**Московский государственный технический университет им. Н.Э. Баумана Кафедра «Системы обработки информации и управления»**

Лабораторная работа №3

по дисциплине

«Методы машинного обучения»

на тему

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

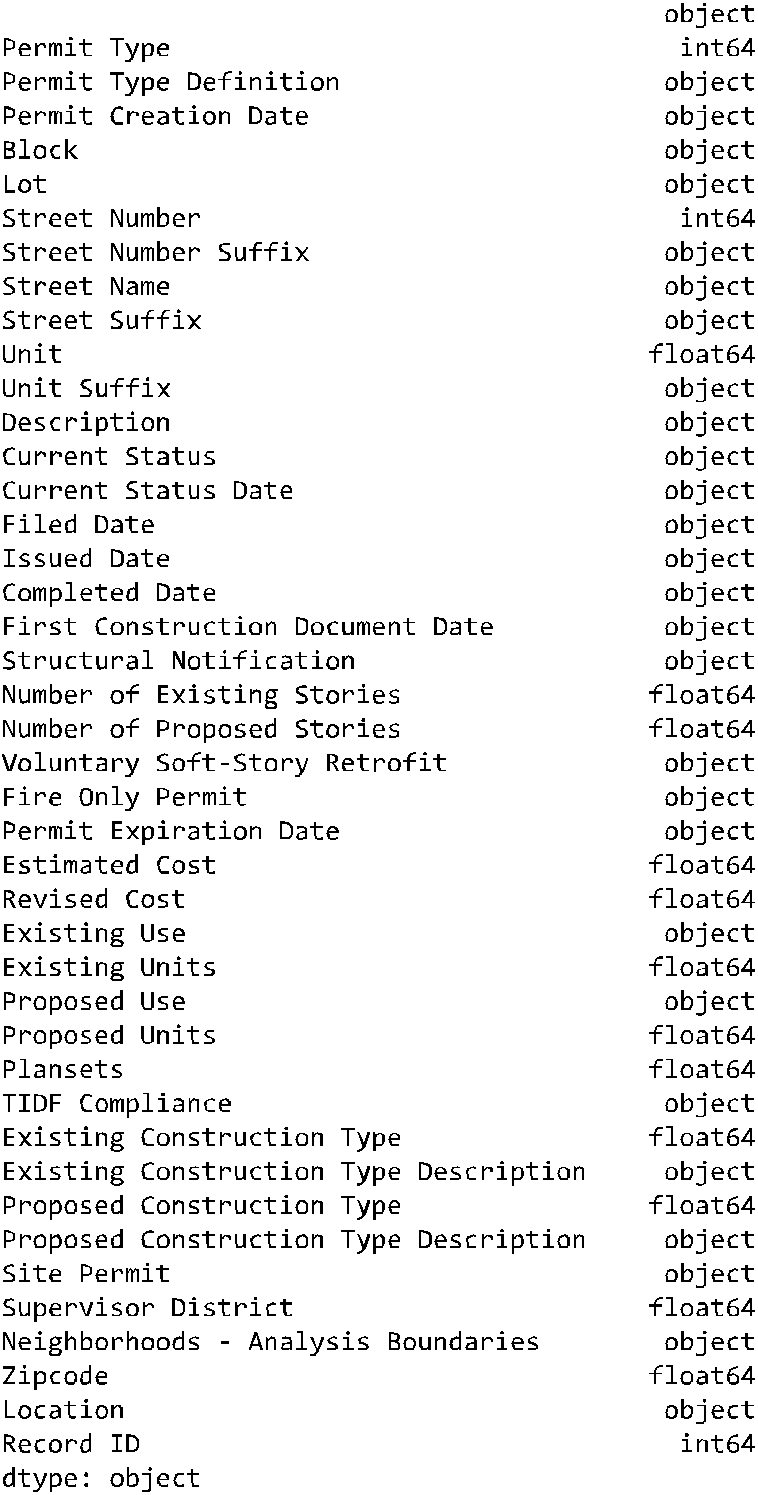
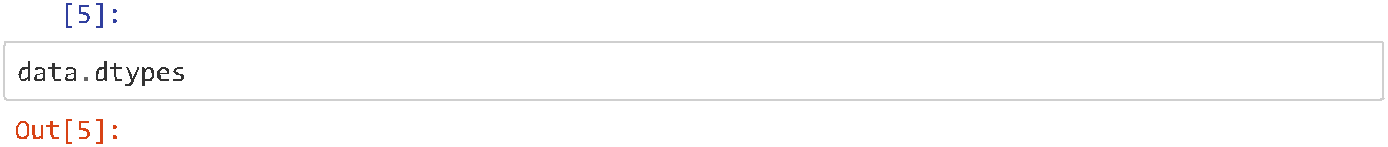
Выполнил:

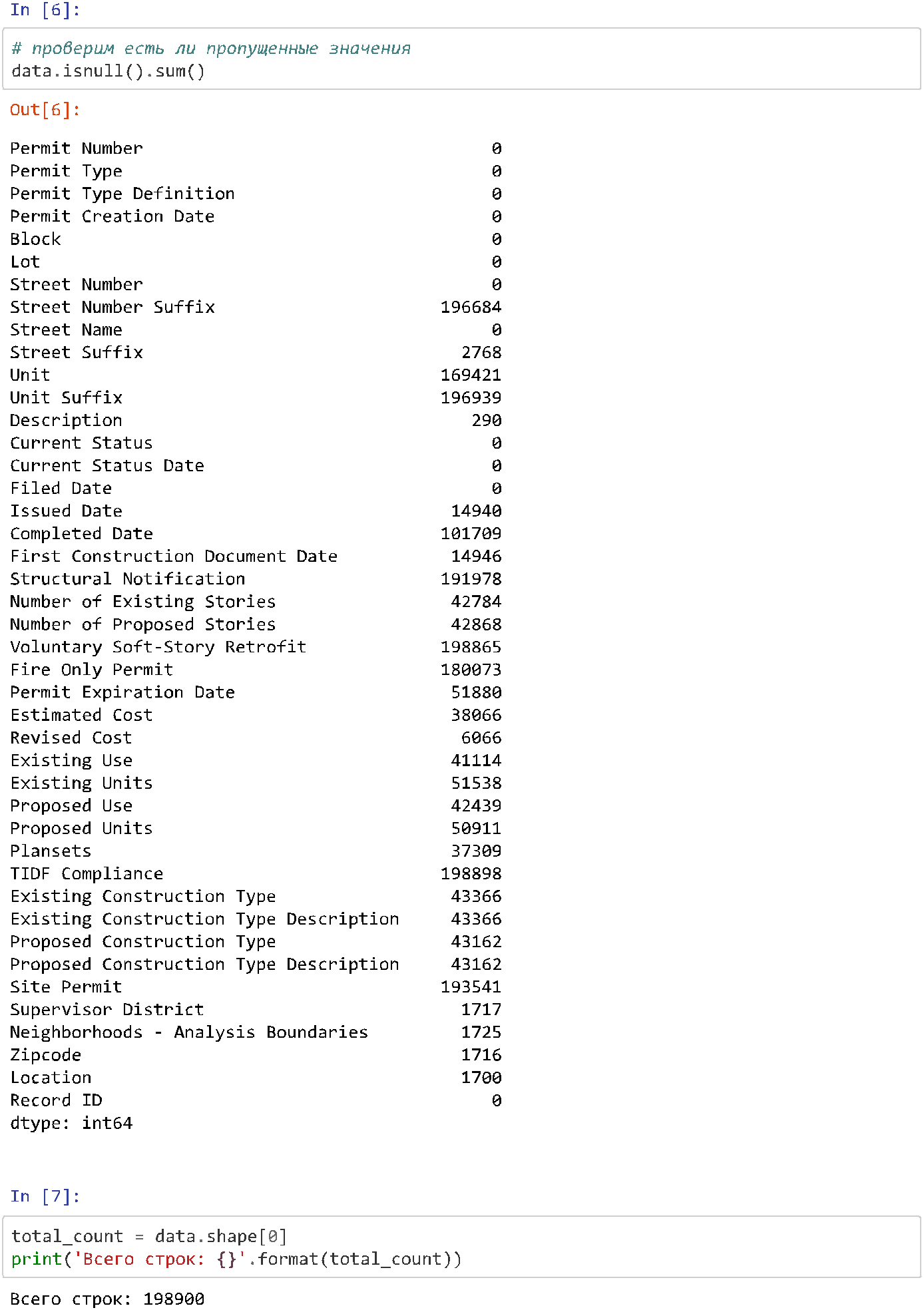
студент группы ИУ5-21М

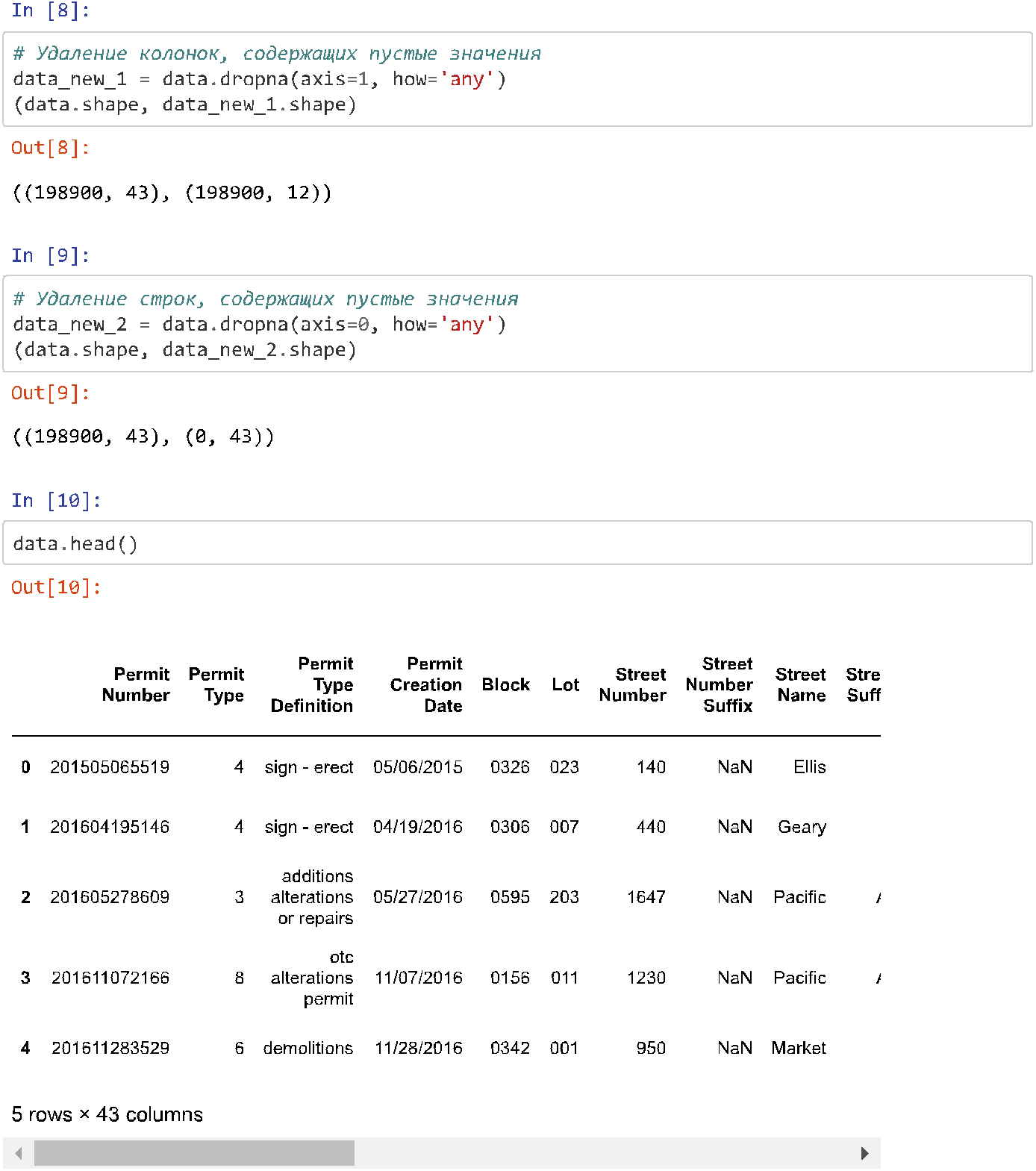
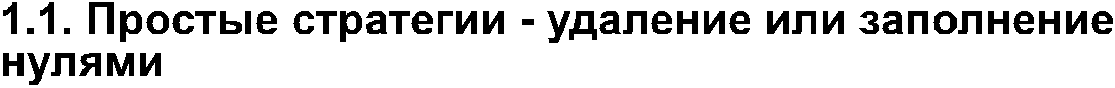
Никитин К.И.

