Московский государственный технический университет им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Системы обработки информации и управления»

*Дисциплина «Технологии машинного обучения»*

**Отчёт**

по лабораторной работе №3

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

*Вариант 12*

Студент:

Крюков Г. М.

Группа ИУ5-61Б

Преподаватель:

Гапанюк Ю. Е.

Москва, 2020 г.

**Цель лабораторной работы:**

Изучение способов предварительной обработки данных для дальнейшего формирования моделей.

**Задание:**

* 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.)
* Для выбранного датасета (датасетов) на основе материалов лекции решить следующие задачи:
* обработку пропусков в данных;
* кодирование категориальных признаков;
* масштабирование данных.

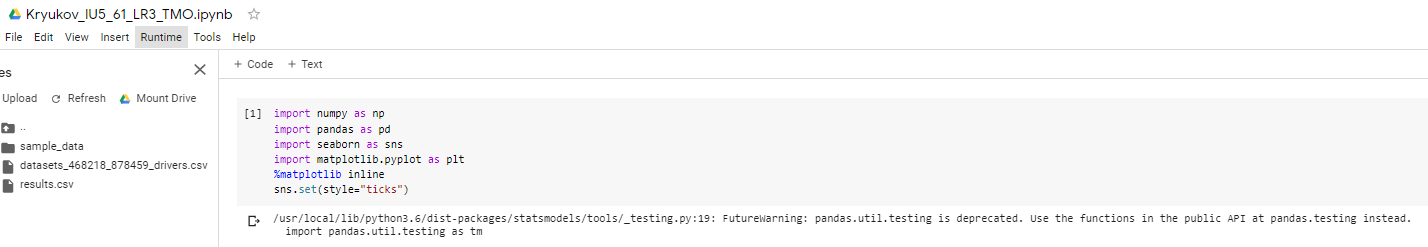
**Выполнение работы:**

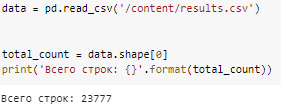
Выбранные датасеты:

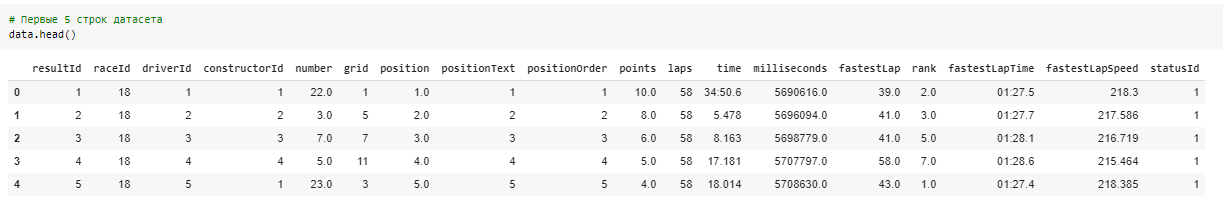
<https://www.kaggle.com/cjgdev/formula-1-race-data-19502017?select=results.csv>

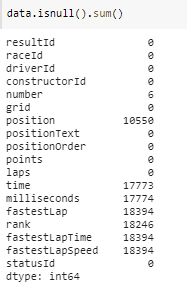
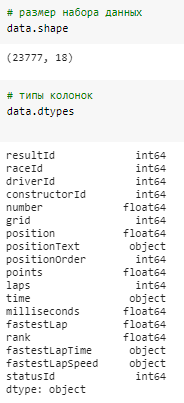
<https://www.kaggle.com/rohanrao/formula-1-world-championship-1950-2020?select=drivers.csv>

