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

Лабораторная работа №3  
по дисциплине  
«Методы машинного обучения»  
на тему

**«Разведочный анализ данных. Исследование и визуализация данных»**

Выполнил:  
студент группы ИУ5И-21М  
Ван Чжэн

Москва — 2024 г.

**1. Цель лабораторной работы**

изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

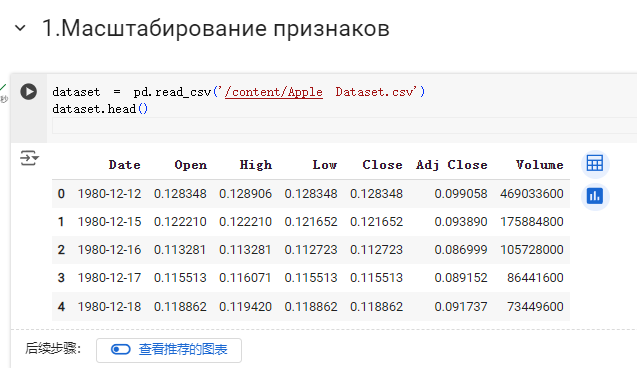
**2. Задание**

1. Выбрать один или несколько наборов данных (датасетов) для решения следующих задач. Каждая задача может быть решена на отдельном датасете, или несколько задач могут быть решены на одном датасете. Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   1. масштабирование признаков (не менее чем тремя способами);
   2. обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
   3. обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
   4. отбор признаков:
      * один метод из группы методов фильтрации (filter methods);
      * один метод из группы методов обертывания (wrapper methods);
      * один метод из группы методов вложений (embedded methods)

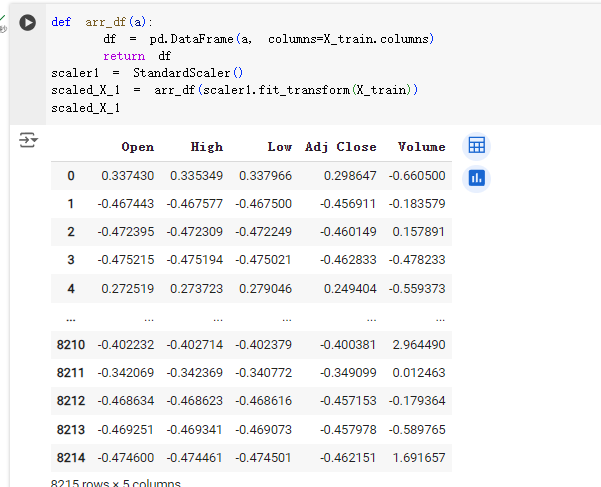
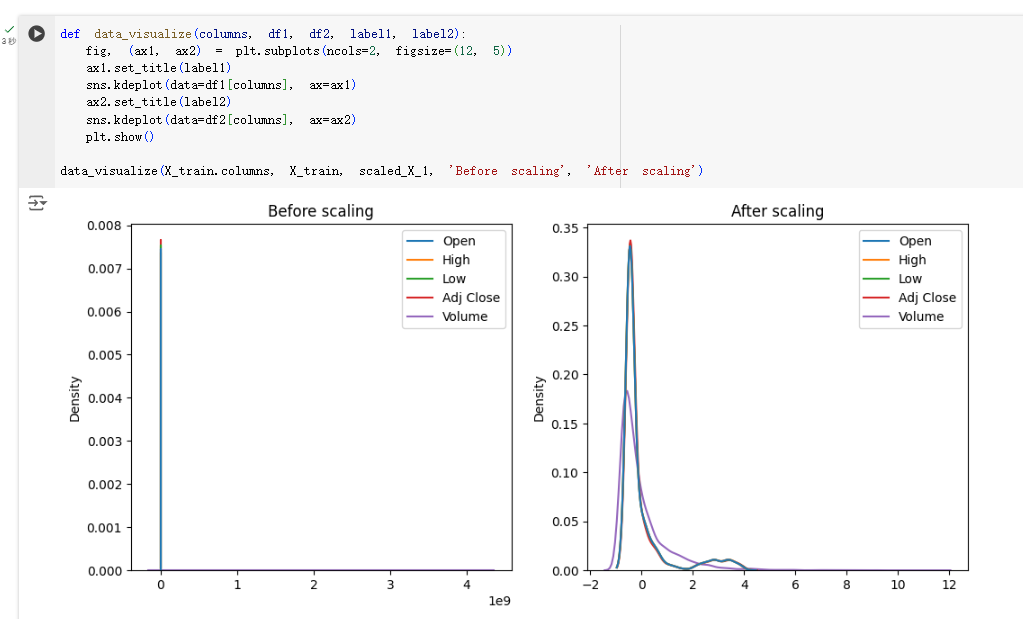
**3. Ход выполнения работы**