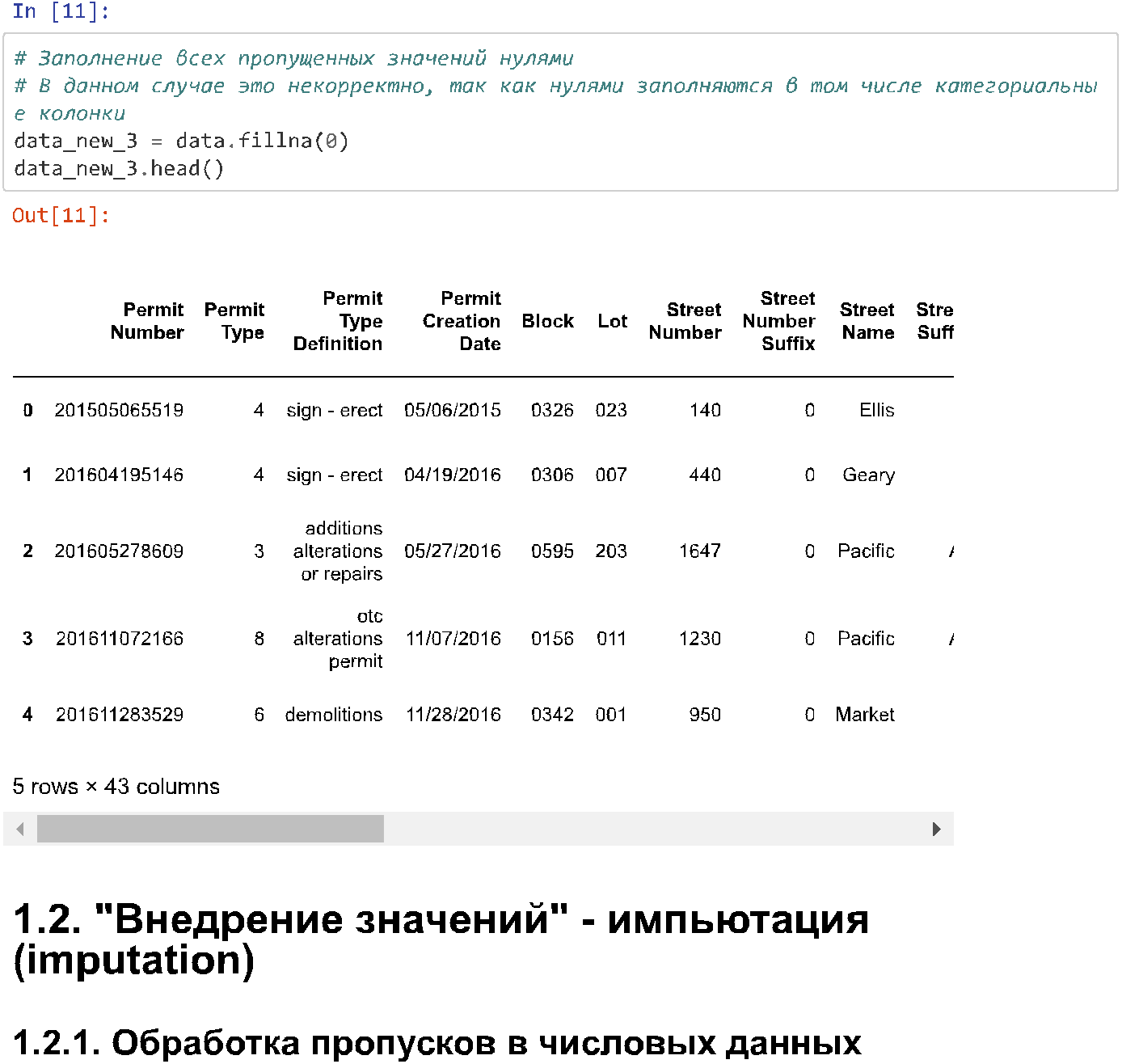
Москва — 2020 г.