1. **Загрузка и первичный анализ данных**

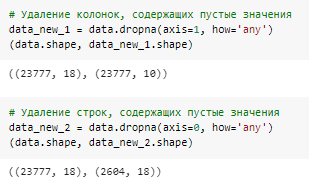


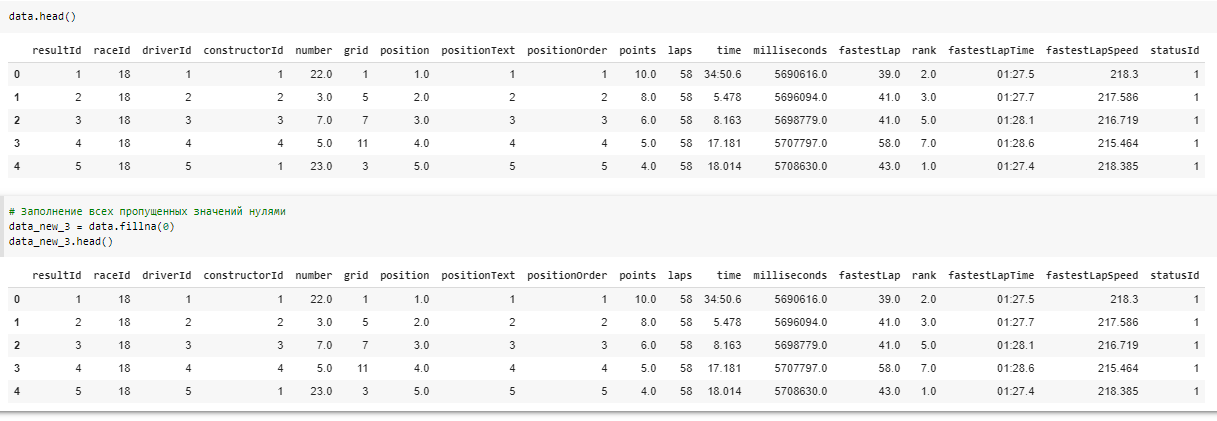




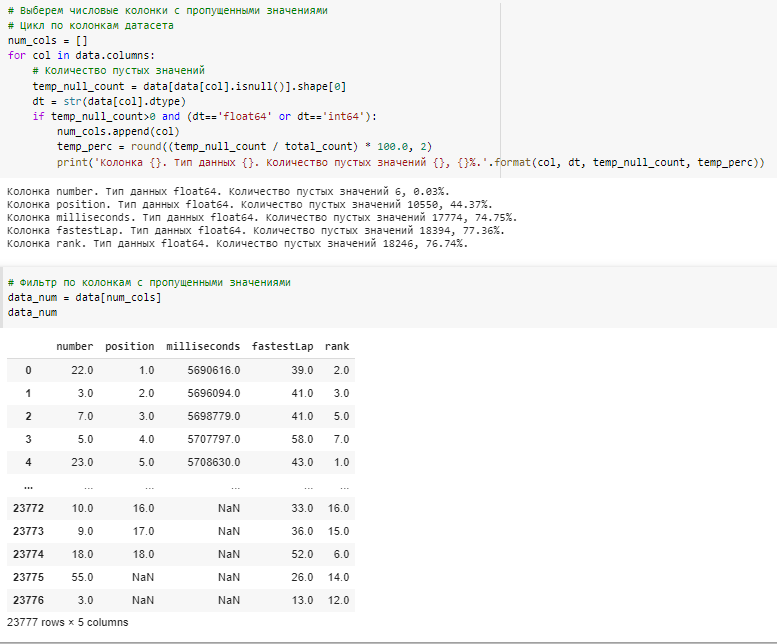


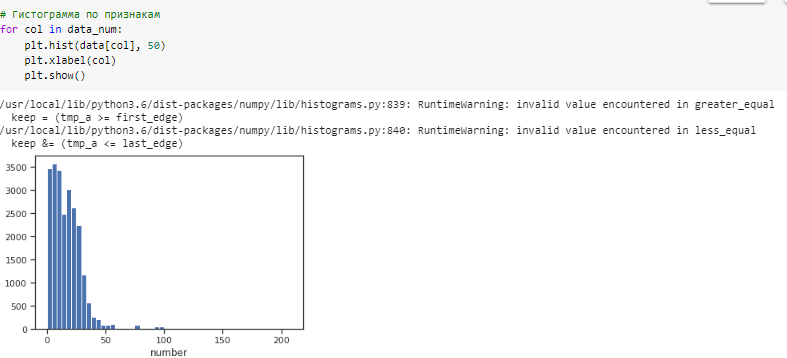
1. **Обработка пропусков в данных**
   1. Простые стратегии - удаление или заполнение нулями

