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

Лабораторная работа №3

по дисциплине «Методы машинного обучения»

на тему«Обработка признаков (часть 2)»

Выполнил:  
студент группы： ИУ5-23М  
Аимань Мухэяти

Москва — 2021 г.

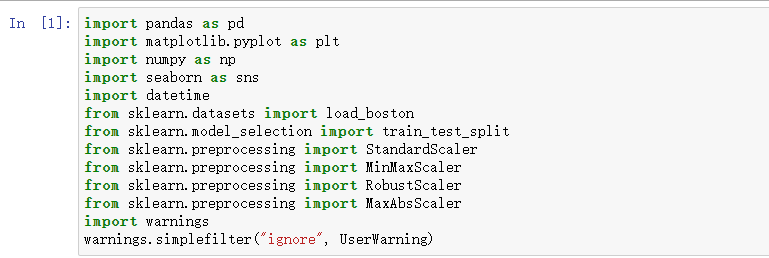
****Цель лабораторной работы:**** изучение продвинутых способов предварительной обработки данных для дальнейшего формирования моделей.

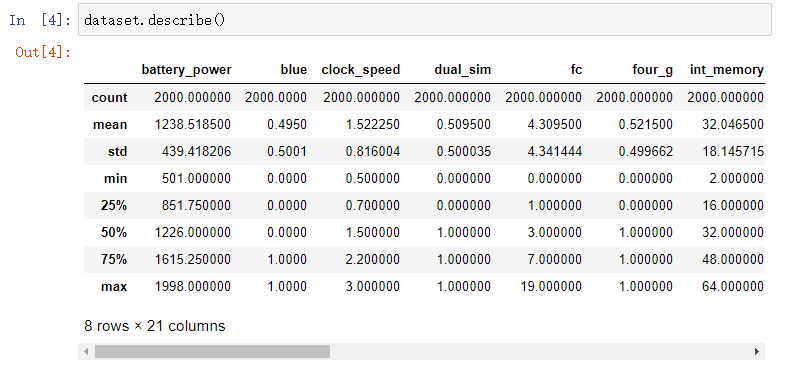
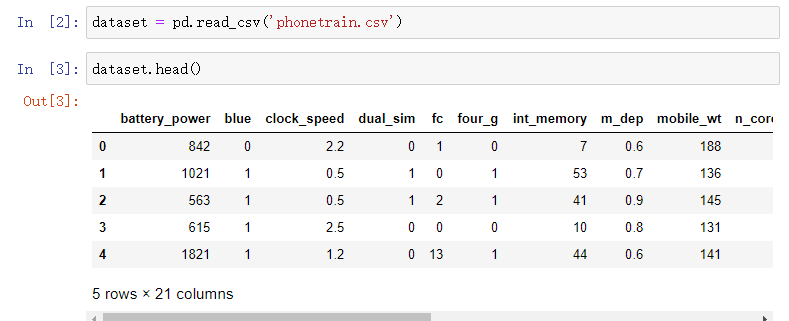
**Задание**: Выбрать набор данных (датасет), содержащий категориальные и числовые признаки и пропуски в данных.

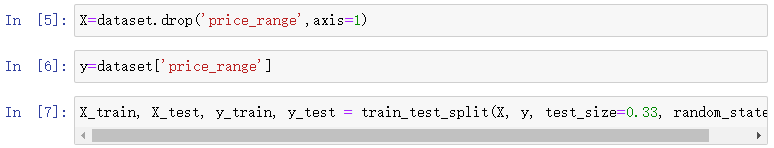
Для выполнения следующих пунктов можно использовать несколько различных наборов данных

Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:

* масштабирование признаков (не менее чем тремя способами);
* обработку выбросов для числовых признаков (по одному способу для удаления выбросов и для замены выбросов);
* обработку по крайней мере одного нестандартного признака (который не является числовым или категориальным);
* 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).



