МГТУ им. Н. Э. Баумана, кафедра ИУ5

курс “Технология машинного обучения”

Рубежный контроль №1

# «Технологии разведочного анализа и обработки данных»

ВЫПОЛНИЛ:

Фонканц Р.В.

Группа: ИУ5-61Б

Вариант: 23

ПРОВЕРИЛ:

Гапанюк Ю.Е.

Москва 2020

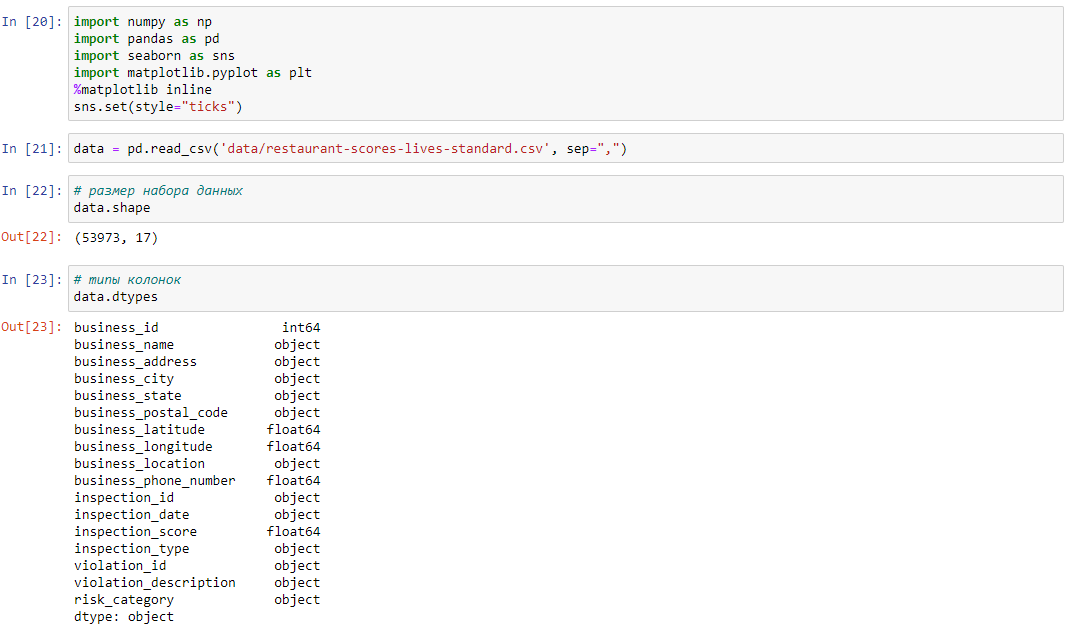
**Задание (№3):**

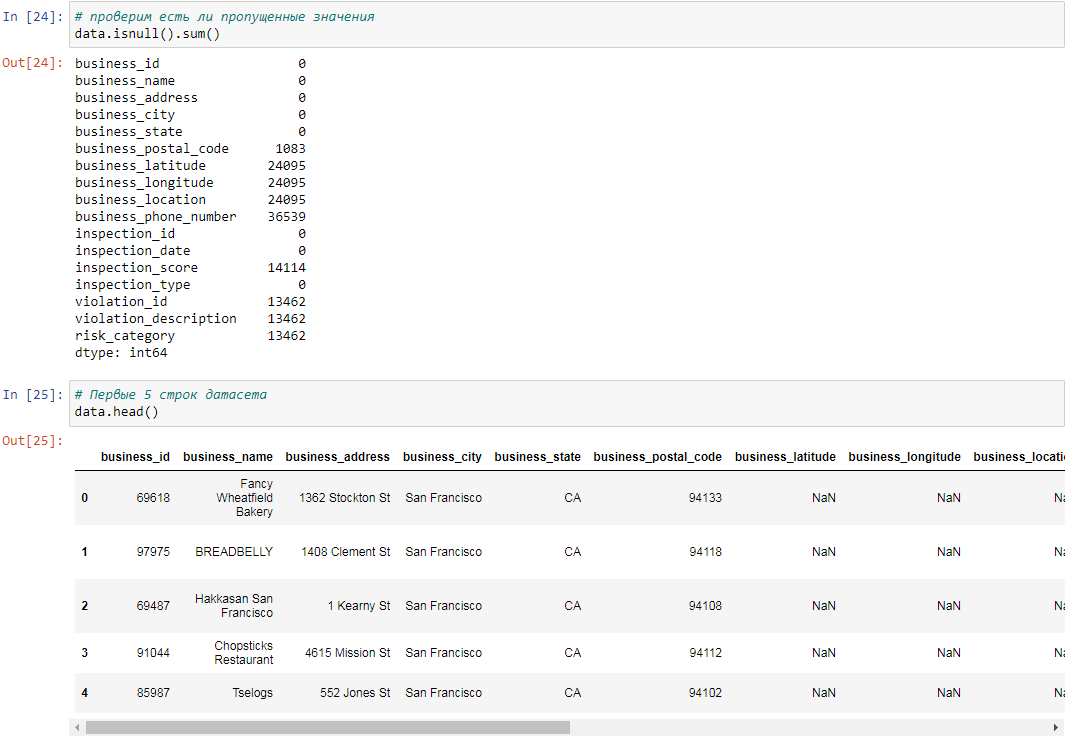
Для заданного набора данных произведите масштабирование данных (для одного признака) и преобразование категориальных признаков в количественные двумя способами (label encoding, one hot encoding) для одного признака. Какие методы Вы использовали для решения задачи и почему?