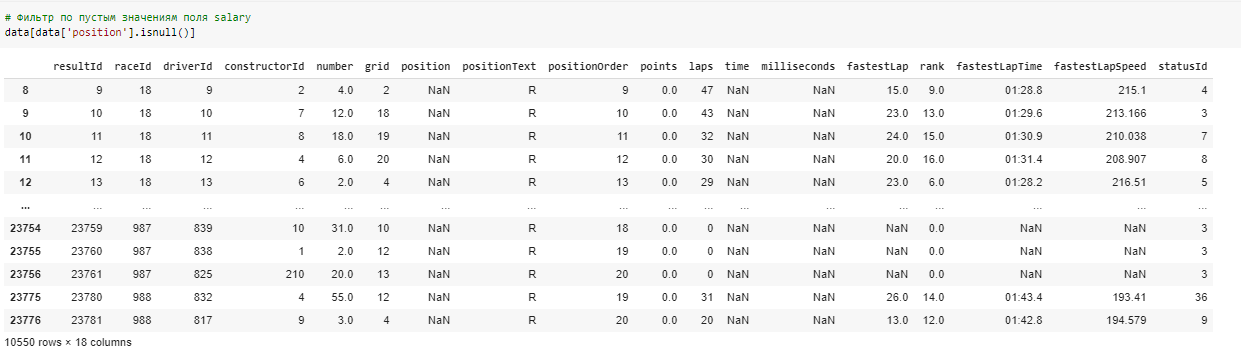


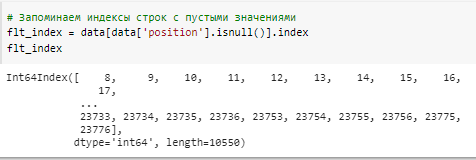


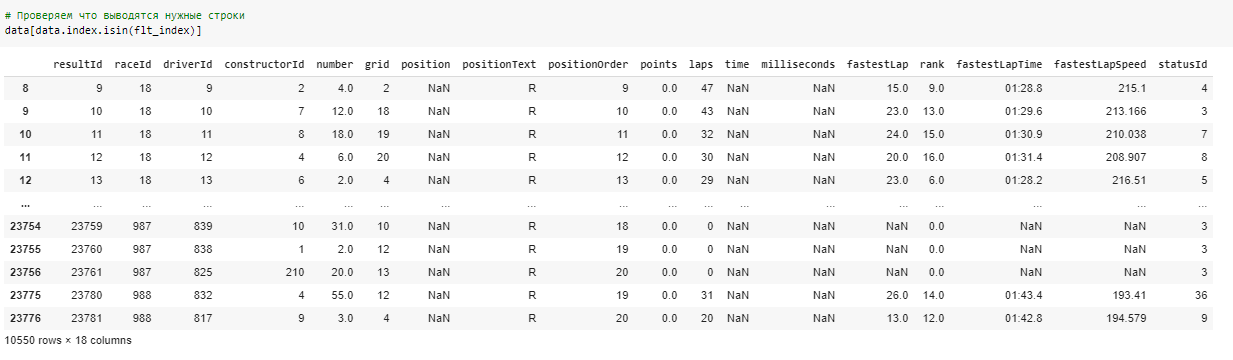
* 1. "Внедрение значений" - импьютация (imputation)