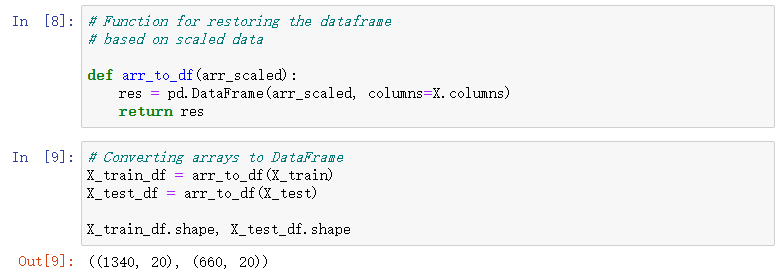


Разделим выборку на обучающую и тестовую

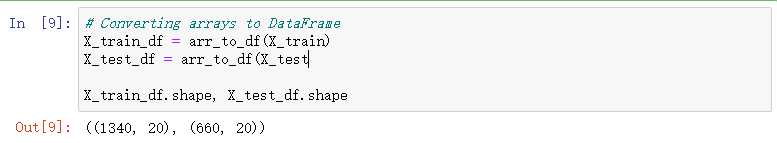
**Масштабирование**

Функция для восстановления датафрейма

на основе масштабированных данных

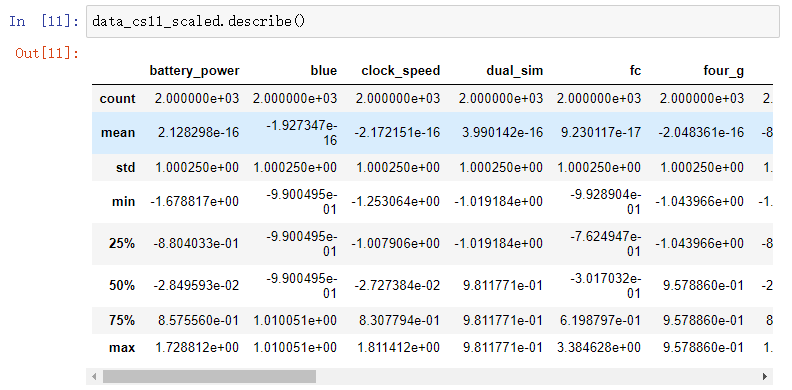


Преобразуем массивы в DataFrame

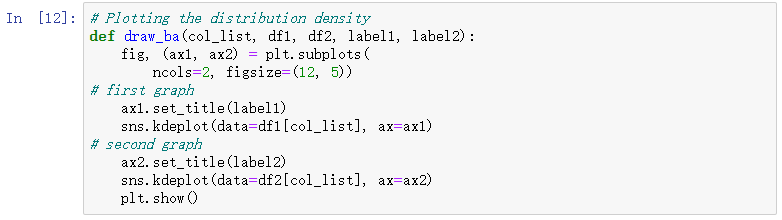


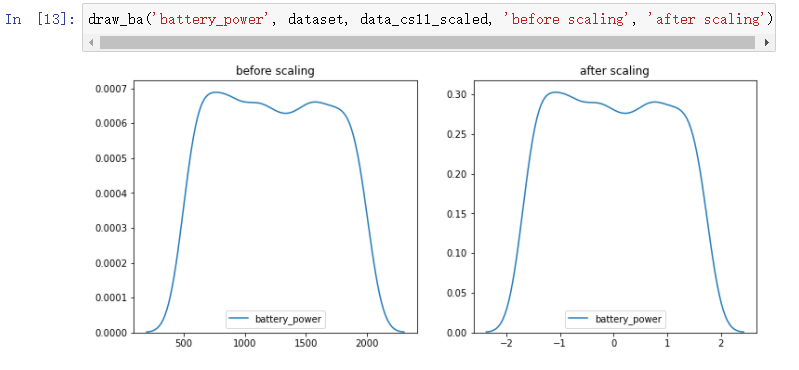
Масштабирование данных на основе Z-оценки

Обучаем StandardScaler на всей выборке и масштабируем

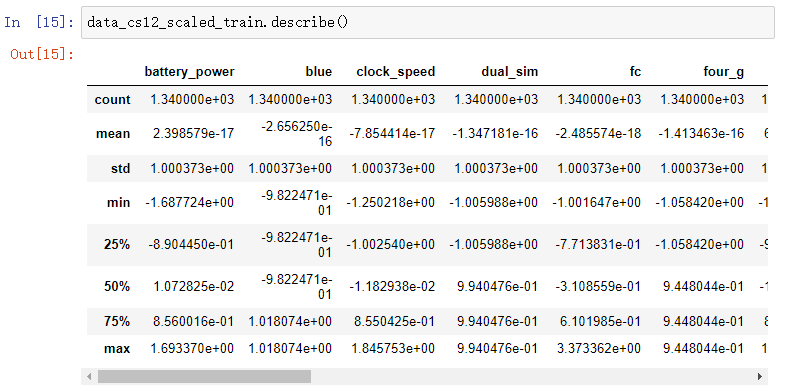
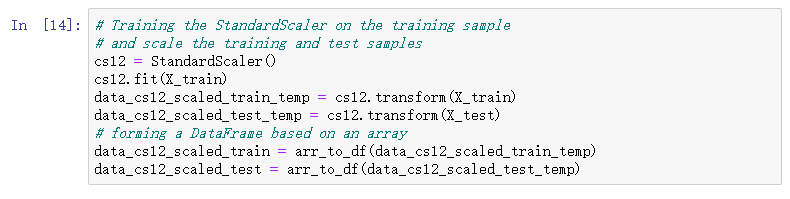


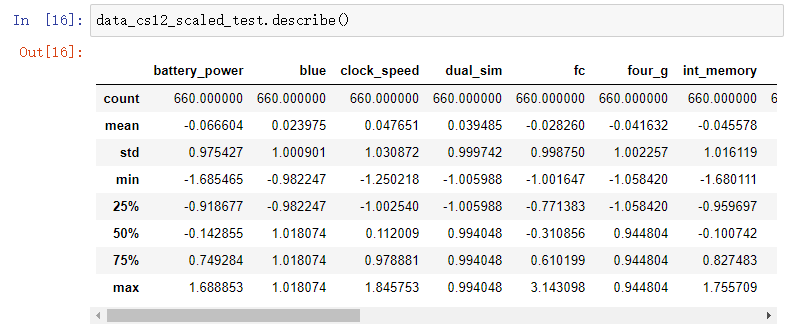
Построение плотности распределения



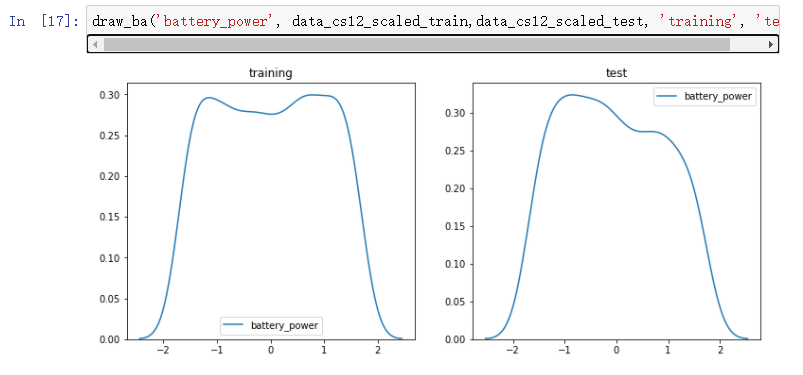


Обучаем StandardScaler на обучающей выборке и масштабируем обучающую и тестовую выборки