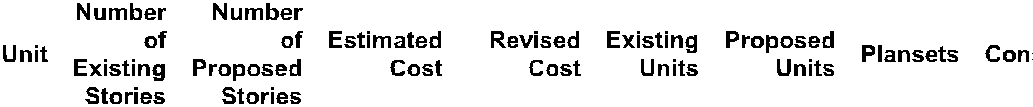


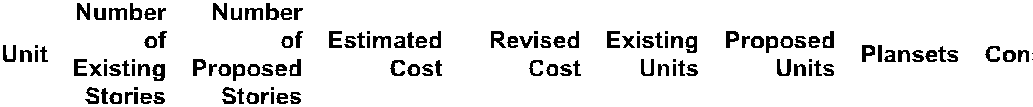


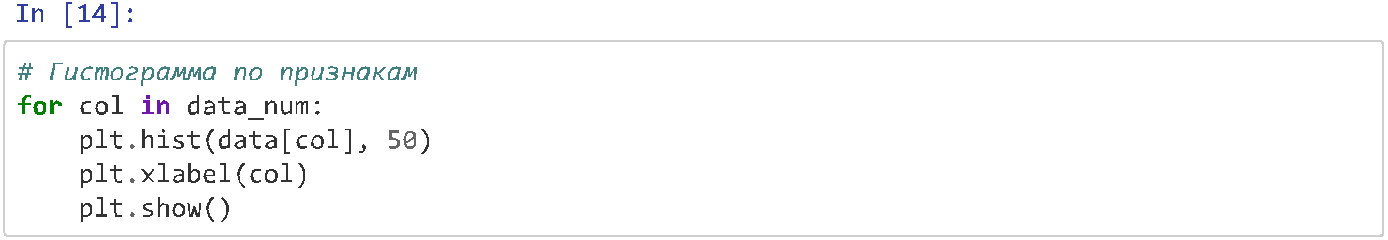


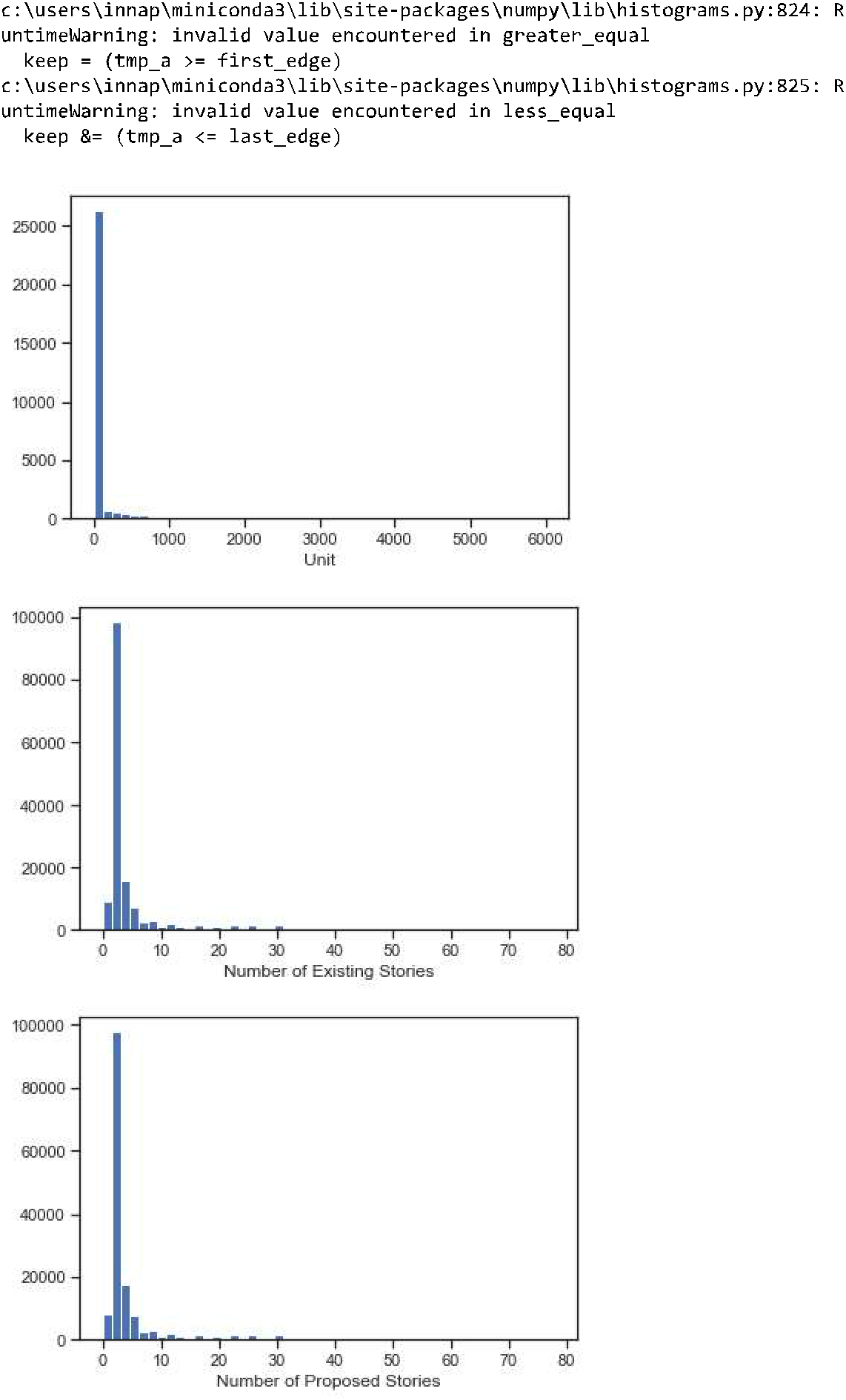


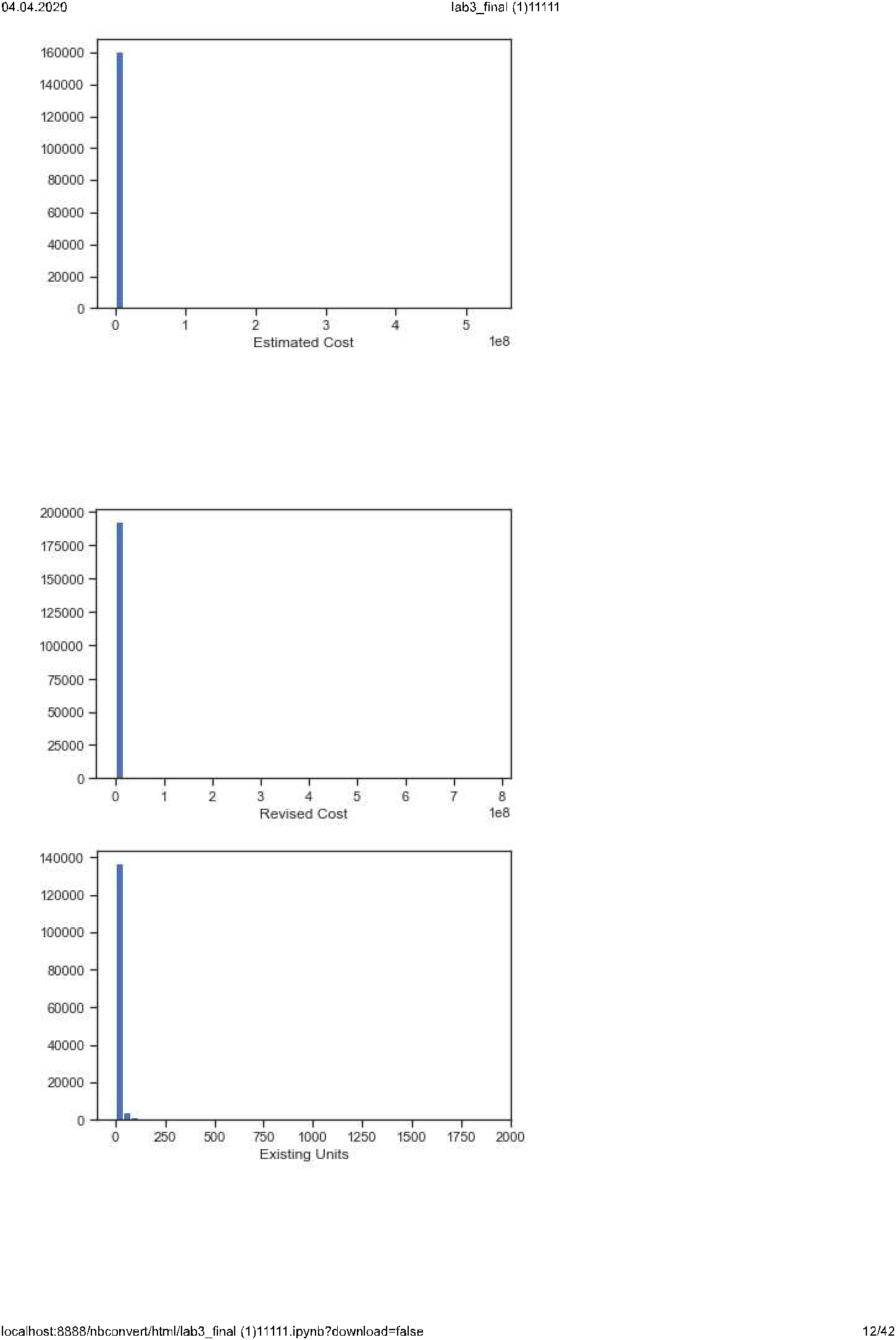


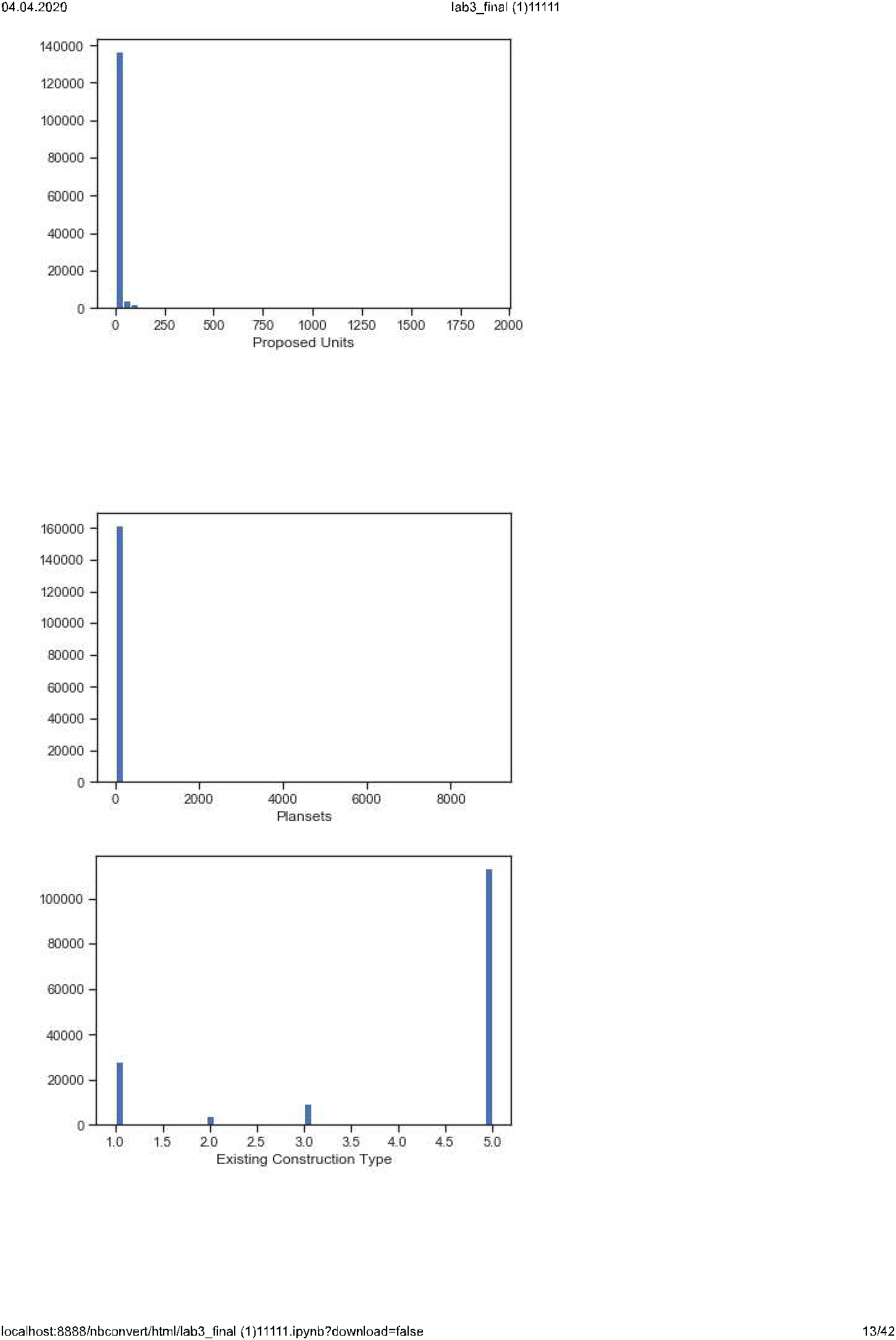


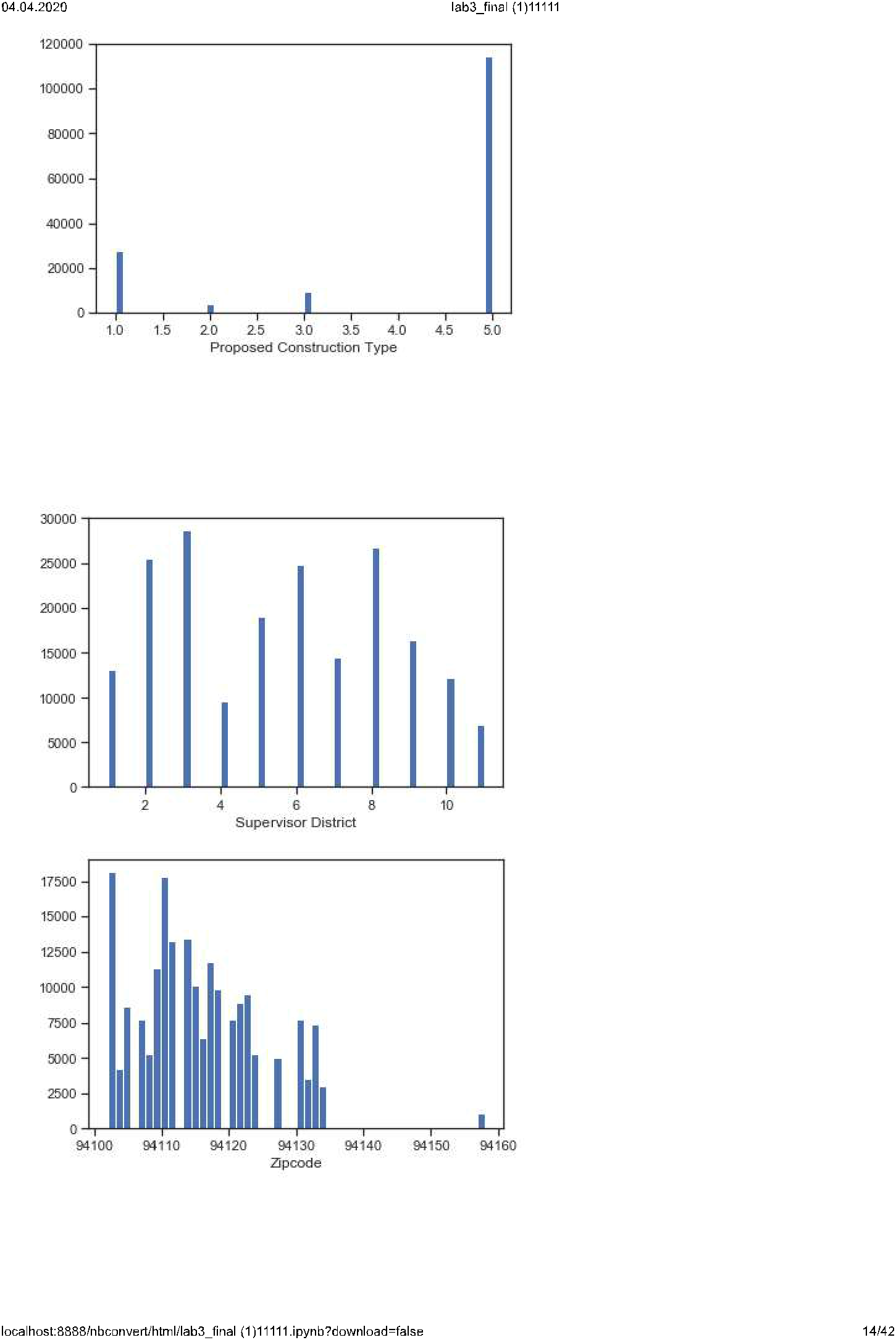




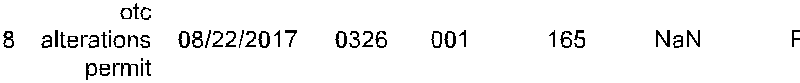
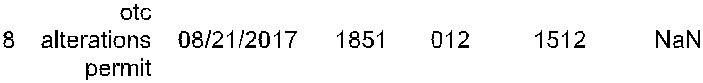
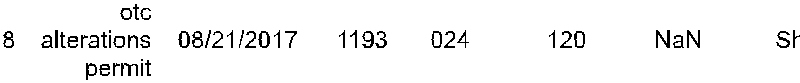
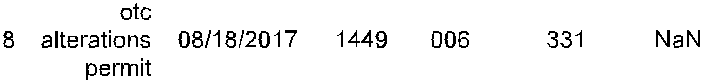
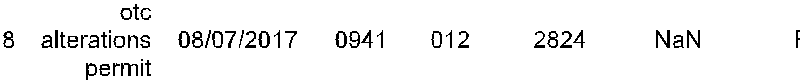
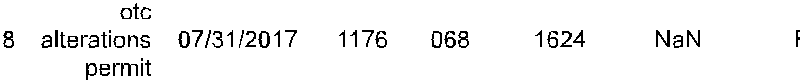
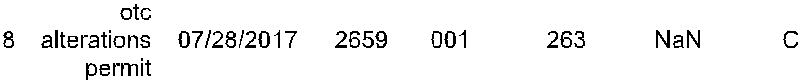
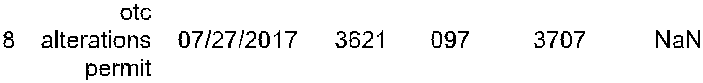
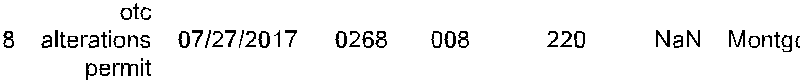
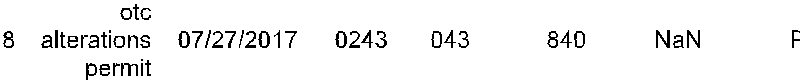
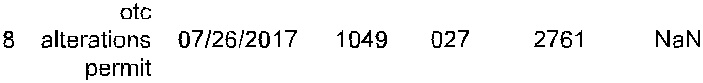
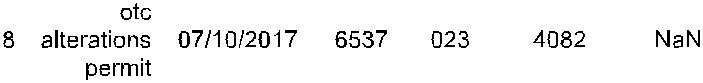
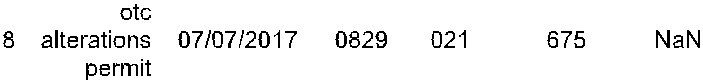
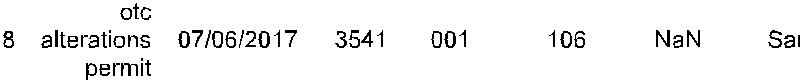
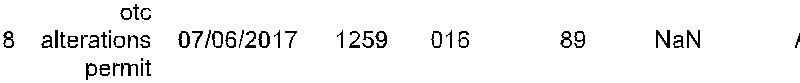
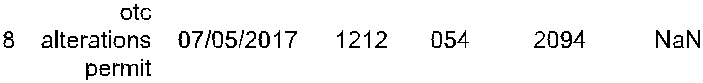
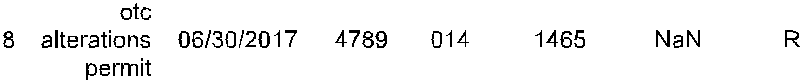
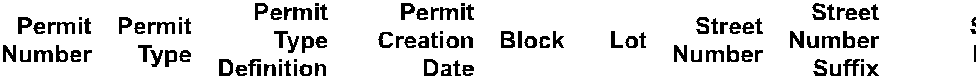


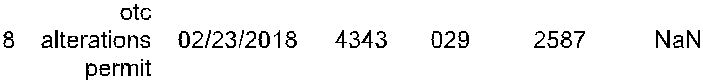
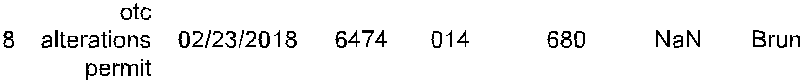
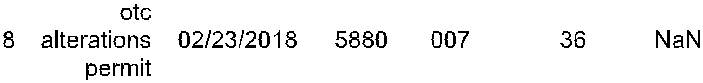
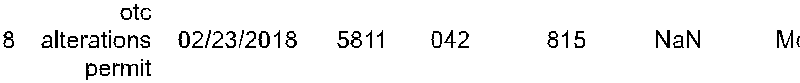
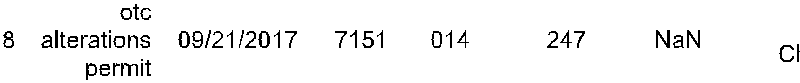
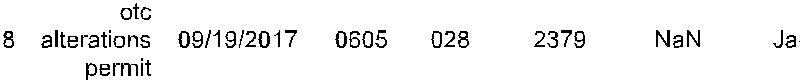
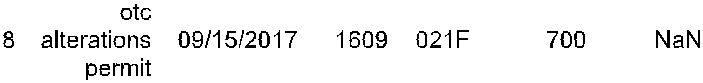
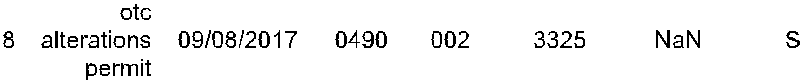
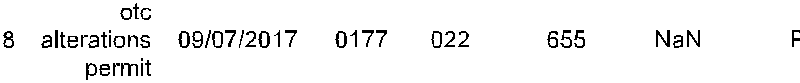
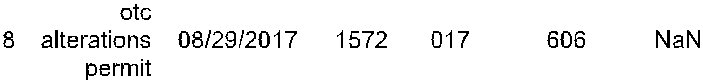
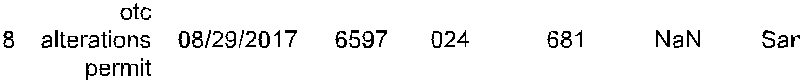
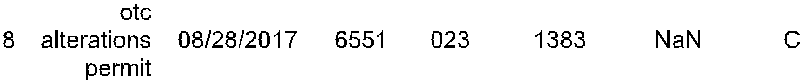
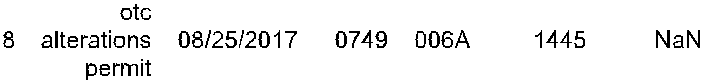
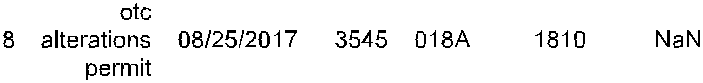
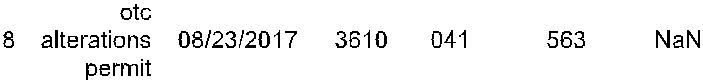
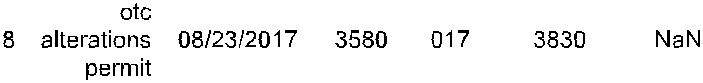
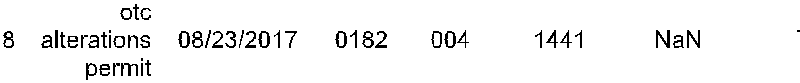
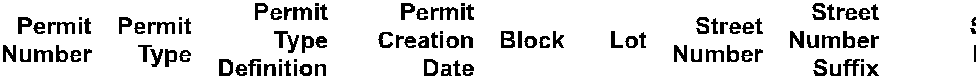