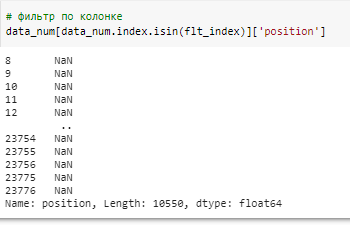


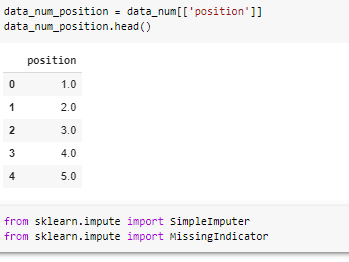


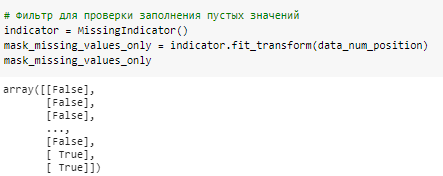


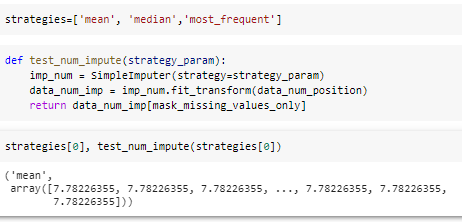


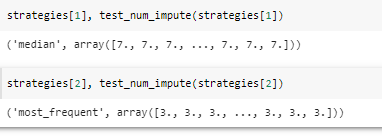


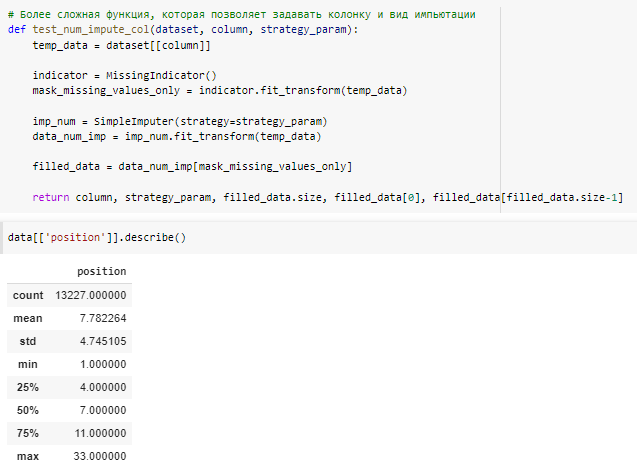


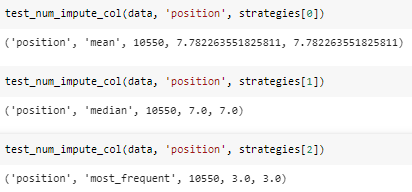




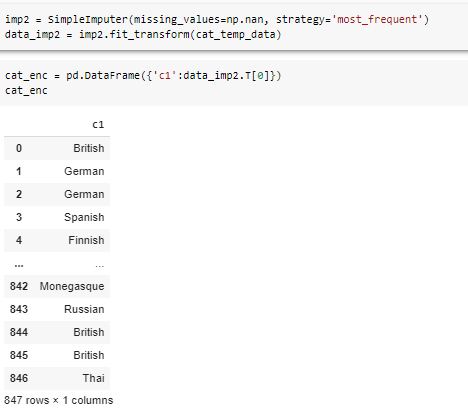
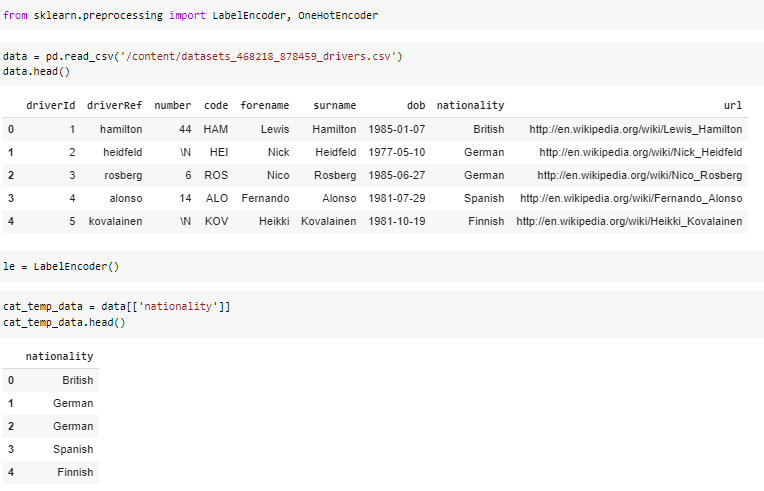




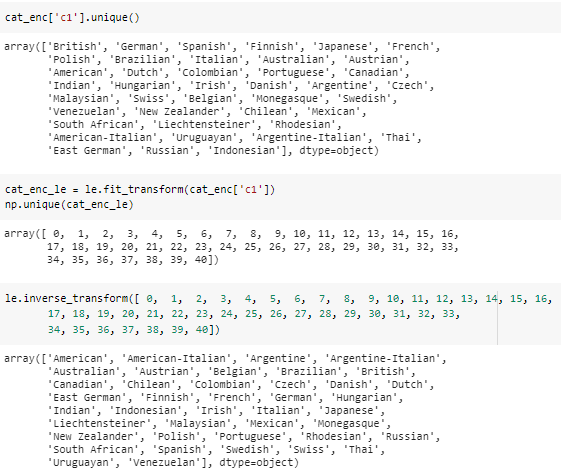




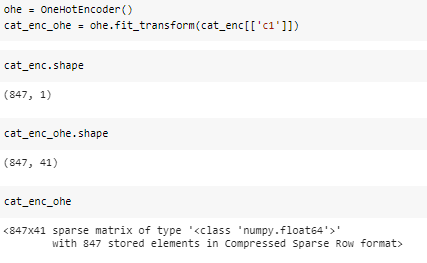
1. **Преобразование категориальных признаков в числовые**