**Набор данных (№7):** <https://www.kaggle.com/san-francisco/sf-restaurant-scores-lives-standard>

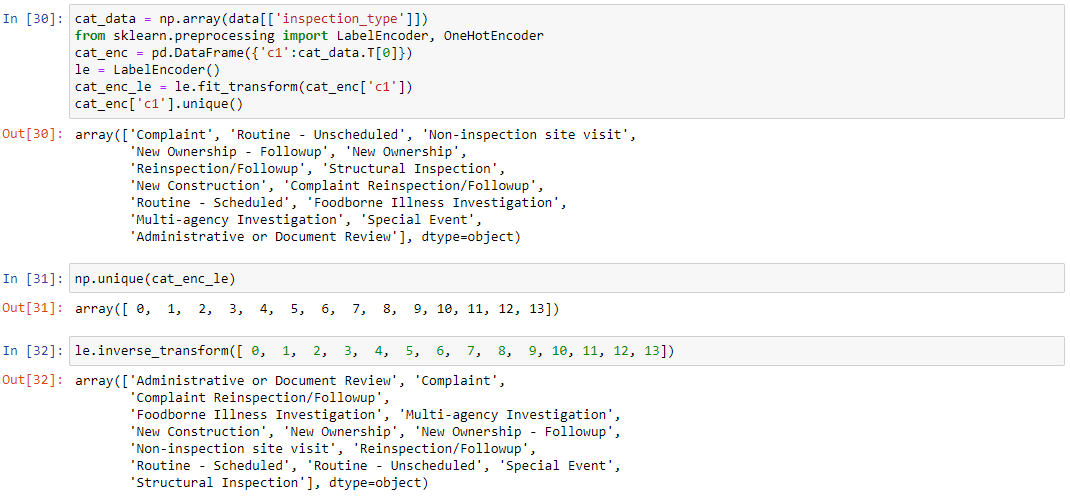
**Выполненная работа:**

## Загрузка и первичный анализ данных

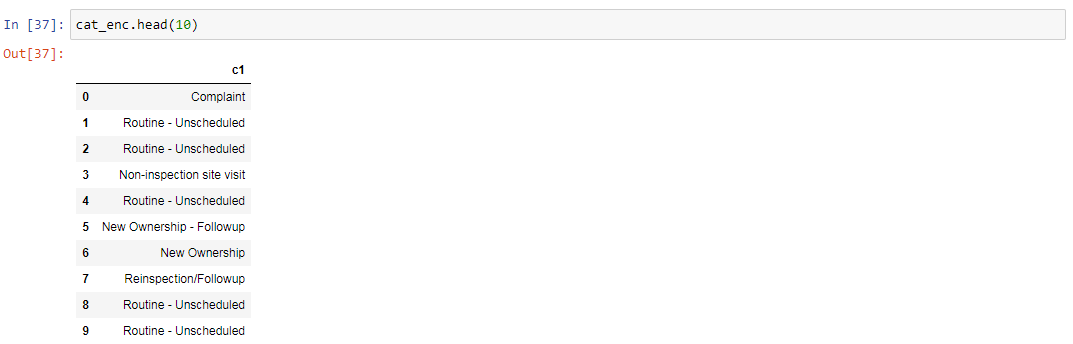