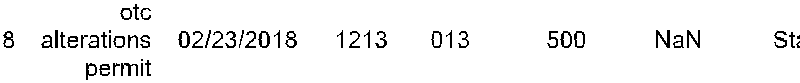
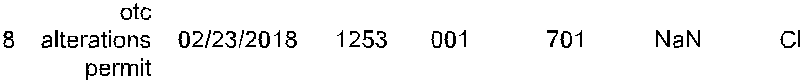
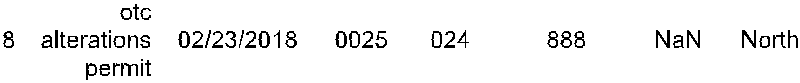
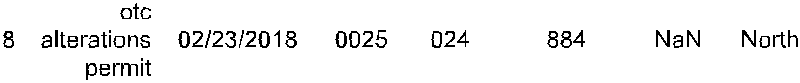
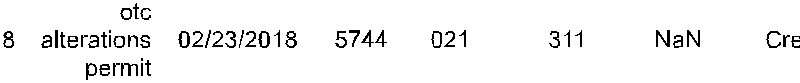
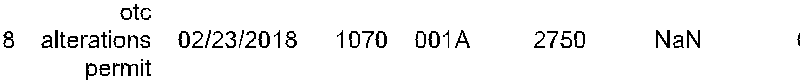
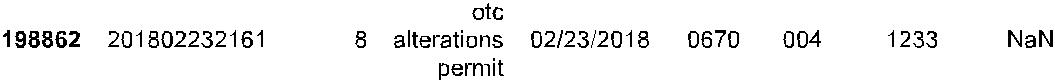
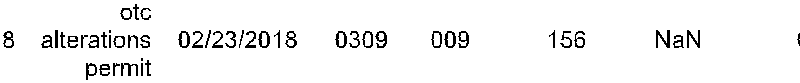
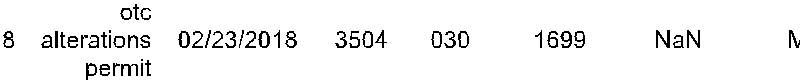
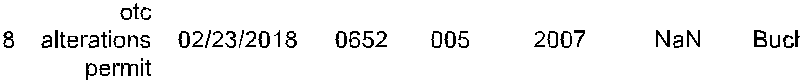
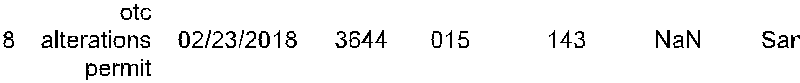
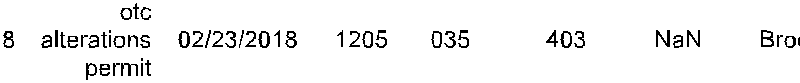
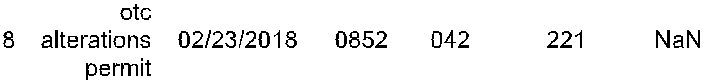
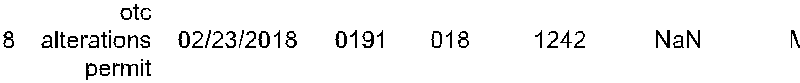
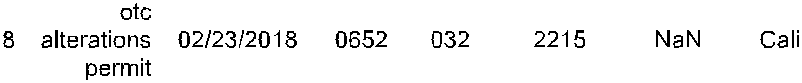
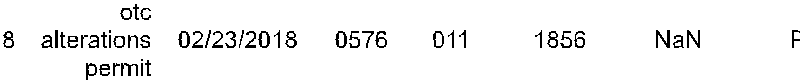
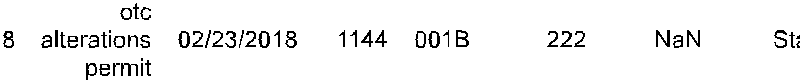
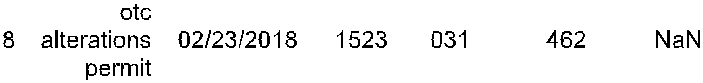
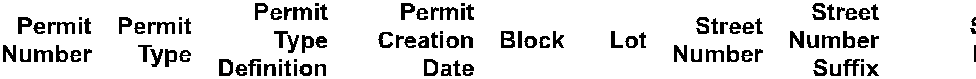


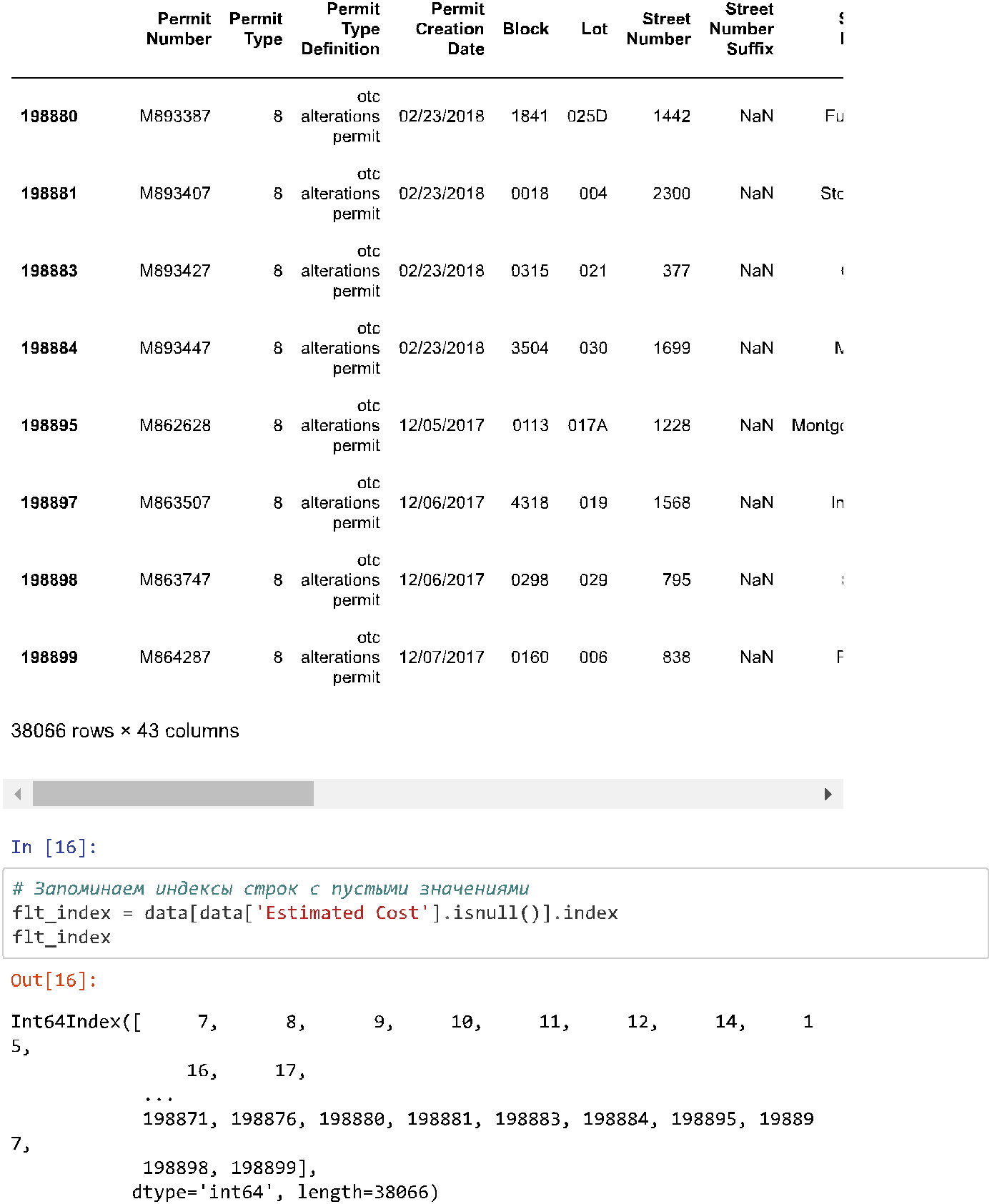


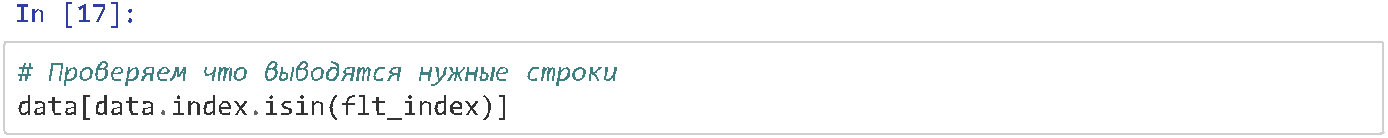


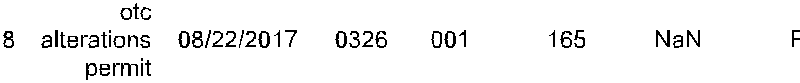
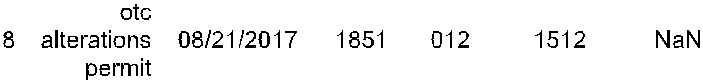
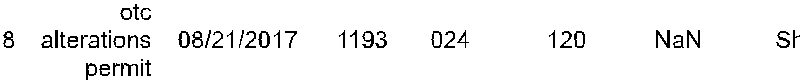
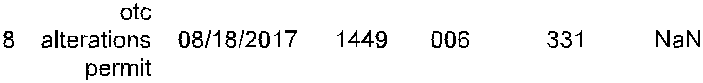
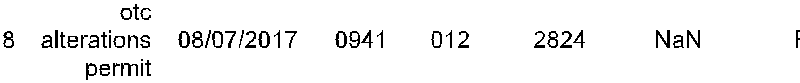
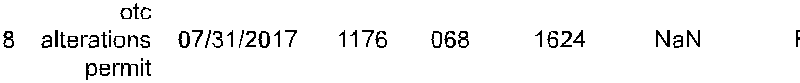
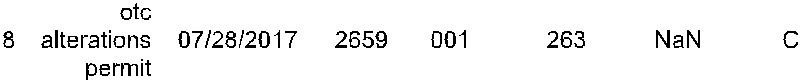
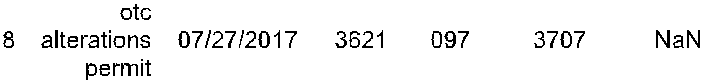
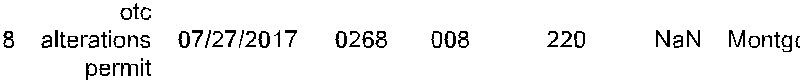
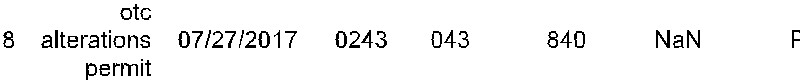
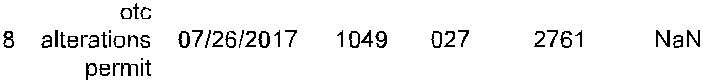
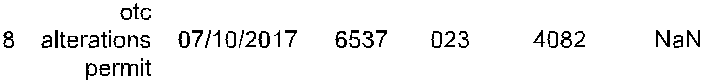
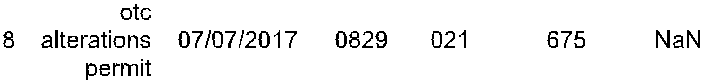
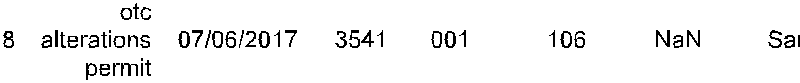
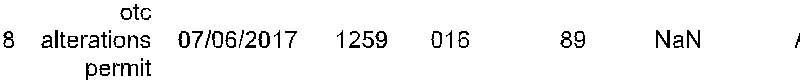
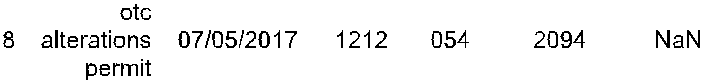
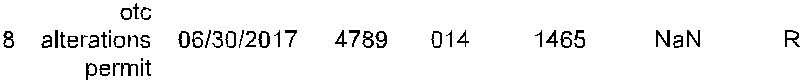
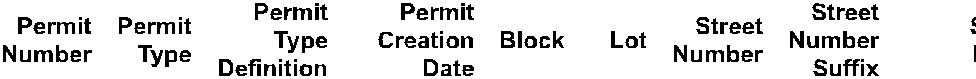