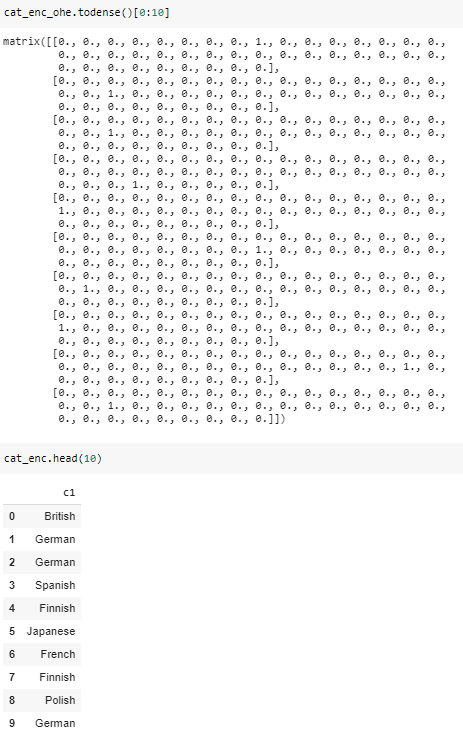


* 1. Кодирование категорий целочисленными значениями - label encoding

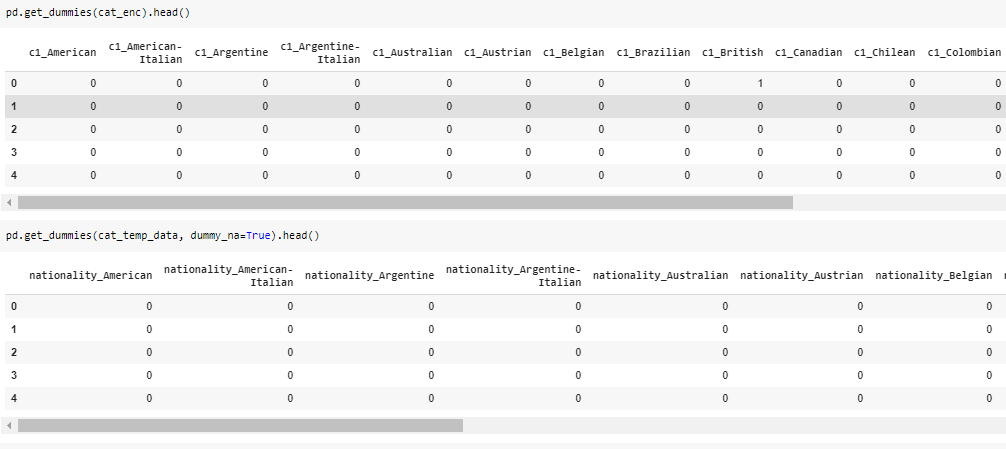


* 1. Кодирование категорий наборами бинарных значений - one-hot encoding





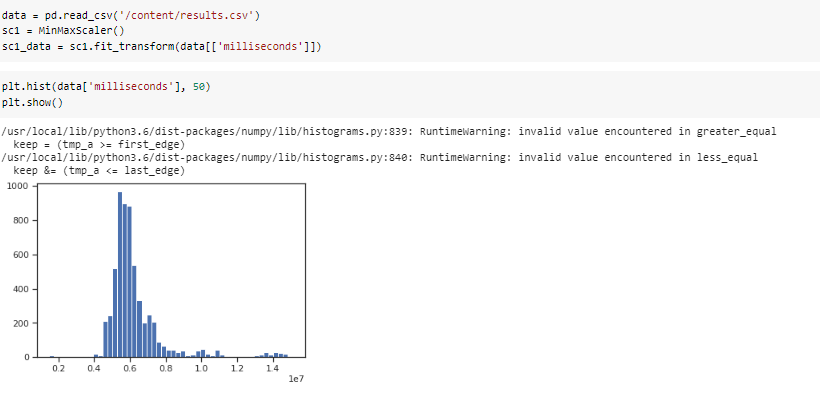
* 1. Pandas get\_dummies - быстрый вариант one-hot кодирования

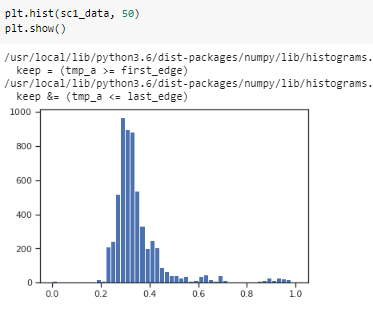


1. **Масштабирование данных**

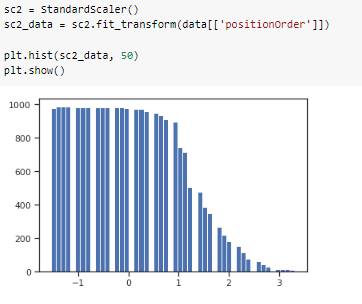
****

* 1. MinMax масштабирование





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



* 1. Нормализация данных