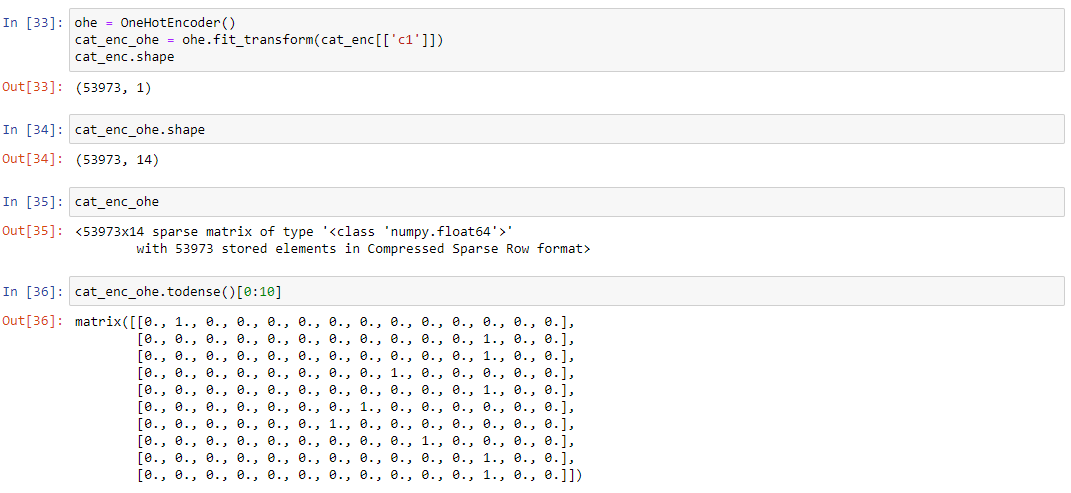




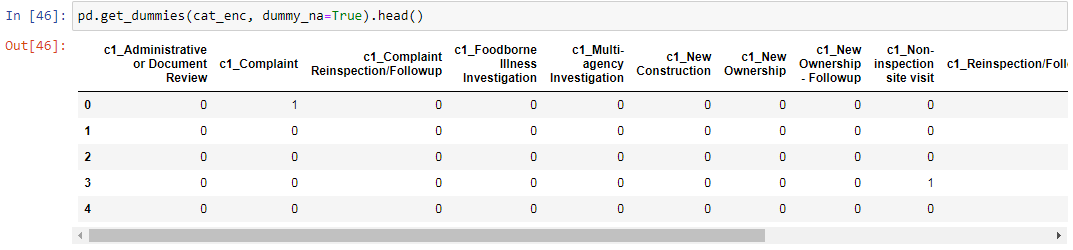
## 2.1. Кодирование категорий целочисленными значениями - [label encoding](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html)