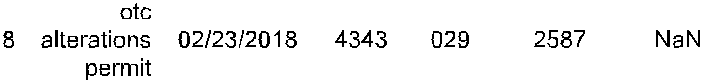
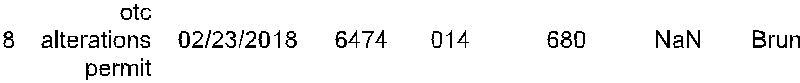
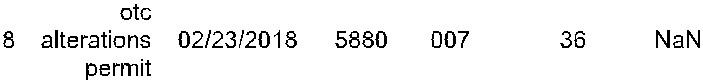
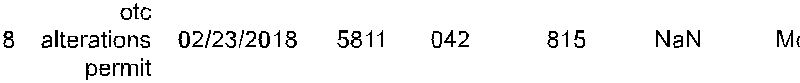
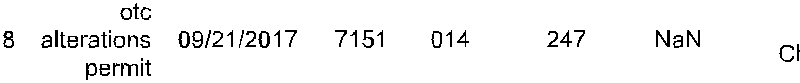
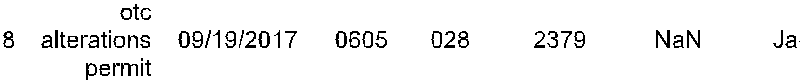
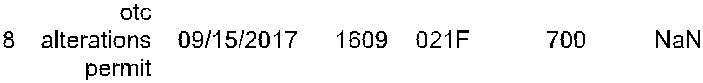
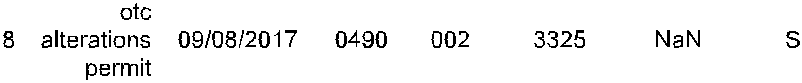
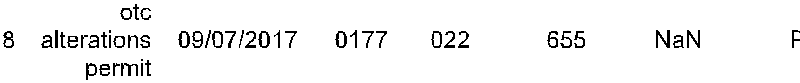
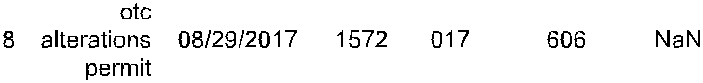
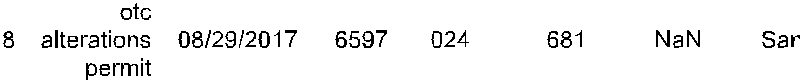
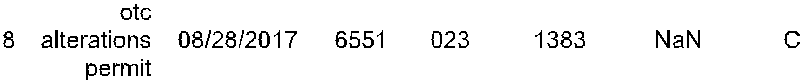
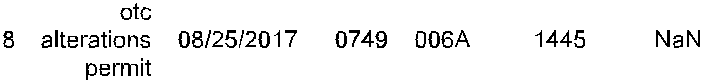
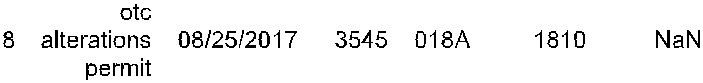
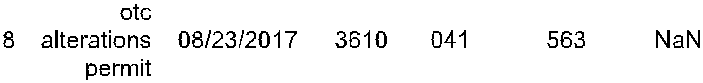
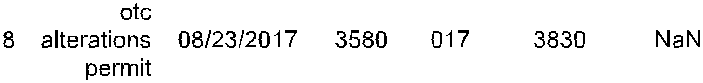
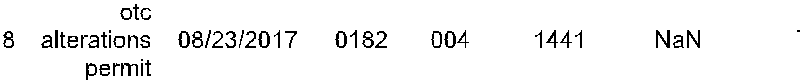
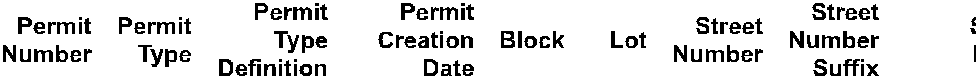


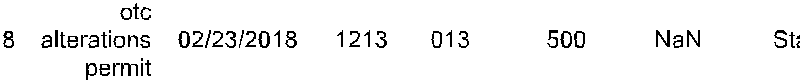
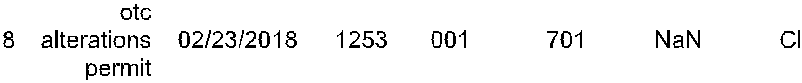
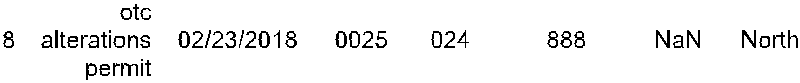
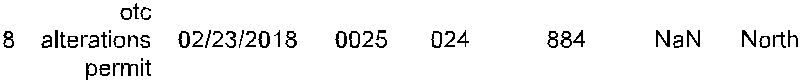
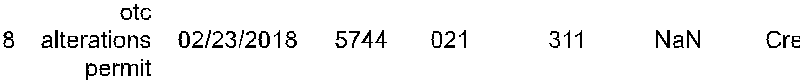
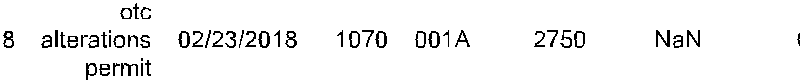
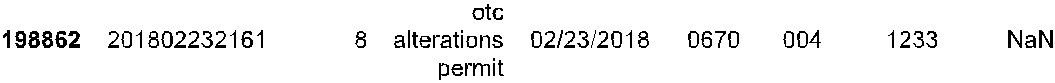
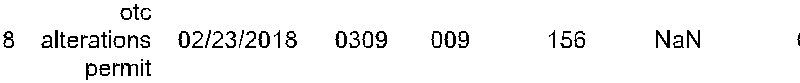
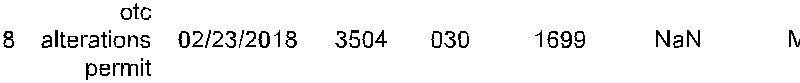
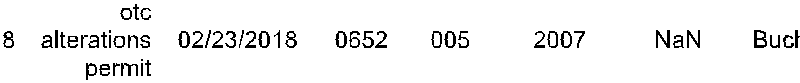
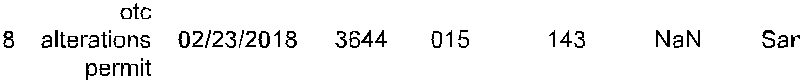
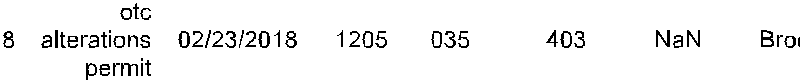
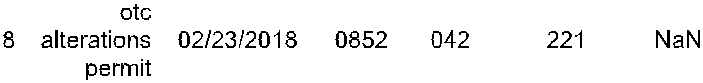
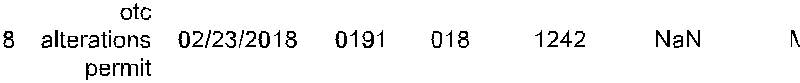
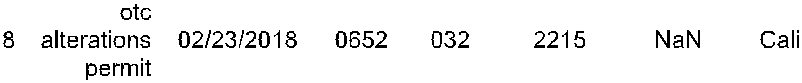
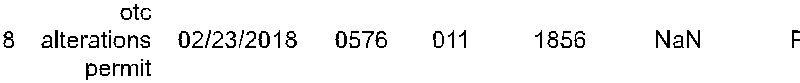
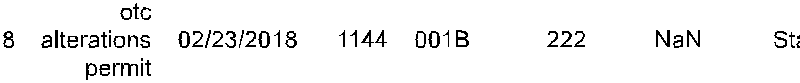
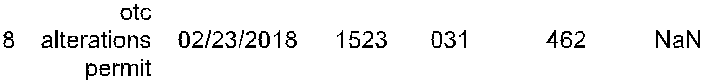
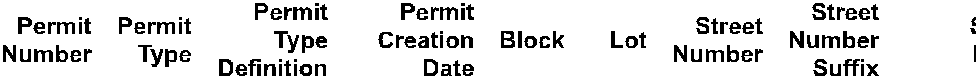


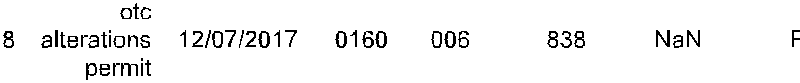
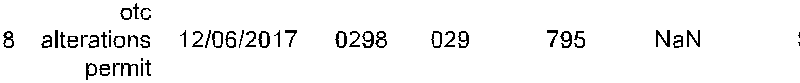
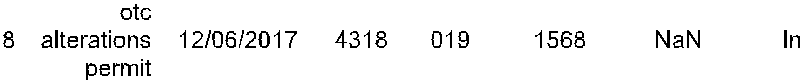
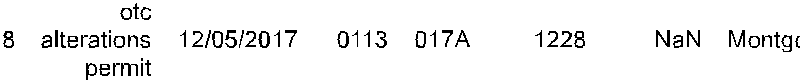
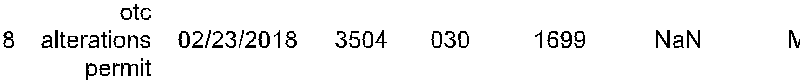
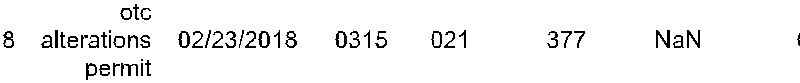
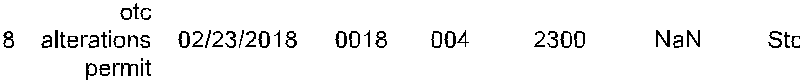
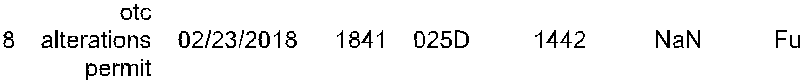
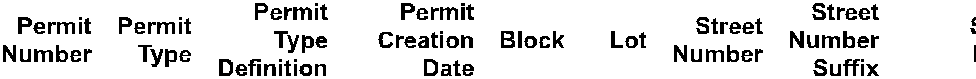