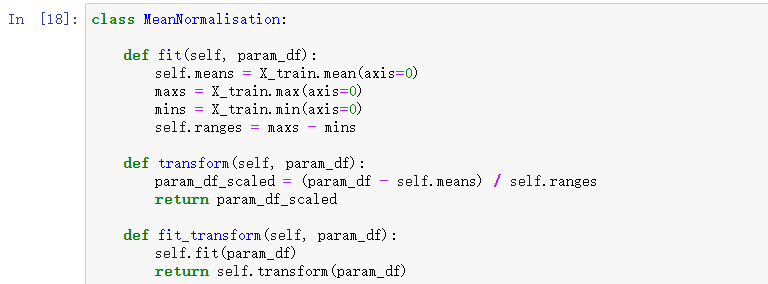


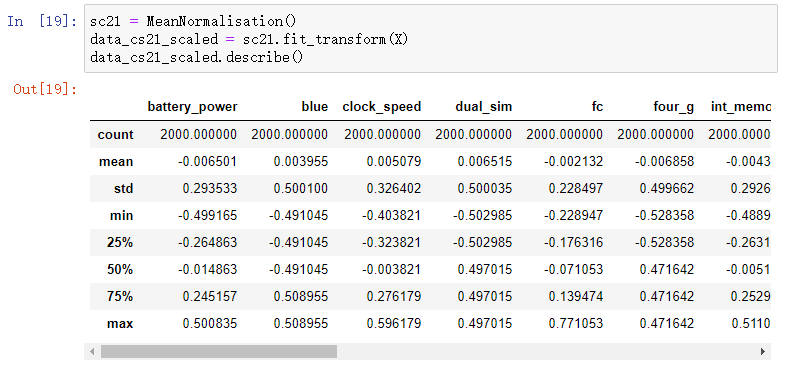


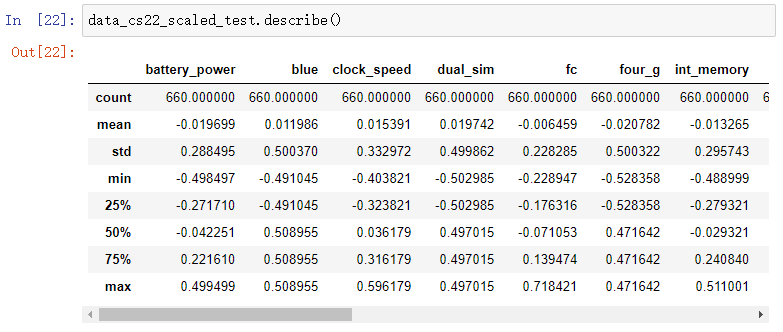
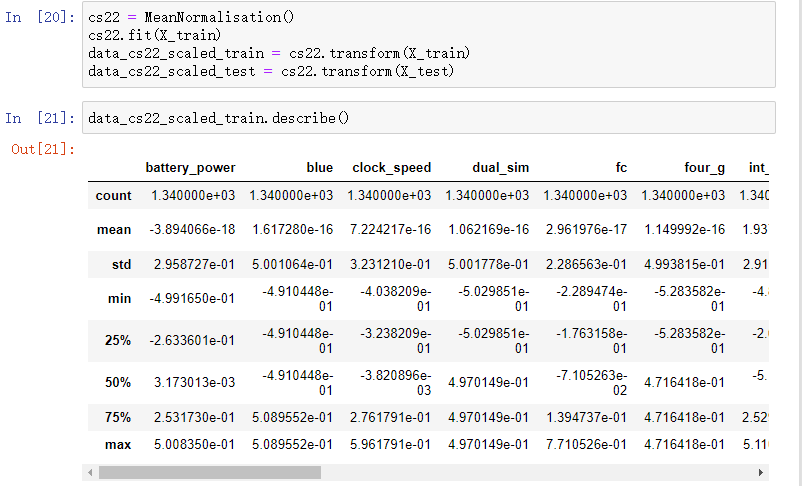
распределения для обучающей и тестовой выборки немного отличаются