**3.1 Масштабирование признаков**

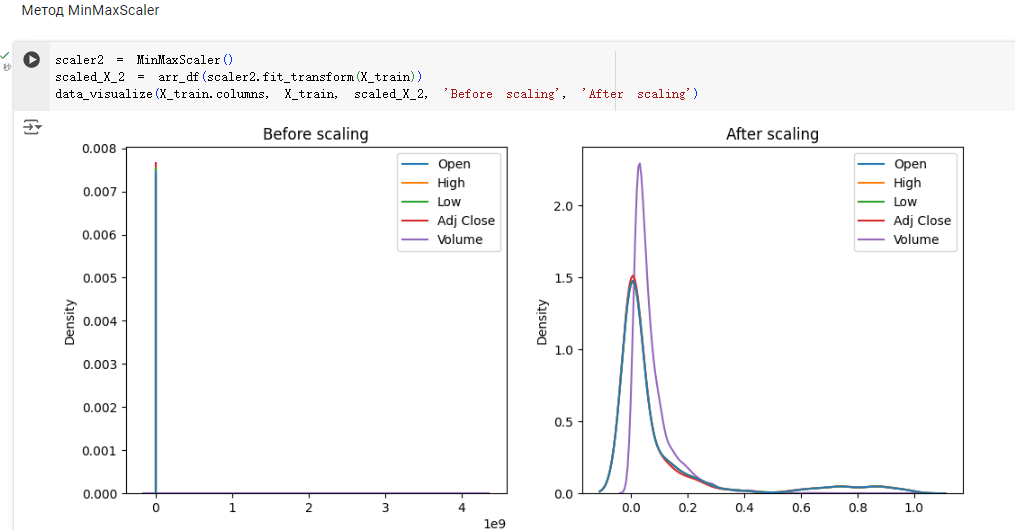
Я выбираю набор данных «Apple Stock Price».



