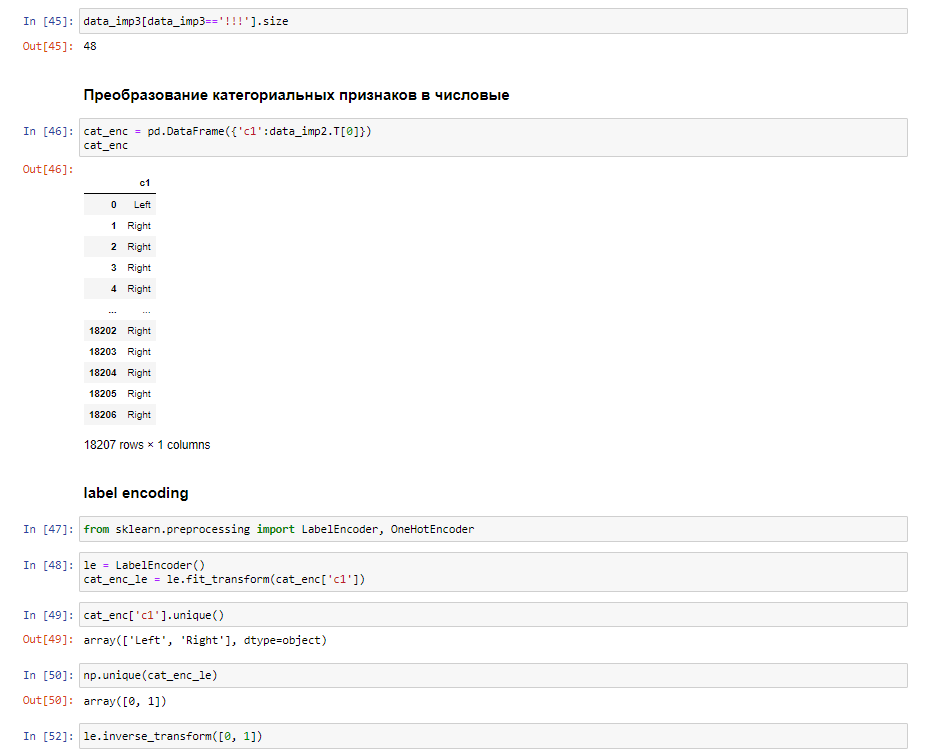
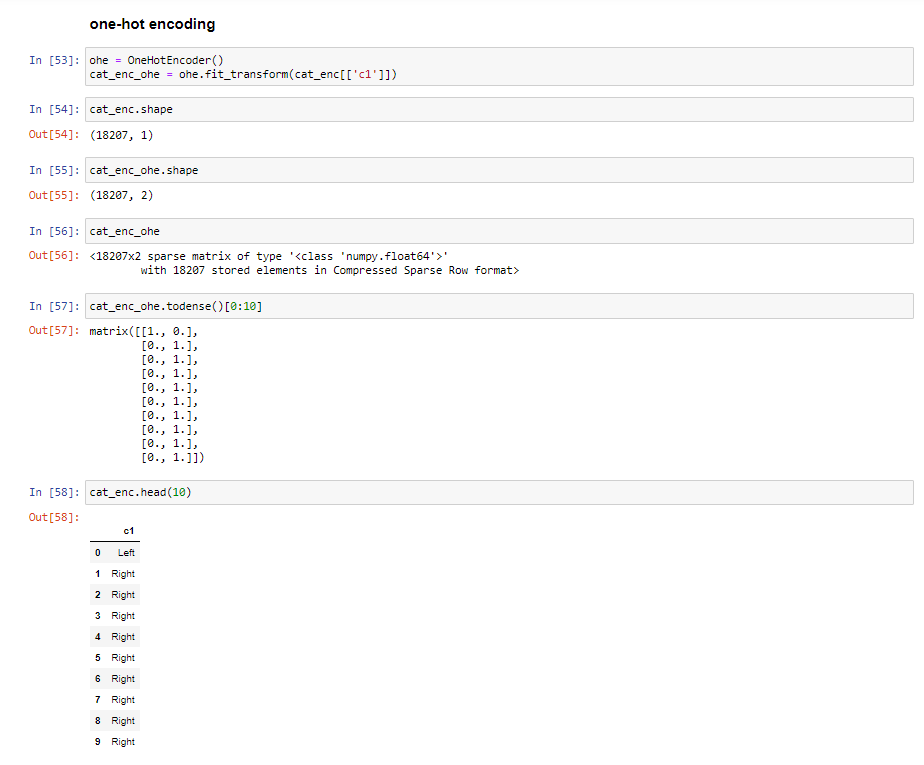
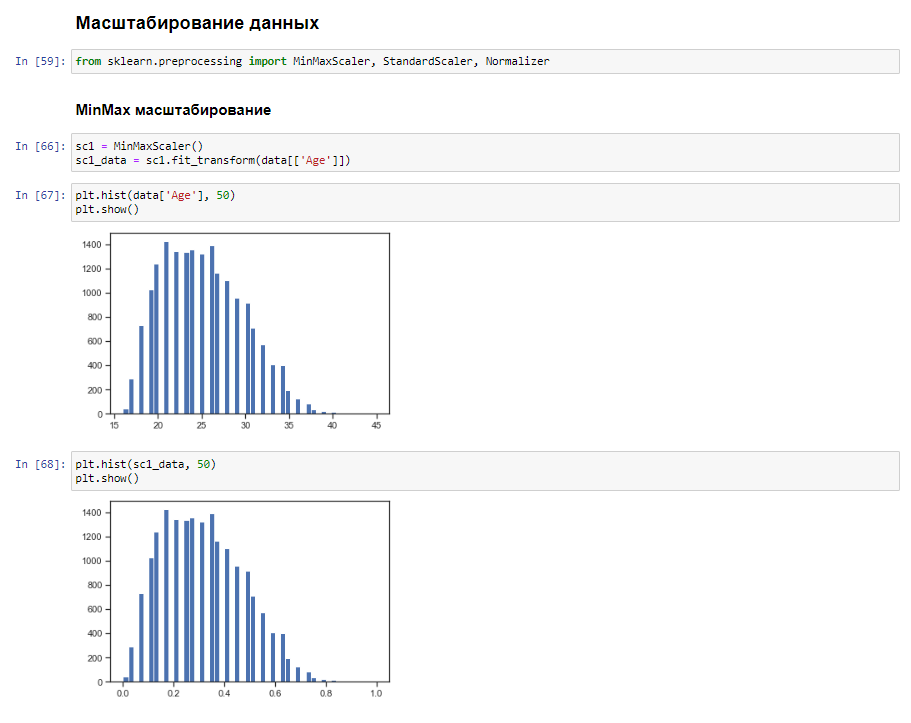
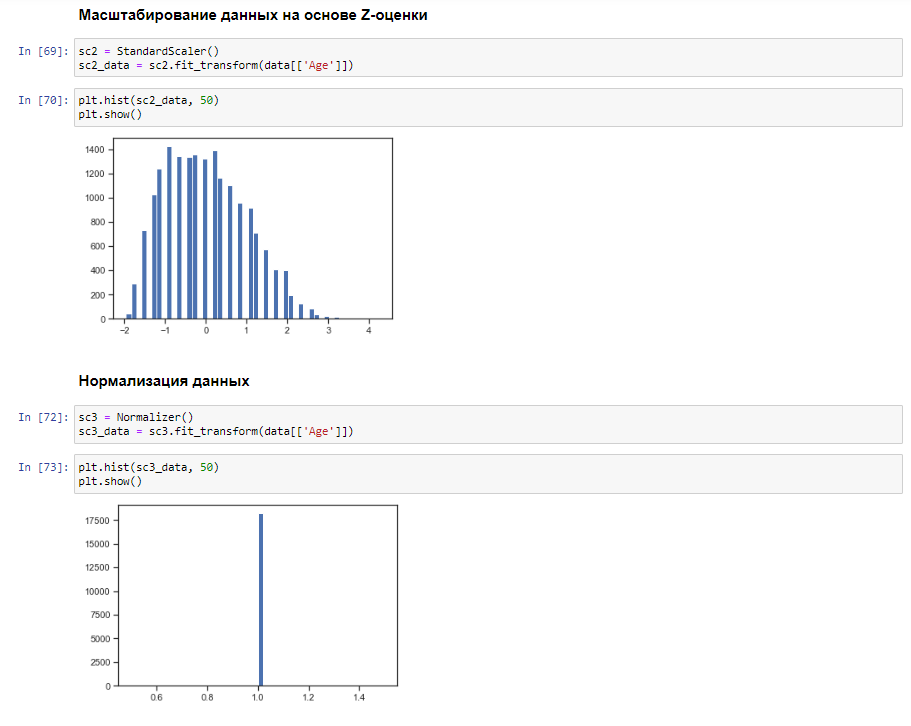
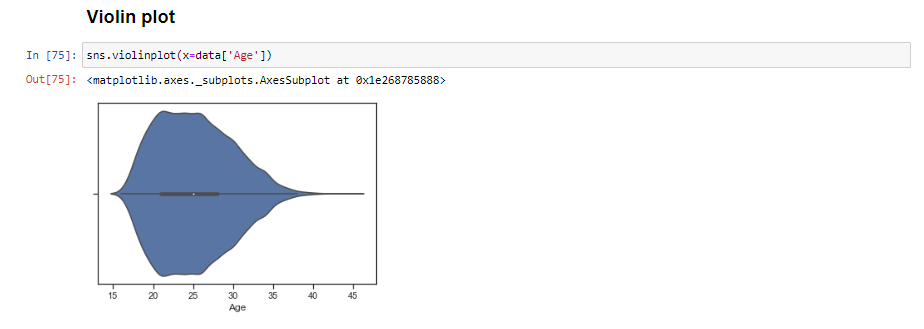
**Задание**

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:

* обработку пропусков в данных (не менее 3 признаков);
* кодирование категориальных признаков (не менее 3 признаков);
* масштабирование данных (не менее 3 признаков).

**Набор данных:**

<https://www.kaggle.com/karangadiya/fifa19>**** **** 

**** ****              Для количественных признаков использовался метод импутации. Для категориальных признаков использовались методы label encoding и one-hot encoding. Для дальнейшего построения ММО буду использовать метод Pandas get\_dummies для категориальных признаков и импьютацию для числовых, так как pandas get dummies является быстрым вариантом one-hot кодирования, а импьютация меньше влияет на данные в целом и не изменит размер датасета по сравнению с удалением или заполнением нулями. Для масштабирования были использованы MinMax масштабирование и масштабирование данных на основе Z-оценки.