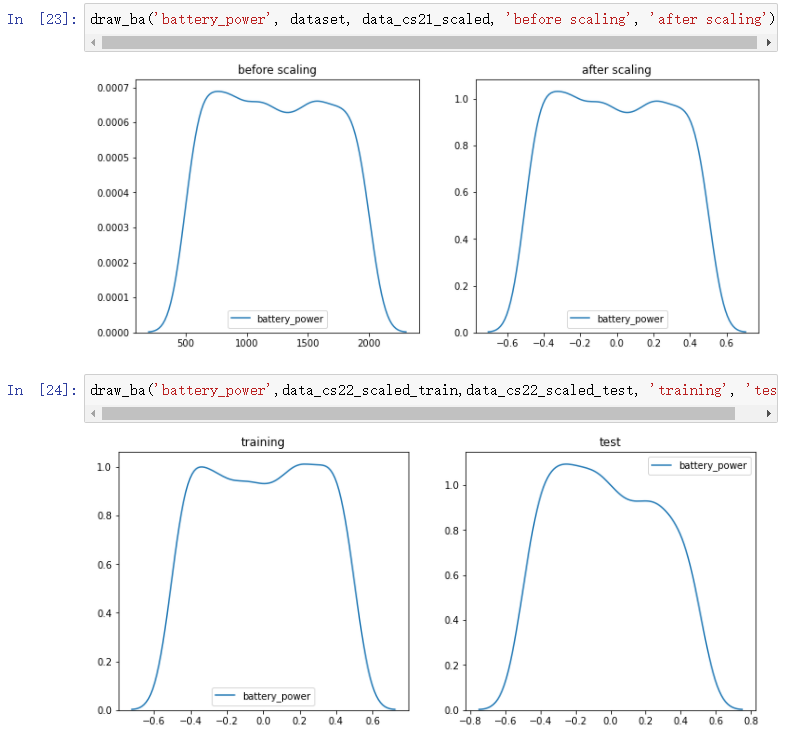


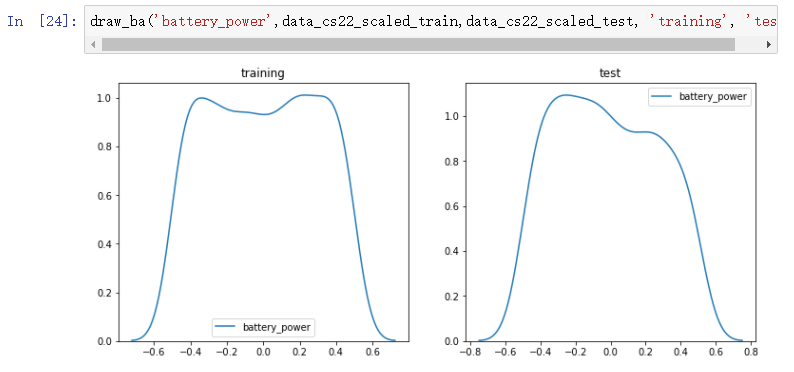
**Масштабирование "Mean Normalisation"**