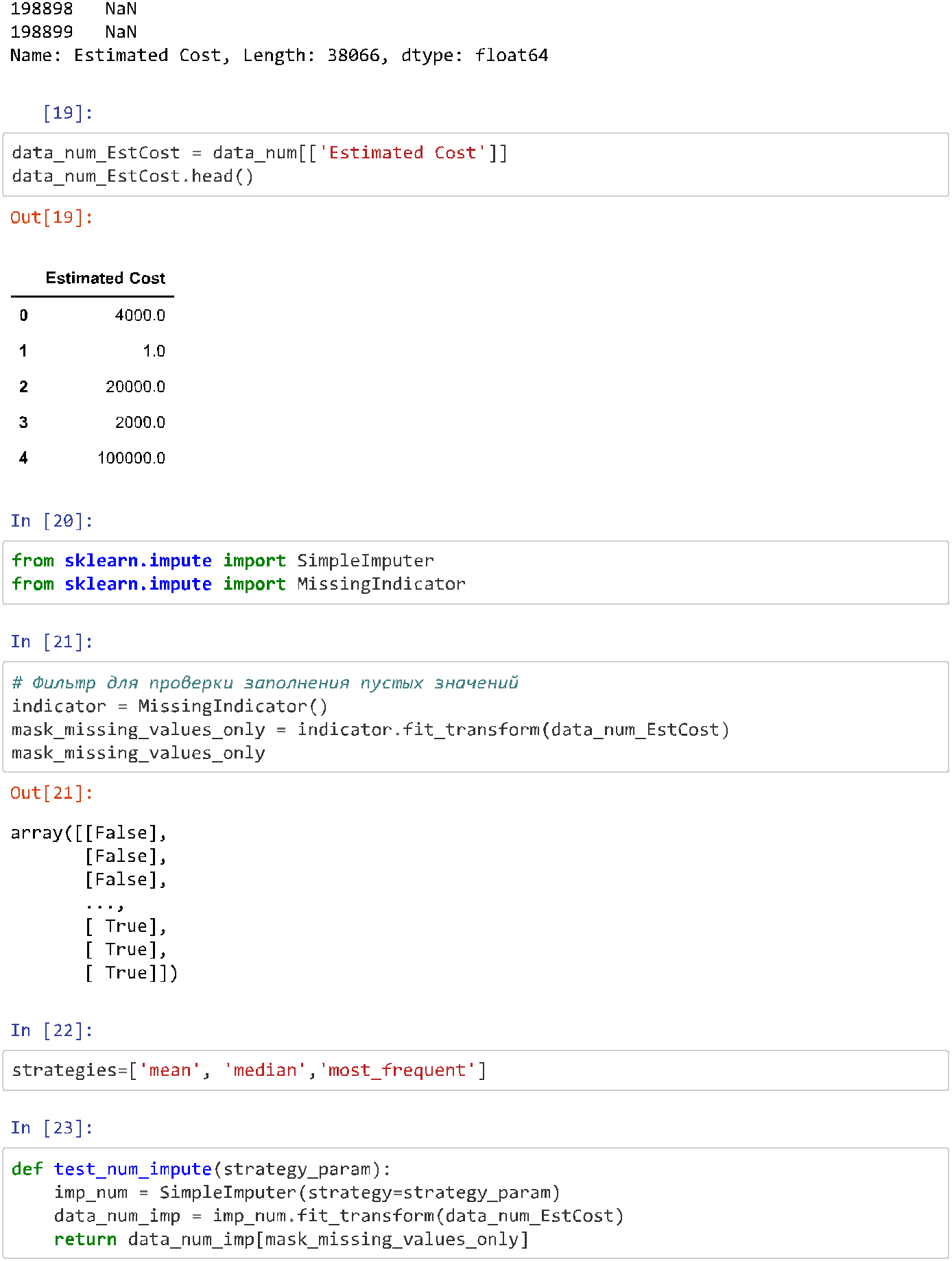


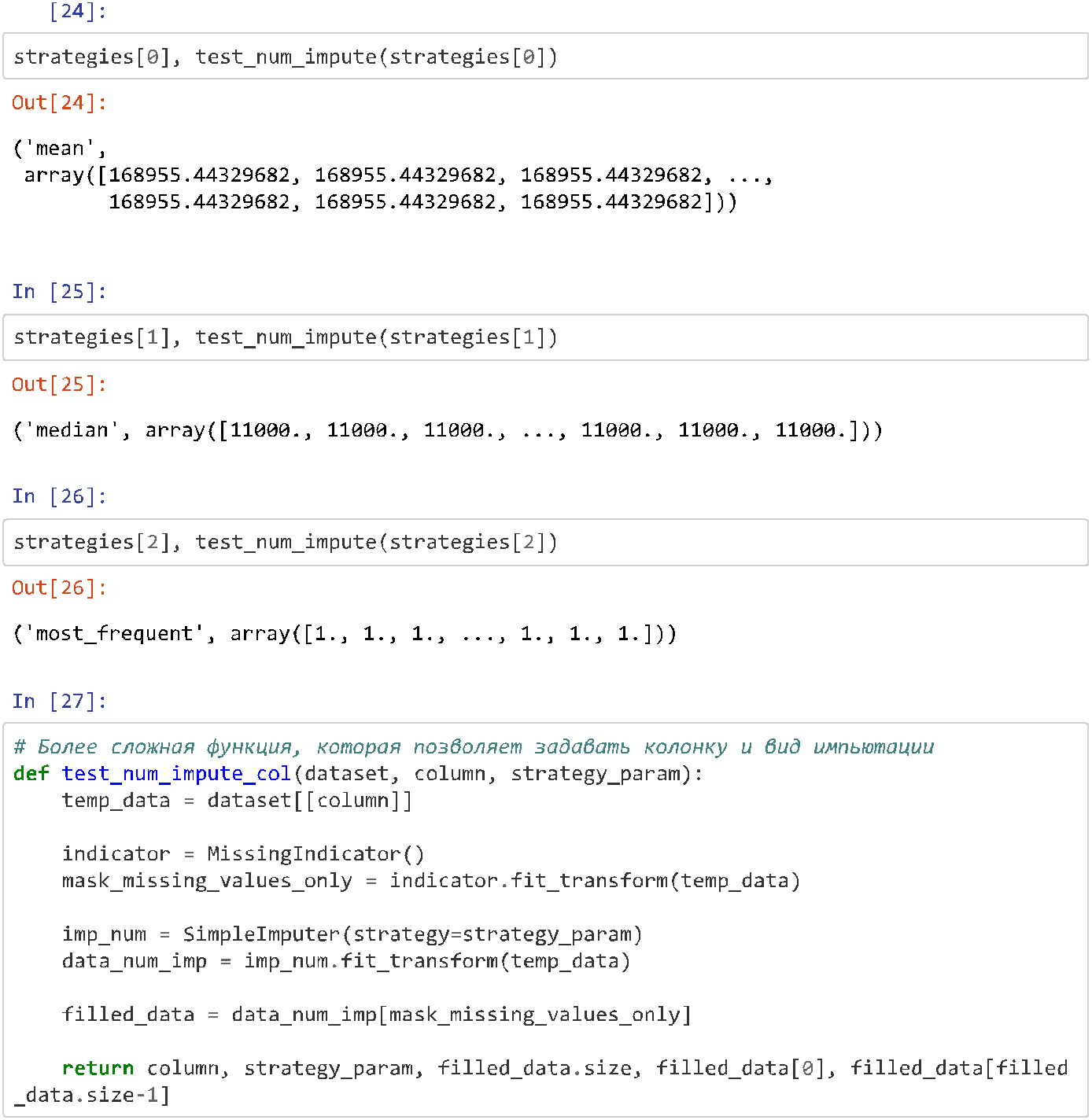


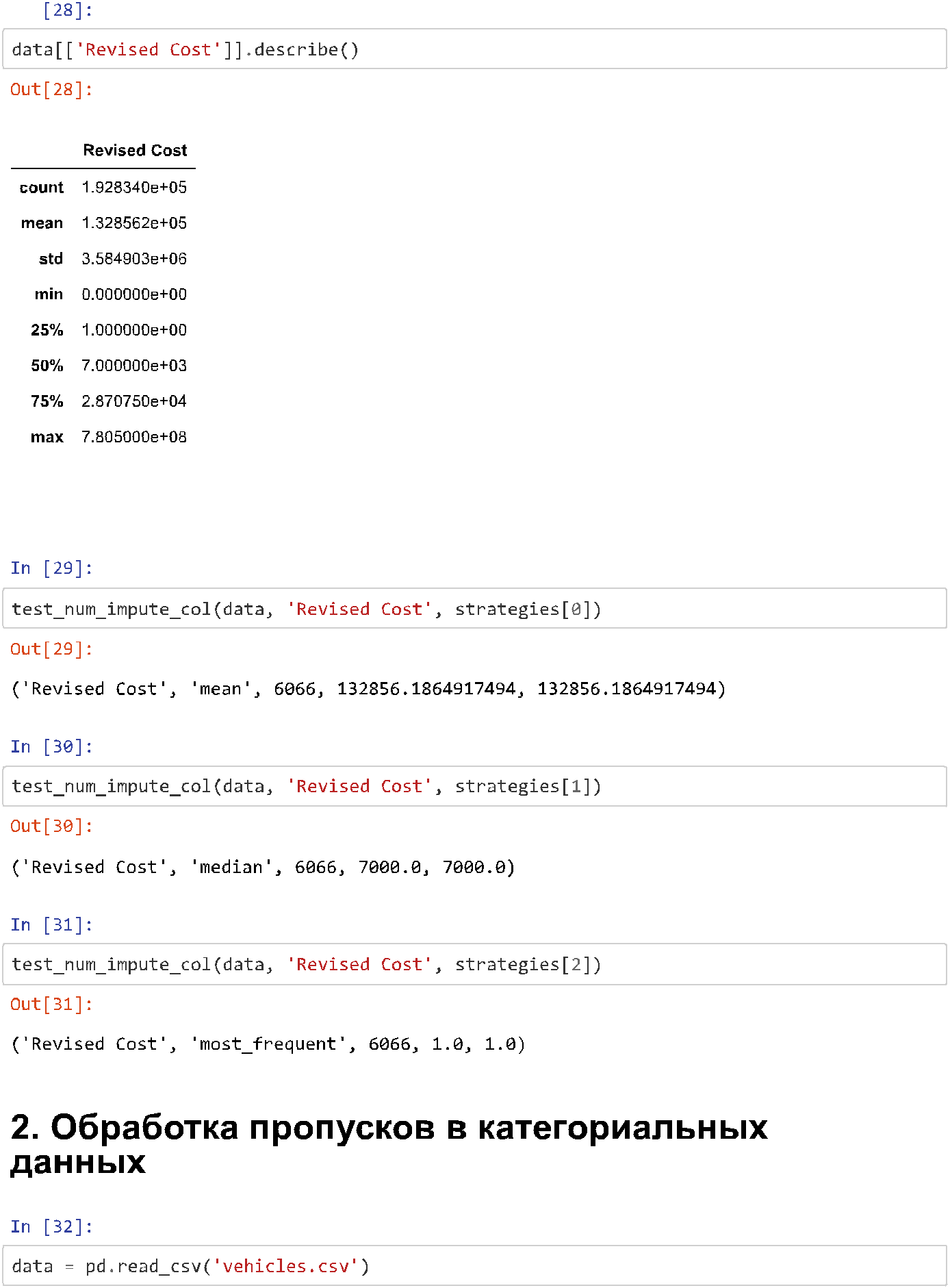


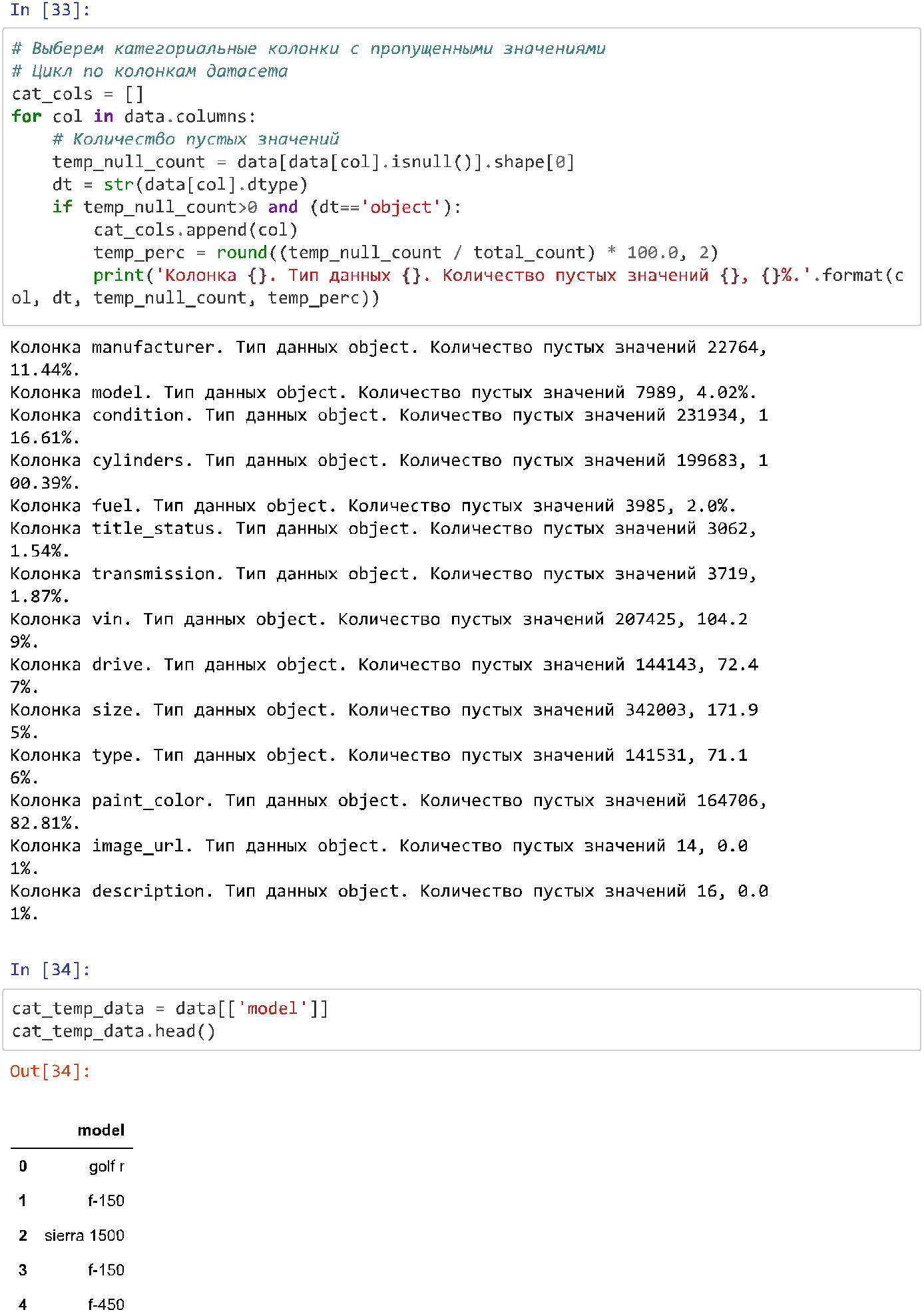


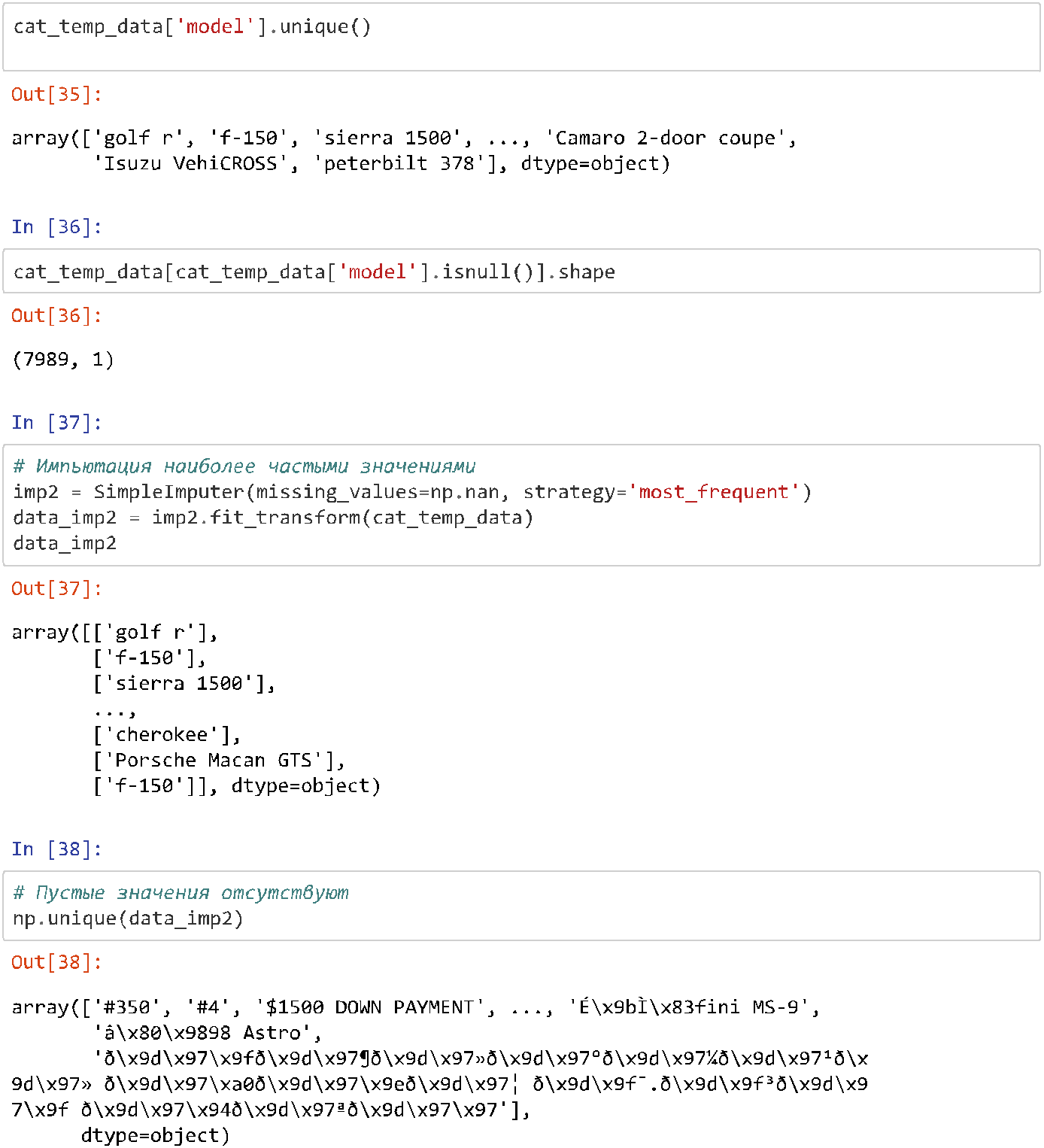


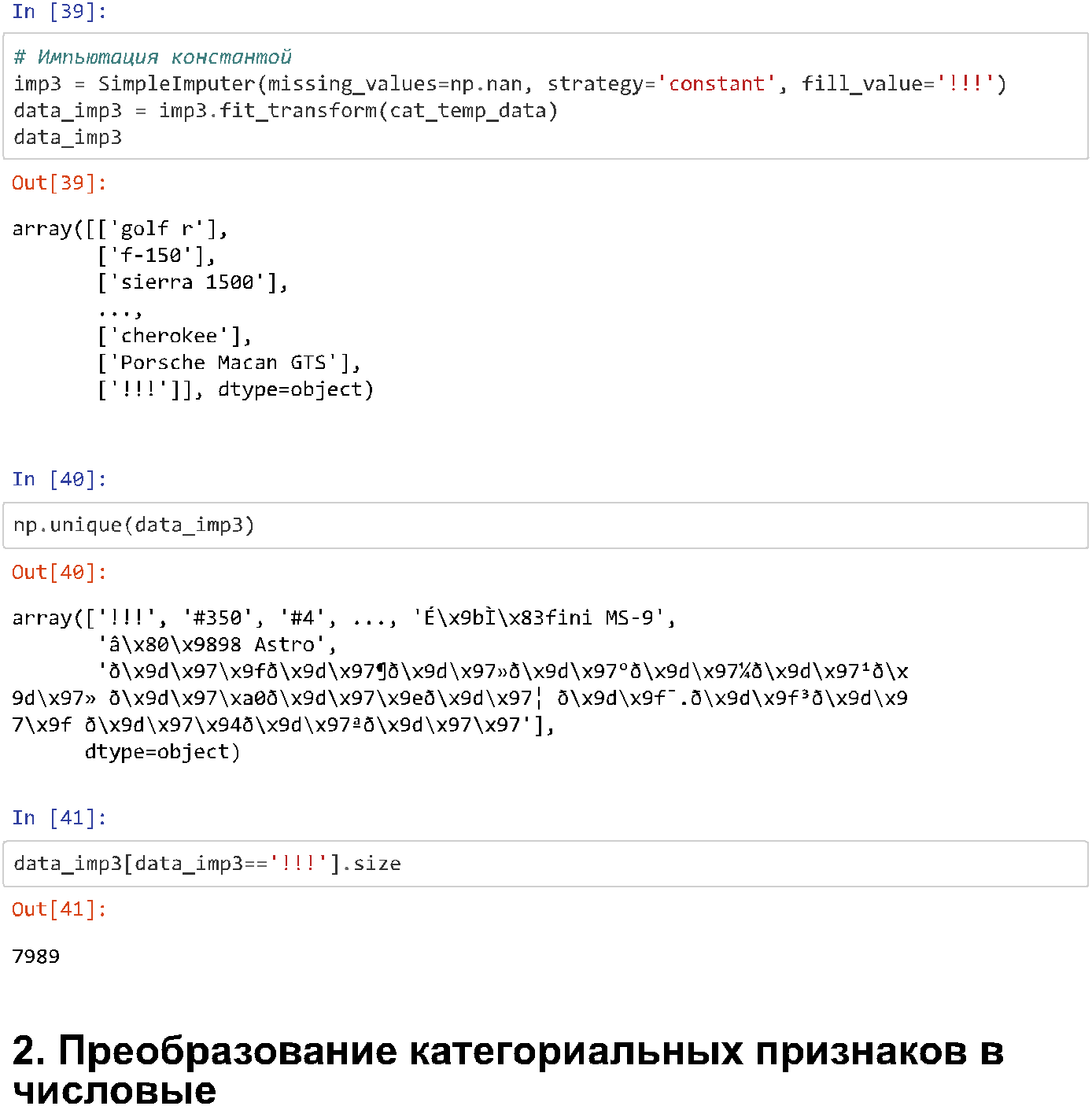






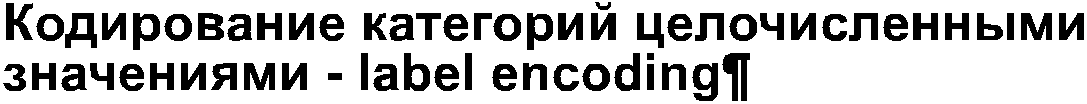


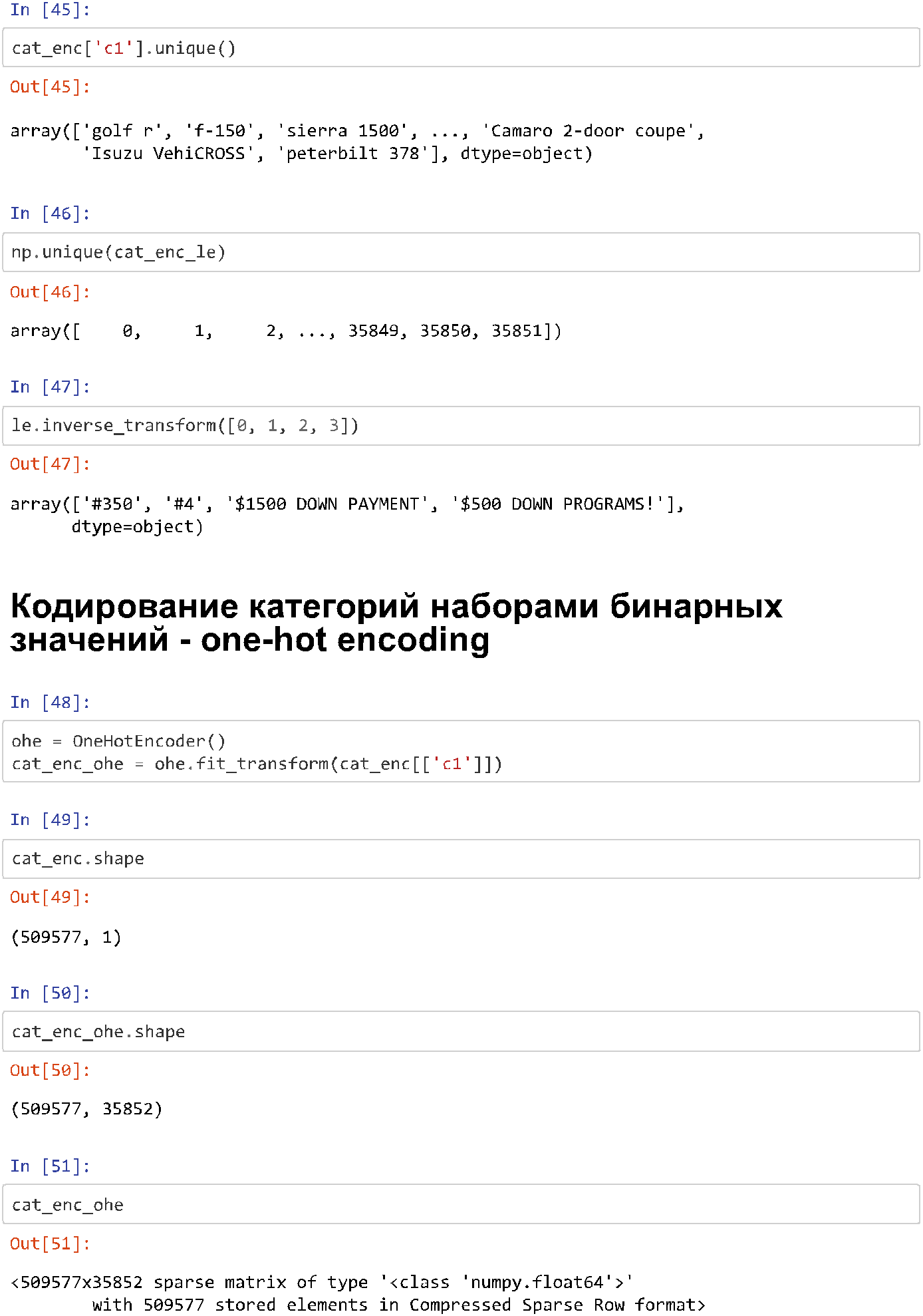




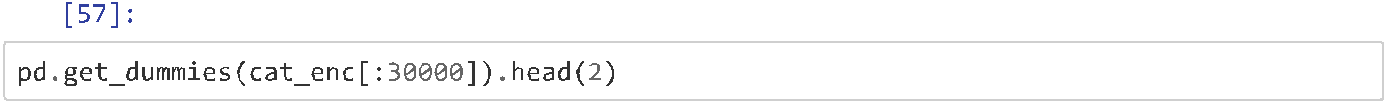


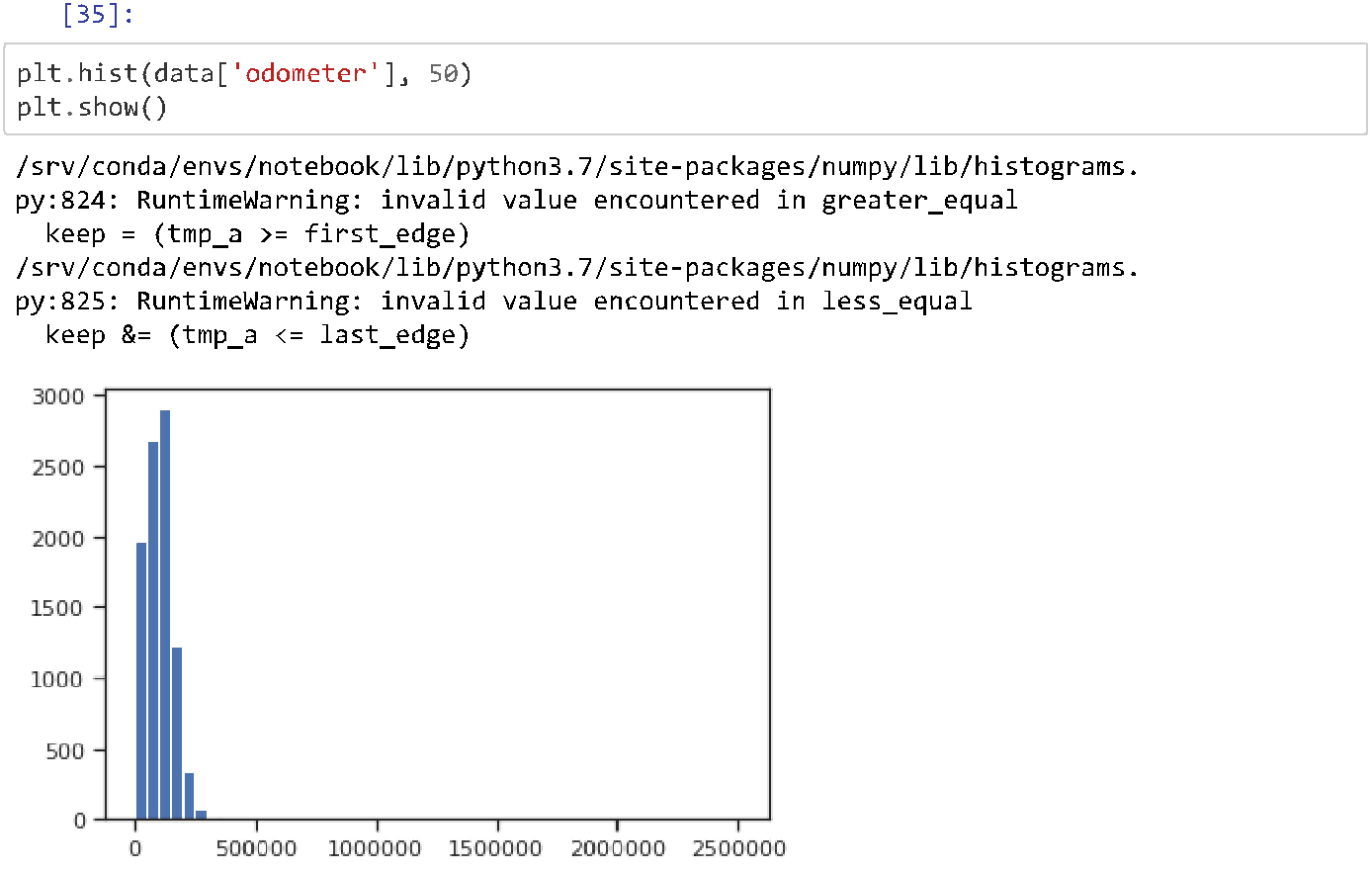


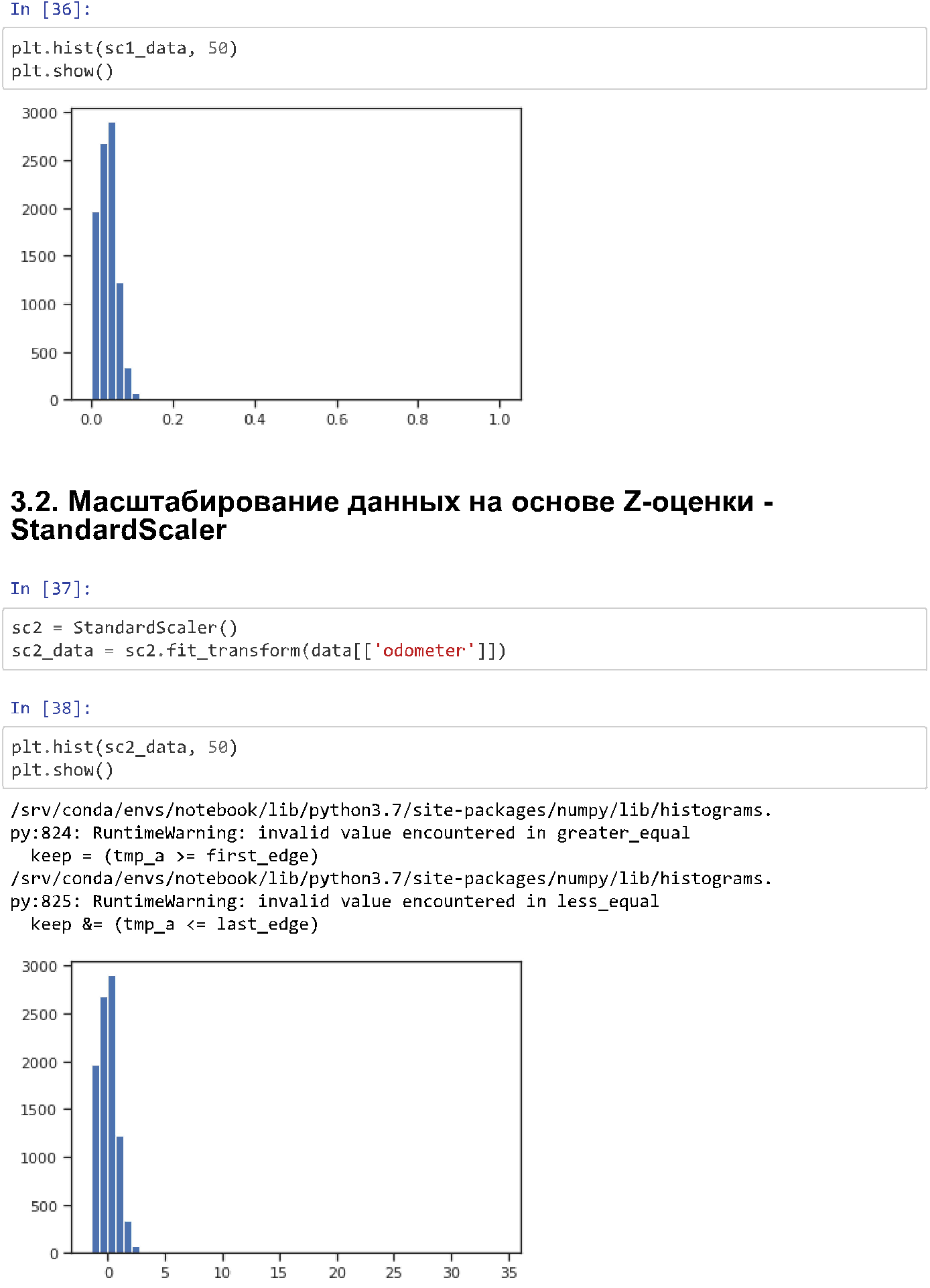


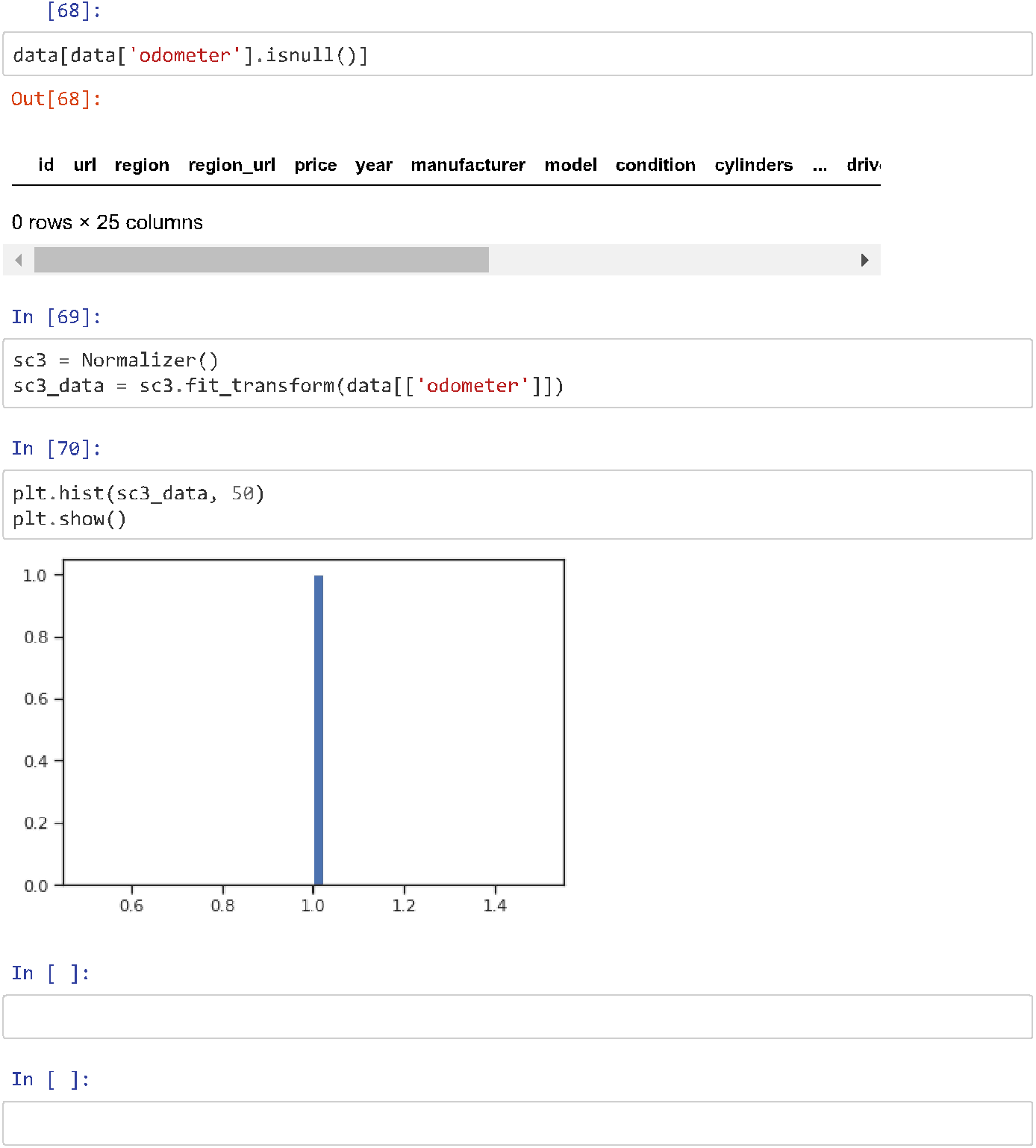
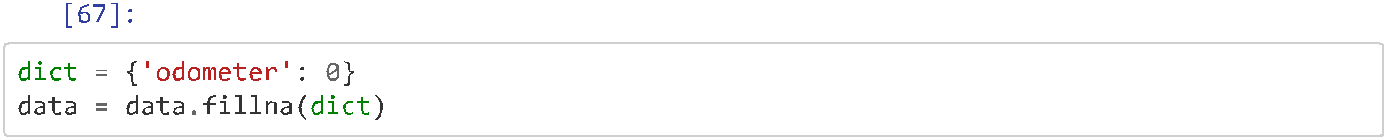














Вывод: в ходе выполнения лабораторной работы были изучены способы обработки пропусков данных, выполнено кодирование категориальных признаков, произведено масштабирование данных.

Обработка пропусков данных может быть выполнена следующими способами:

* + Удаление или заполнение нулями недостающих данных;
  + Внедрение значений (импьюация).