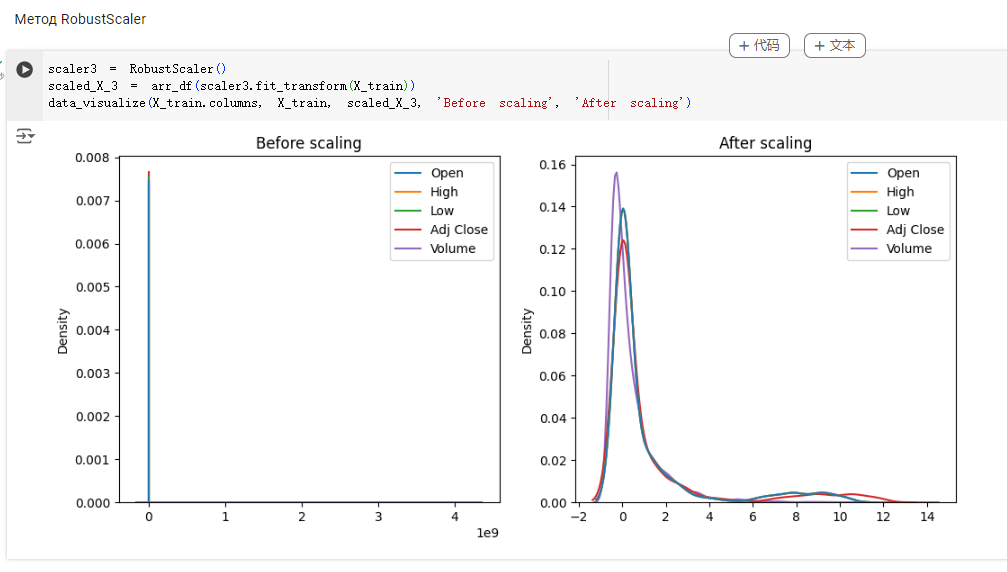
Метод Z-оценки

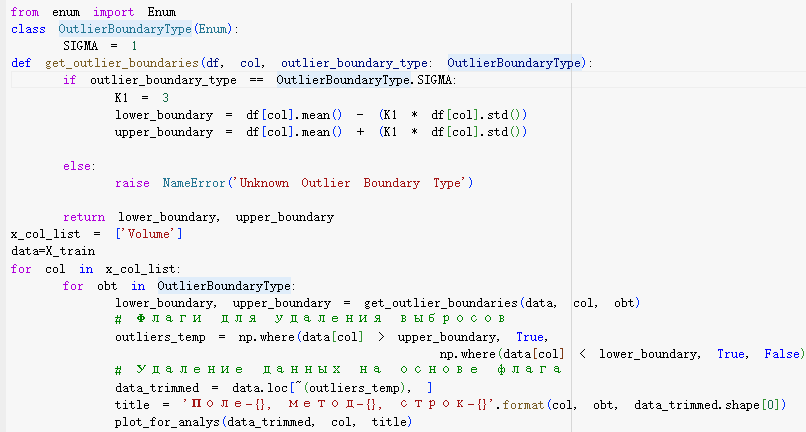
Метод MinMaxScaler

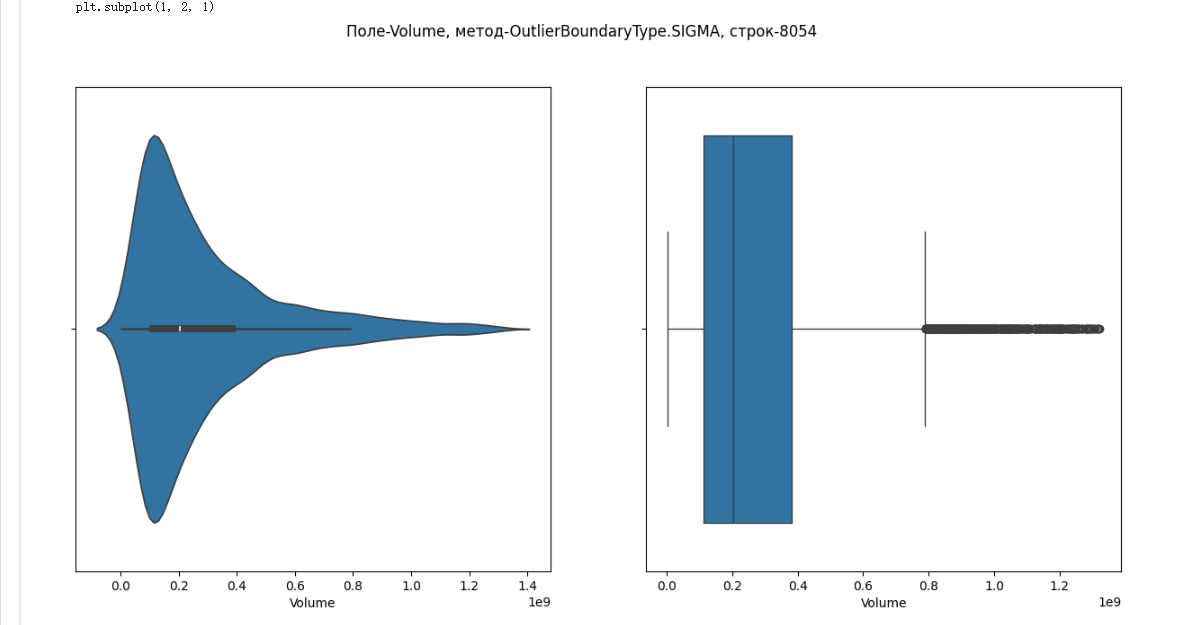


Метод RobusterScaler



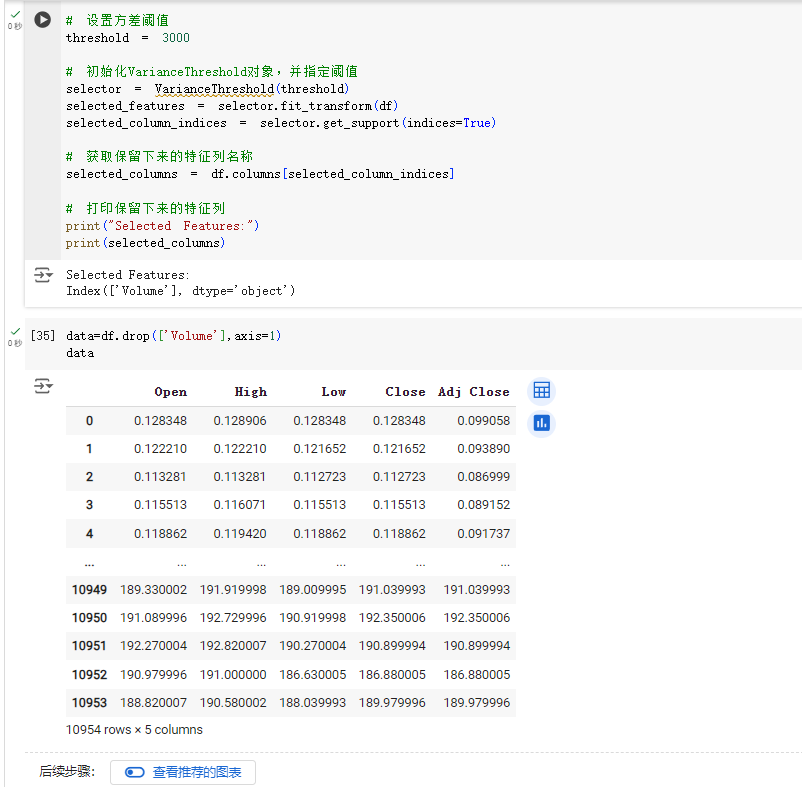
**3.2 Обработка выбросов для числовых признаков(удаление выбросов)**



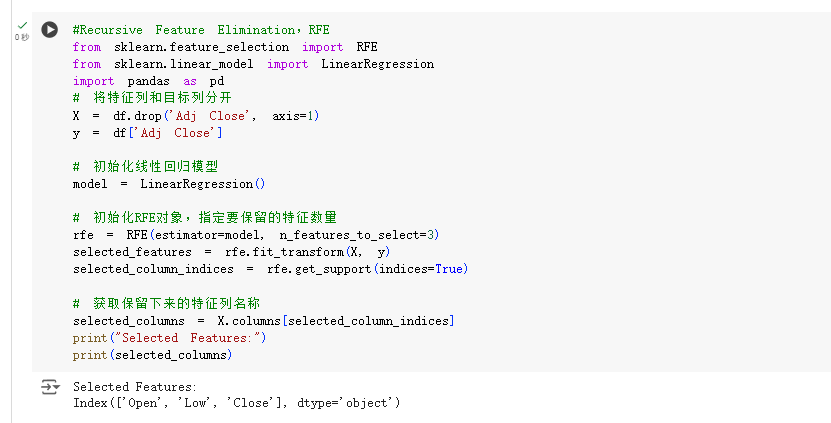


**3.3 Отбор признаков**

Метод из группы методов фильтрации- Variance Thresholding



Метод из группы методов обертывания- Recursive Feature Elimination，RFE





Метод из группы методов вложений



**Список литературы**

[1] Гапанюк Ю. Е. Лабораторная работа «Подготовка обучающей и тестовой выборки,

кросс-валидация и подбор гиперпараметров на примере метода ближайших соседей»

[Электронный ресурс] // GitHub. –– 2019. –– Режим доступа: https://github.com/

ugapanyuk/ml\_course/wiki/LAB\_KNN (дата обращения: 05.04.2019).

[2] 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).

[3] Waskom M. seaborn 0.9.0 documentation [Electronic resource] // PyData. — 2018. —

Access mode: https://seaborn.pydata.org/ (online; accessed: 20.02.2019).

[4] pandas 0.24.1 documentation [Electronic resource] // PyData. — 2019. — Access mode:

http://pandas.pydata.org/pandas-docs/stable/ (online; accessed: 20.02.2019).

[5] dronio. Solar Radiation Prediction [Electronic resource] // Kaggle. — 2017. — Access

mode: https://www.kaggle.com/dronio/SolarEnergy (online; accessed: 18.02.2019).

[6] Chrétien M. Convert datetime.time to seconds [Electronic resource] // Stack Overflow.

— 2017. — Access mode: https://stackoverflow.com/a/44823381 (online; accessed:

20.02.2019).

[7] scikit-learn 0.20.3 documentation [Electronic resource]. — 2019. — Access mode: https:

//scikit-learn.org/ (online; accessed: 05.04.2019).