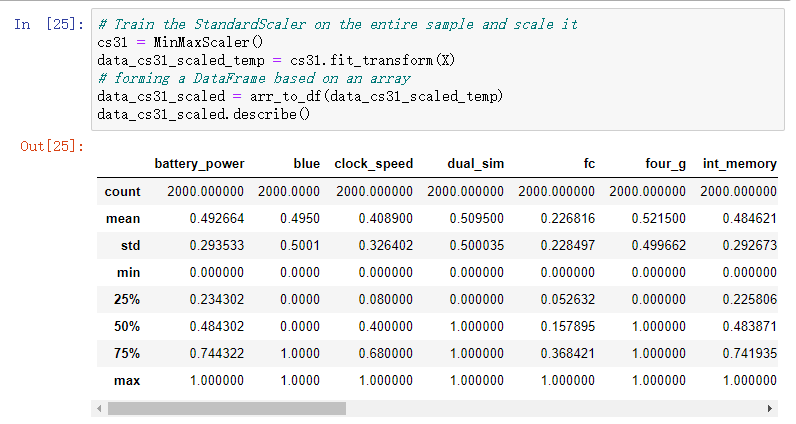




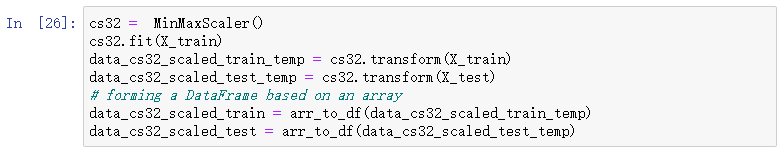


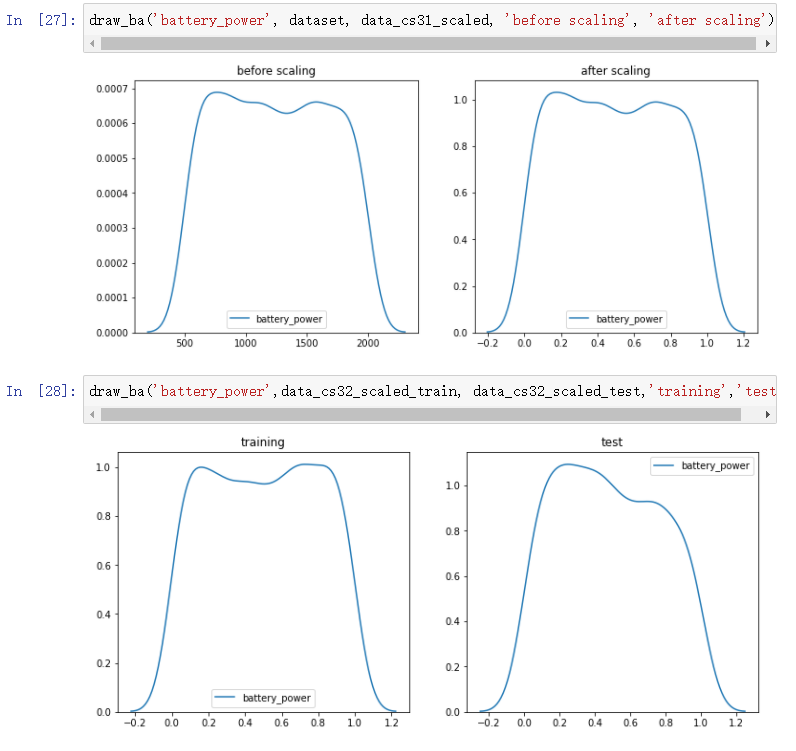
**MinMax-масштабирование**

Обучаем StandardScaler на всей выборке и масштабируем



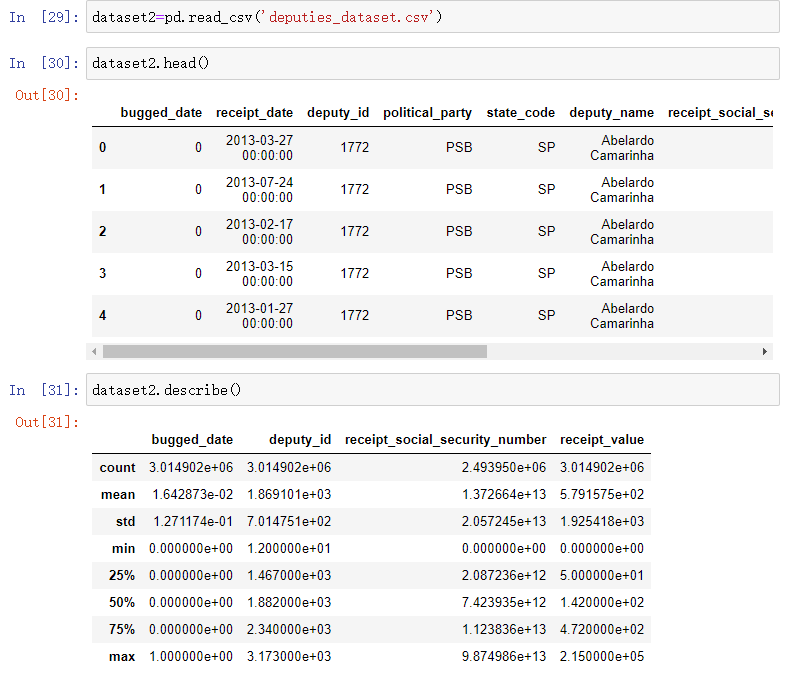
формируем DataFrame на основе массива

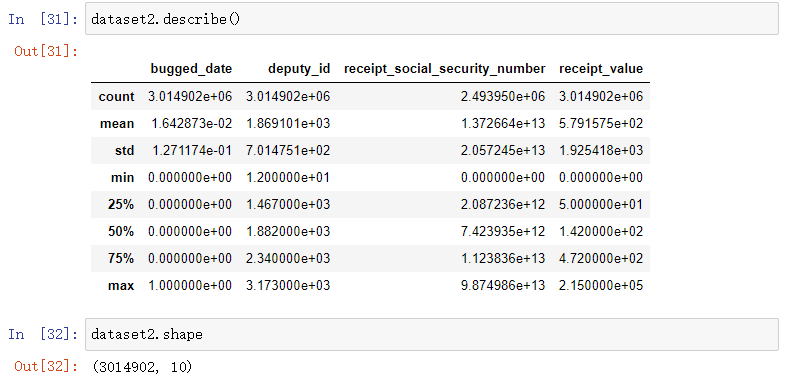




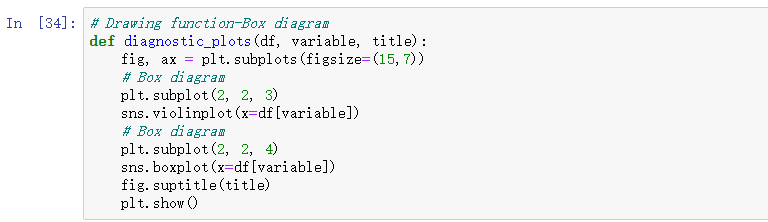
### **Обработка выбросов**

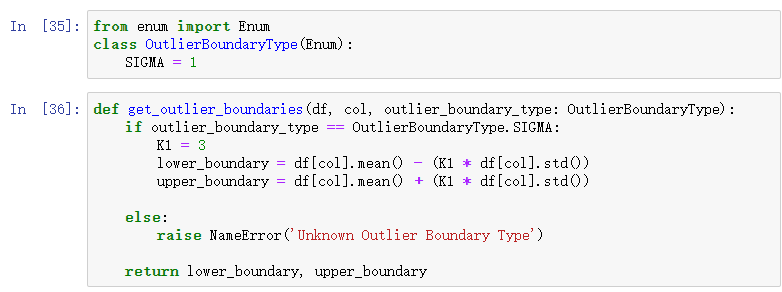