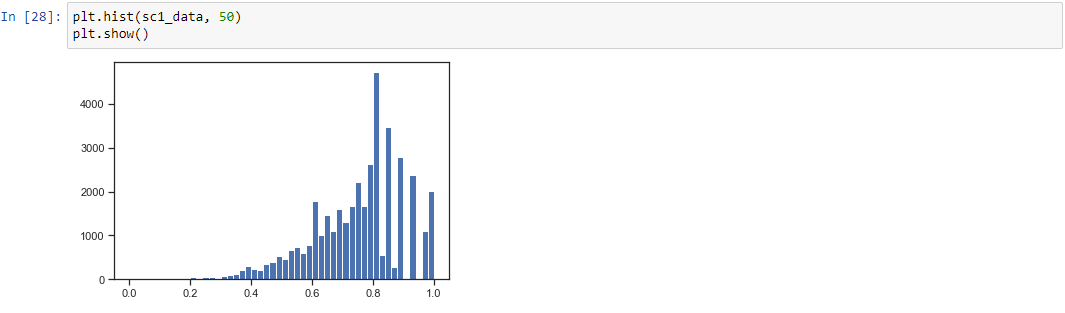
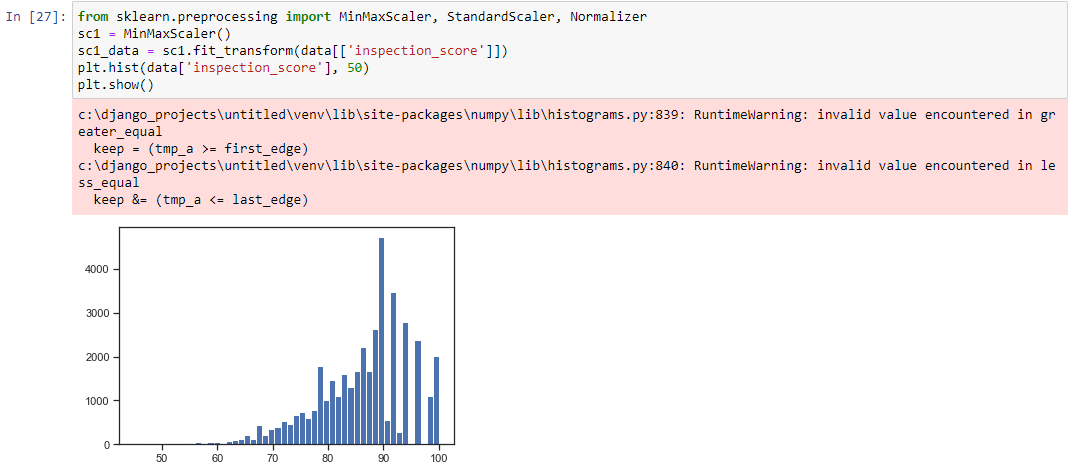
## 2.2. Кодирование категорий наборами бинарных значений - [one-hot encoding](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html)



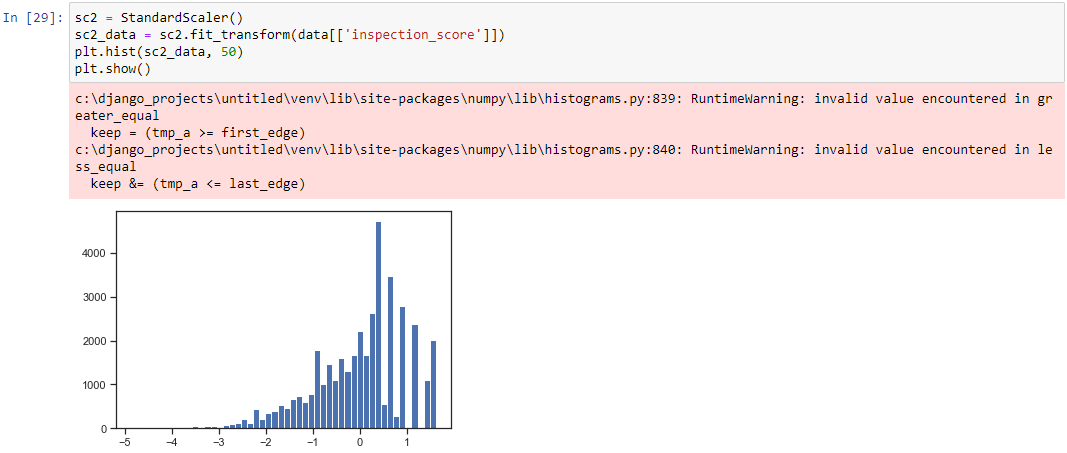
## 2.3. [Pandas get\_dummies](https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.get_dummies.html) - быстрый вариант one-hot кодирования