* помощью импьютации можно обрабатывать числовые и категориальные данные, цель метода – заполнить пропуски в данных усредненными значениями, заданными значениями и т.д. При выполнении лабораторной работы использовался класс SimpleImputer библиотеки sklearn, использовались следующие стратегии: 'mean', 'median','most\_frequent'.

Кодирование категориальных признаков выполняется при помощи целочисленных значений (label encoding) и бинарных значений (one-hot encoding). Используются классы библиотеки sklearn LabelEncoder и OneHotEncoder соответственно. Также был рассмотрен Pandas get\_dummies - быстрый вариант one-hot кодирования.

Масштабирование данных - изменение диапазона измерения величины.

Если признаки лежат в различных диапазонах, то необходимо их нормализовать. Как правило, применяют два подхода:

* + MinMax масштабирование;
  + Масштабирование данных на основе Z-оценки.

MinMax масштабирование реализовано при помощи класса MinMaxScaler библиотеки sklearn, а масштабирование данных на основе Z-оценки - StandardScaler. Нормализация данных предполагает изменение распределения данных. Доступна в sklearn при помощи класса Normalizer.

**Список литературы**

1. Гапанюк Ю. Е. Лабораторная работа «Обработка пропусков в данных, кодиро-вание категориальных признаков, масштабирование данных» [Электронный ре-

сурс] // GitHub. **–** 2020. **–** Режим доступа: https://github.com/ugapanyuk/ml\_

course/wiki/LAB\_MISSING (дата обращения: 25.03.2020).

1. Team The IPython Development. IPython 7.3.0 Documentation [Electronic resource] // Read the Docs. — 2019. — Access mode: https://ipython.readthedocs.io/en/ stable/ (online; accessed: 20.02.2019).
2. Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. — Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).
3. pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode: http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).
4. Gupta L. Google Play Store Apps [Electronic resource] // Kaggle. — 2019. — Access mode: https://www.kaggle.com/lava18/google-play-store-apps (online; accessed: 05.04.2019).