**Удаление выбросов**



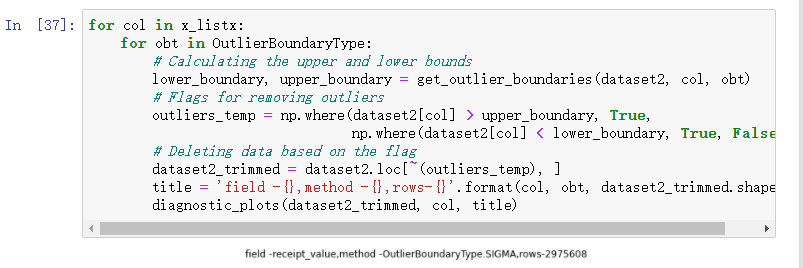


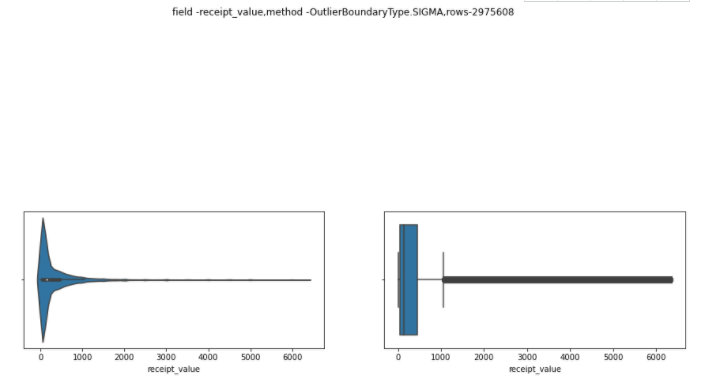
Функция построения графиков - ящики с усами





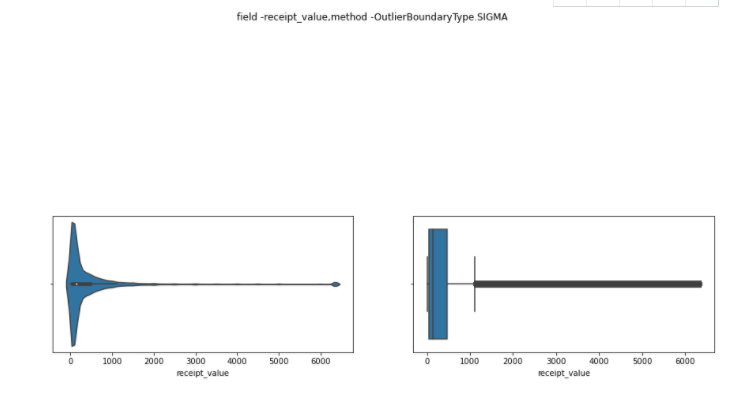
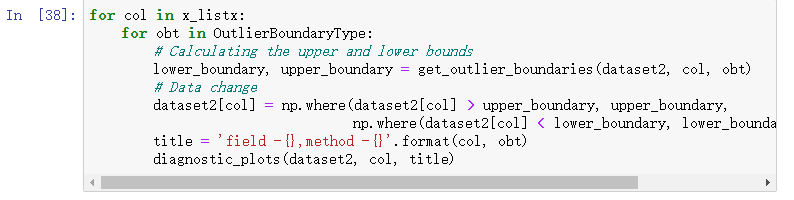
Вычисление верхней и нижней границы





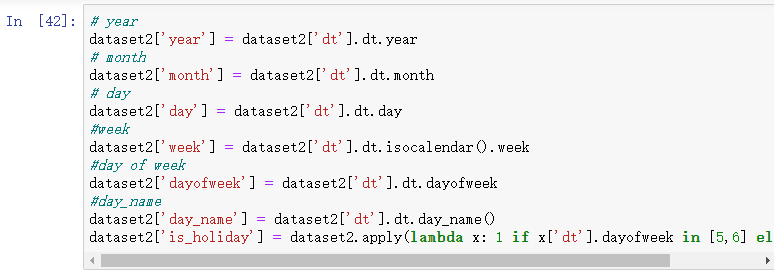
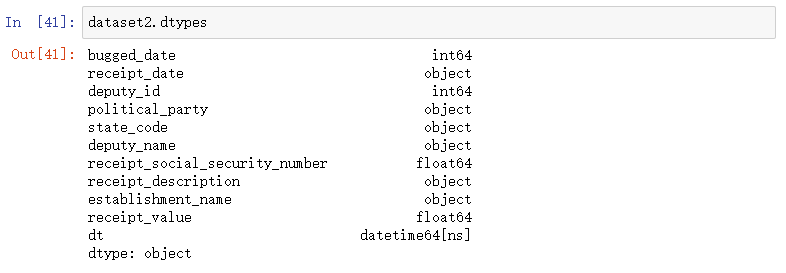
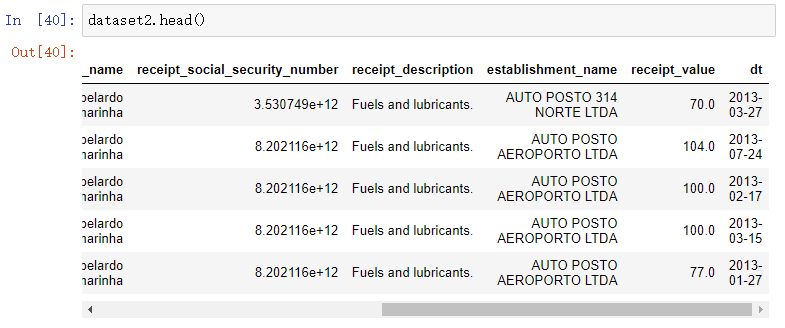
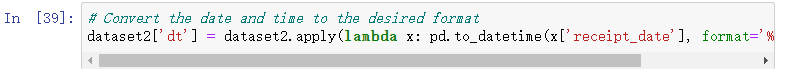
### **Замена выбросов**