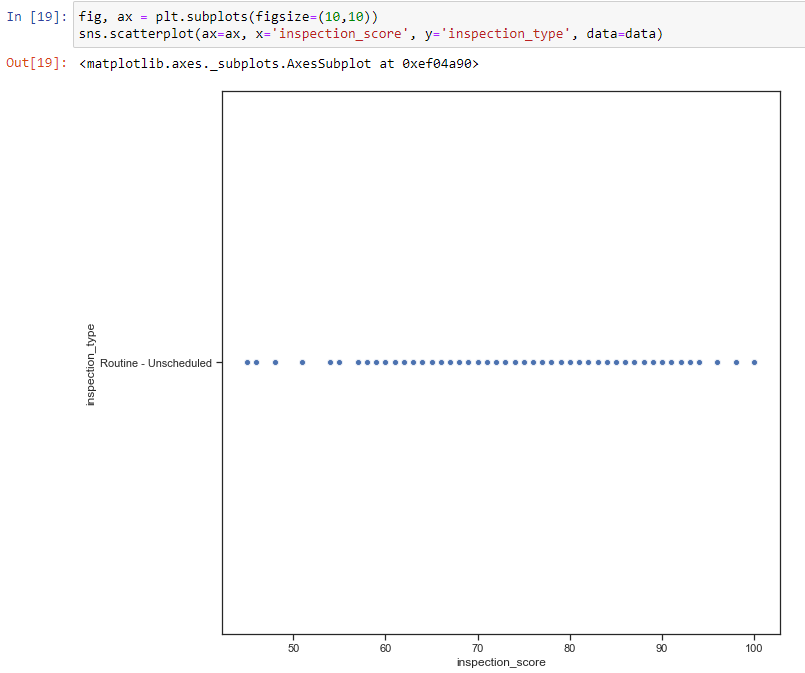
## 3. Масштабирование данных. [MinMax масштабирование](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html)



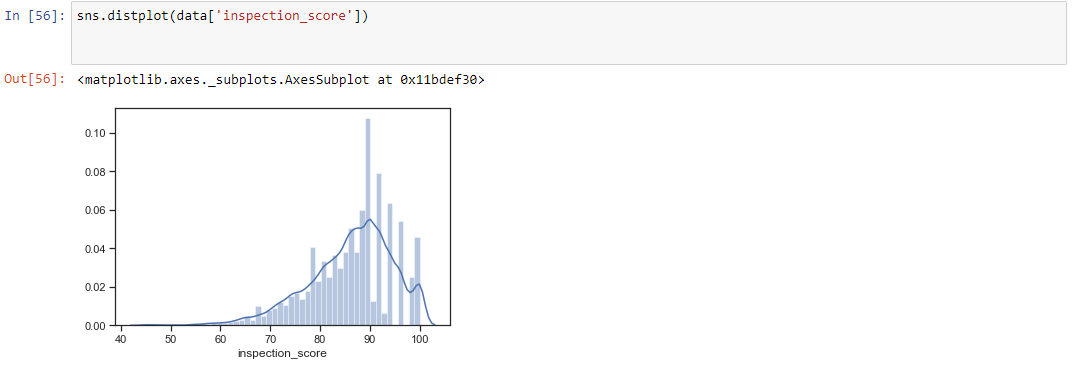
## 3.2. Масштабирование данных на основе [Z-оценки](https://ru.wikipedia.org/wiki/Z-%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0) - [StandardScaler](https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html" \l "sklearn.preprocessing.StandardScaler)



4. Дополнительное задание по группам: «Диаграмма рассеяния»



Т.к. диаграмма рассеяния не дает корректного представления о данных, применим гистограмму:



В данной работе были применены различные методы обработки данных. Для выбранного датасета лучшим методом масштабирования является MinMax масштабирования, т.к. выбранный признак изменяется в диапазоне от 0 до 100. В качестве метода для преобразования категориальных признаков в количественные лучше всего себя показал one-hot encoding, т.к. выбранный признак имеет целых 14 различных значений (label encoding неплох для небольшого количества возможных значений признака).

Данный отчет размещен в репозитории: <https://github.com/Rauger474/TMO>