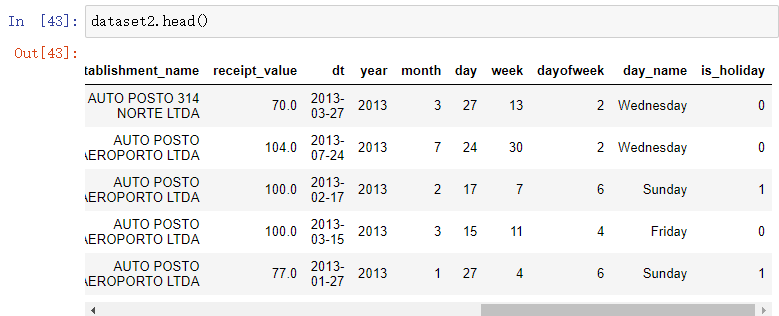
Вычисление верхней и нижней границы

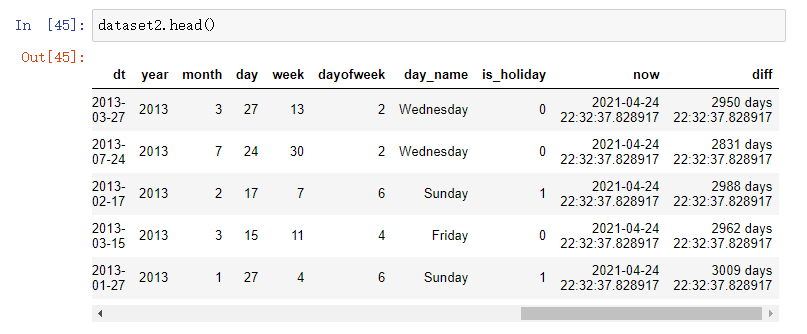
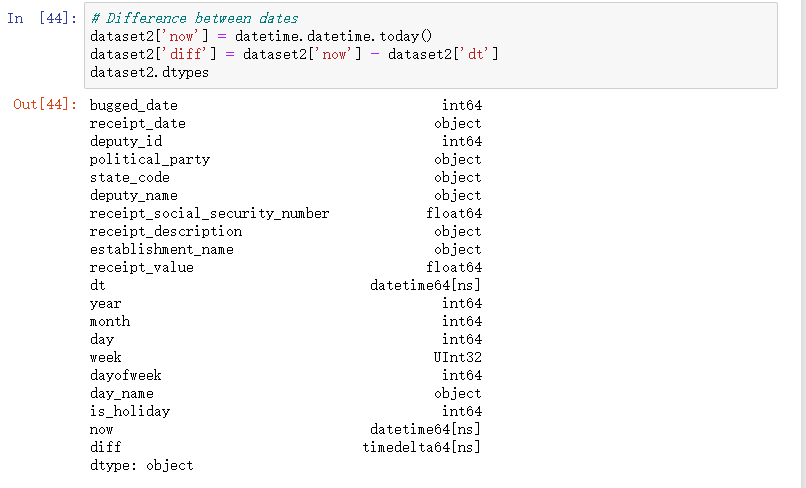


**Обработка нестандартного признака**

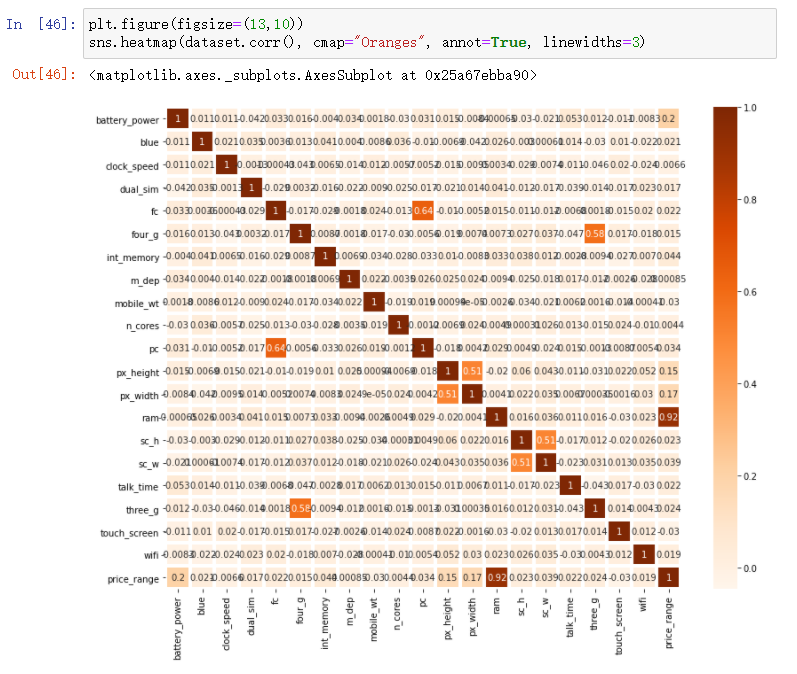
Сконвертируем дату и время в нужный формат



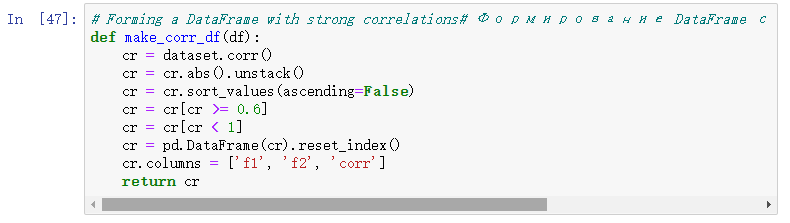


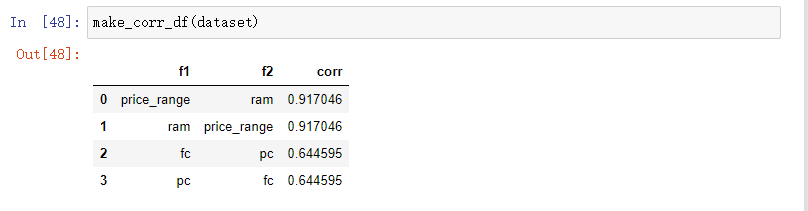
Разница между датами 

### **Отбор признаков из группы методов фильтрации (корреляция признаков**

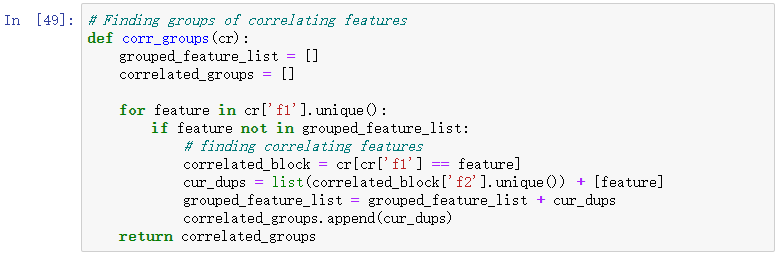


Формирование DataFrame с сильными корреляциями

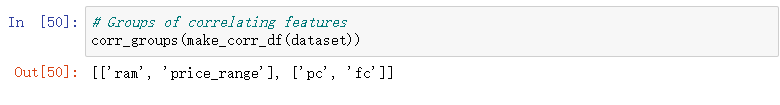




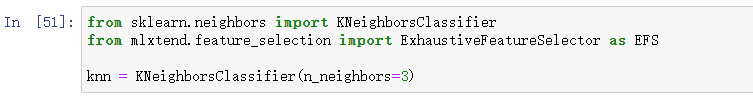
Обнаружение групп коррелирующих признаков

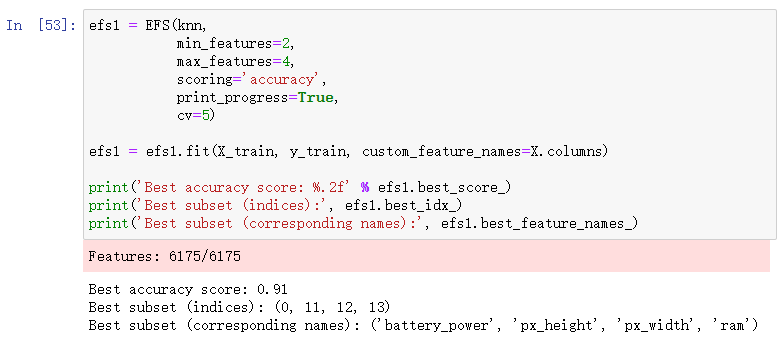


Группы коррелирующих признаков

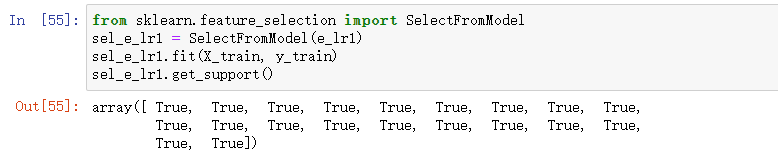
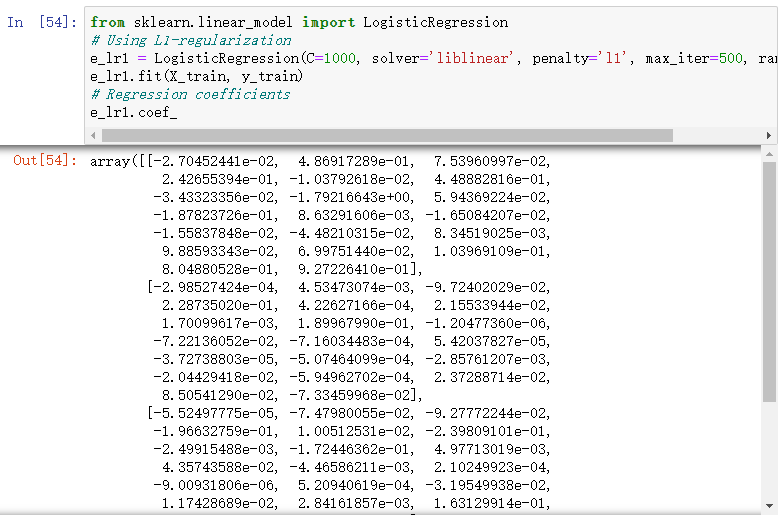


**Отбор признаков из группы методов обертывания (алгоритм полного перебора)**





**Отбор признаков из группы методов вложения (логистическая регрессия)**



### Список литературы

## [1] Гапанюк Ю. Е. Лабораторная работа «Обработка признаков (часть2)» [Электронный ресурс] https://github.com/ugapanyuk/ml\_course\_2021/wiki/LAB\_MMO\_\_FEATURES