МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

им. Н.Э. Баумана

Факультет «Информатика и системы управления»

Кафедра «Систем обработки информации и управления»

ОТЧЕТ

**Лабораторная работа №\_\_2\_**

по дисциплине«Методы машинного обучения в автоматизированных системах обработки информации и управления»

Тема: « Устранение пропусков в данных»

ИСПОЛНИТЕЛЬ: \_\_Се Цзявэнь\_\_\_\_

ФИО

группа ИУ5И-22М\_ \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_\_"\_\_\_\_\_\_\_\_\_202\_ г.

ПРЕПОДАВАТЕЛЬ: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

ФИО

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

подпись

"\_15\_"\_\_\_\_\_02\_\_\_\_2024\_ г.

Москва - 2024

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Задания

1. Выбрать набор данных (датасет), содержащий категориальные и числовые 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.) Просьба не использовать датасет, на котором данная задача решалась в лекции.
2. Для выбранного датасета (датасетов) на основе материалов лекций решить следующие задачи:
   1. устранение пропусков в данных;
   2. кодирование категориальных признаков;
   3. нормализация числовых признаков.

# Программы

* Текстовое описание набора данных:

Был выбран набор данных о диабете из sklearn. Для каждого из n = 442 пациентов с диабетом были получены десять исходных переменных: возраст, пол, индекс массы тела, среднее артериальное давление и шесть измерений сыворотки крови, а также интересующий ответ - количественный показатель прогрессирования заболевания через год после исходного уровня. Переменные называются «AGE SEX BMI BP S1 S2 S3 S4 S5 S6 Y».

Подробности набора данных следующие:

|  |  |
| --- | --- |
| Количество выборочных данных | 442 |
| Количество функций (размерность данных) | 10 |
| Значение каждой функции | Возраст, пол, ИМТ, среднее артериальное давление, S1, S2, S3, S4, S5, S6. |
| Диапазон значений функции | (-0.2,0.2) |
| Значение тега | На основе количественных измерений через год после прогрессирования заболевания. |

# Устранение пропусков в данных

## Удаление пропущенных значений

**Допущение:** пропуски распределены случайным образом.

**Когда рекомендуется использовать?**

* Если пропущенных данных слишком много и возникает опасность нарушить распределение исходных данных при заполнении пропусков. Рекомендуется удалять признак (колонку) целиком.
* Если датасет большой и пропущенных данных относительно немного, то рекомендуется удалять строки, содержащие пропуски в данных.
* Под "немного" в идеальном случае понимается 5% от выборки.

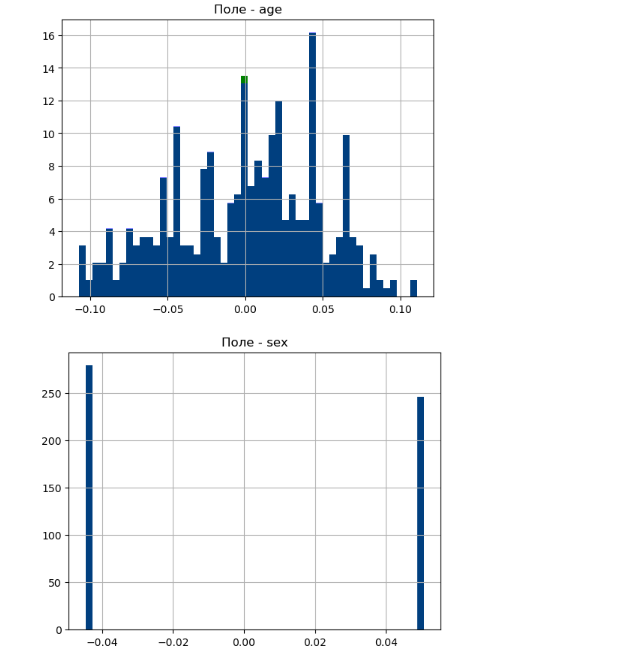
**Преимущества:**

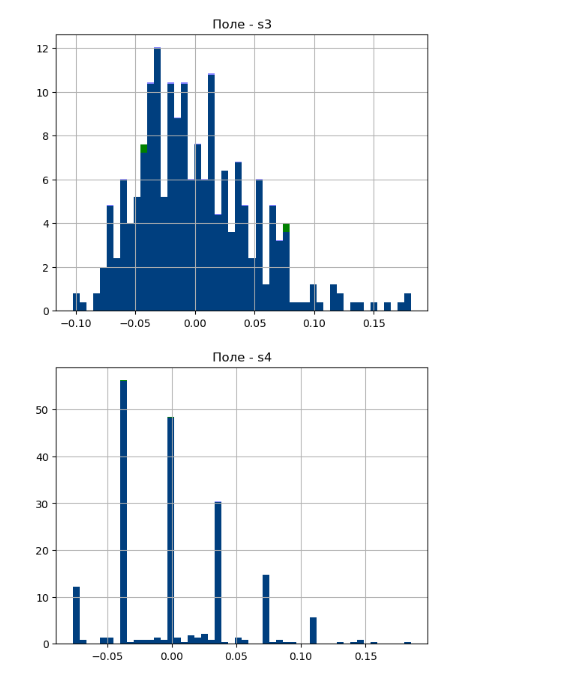
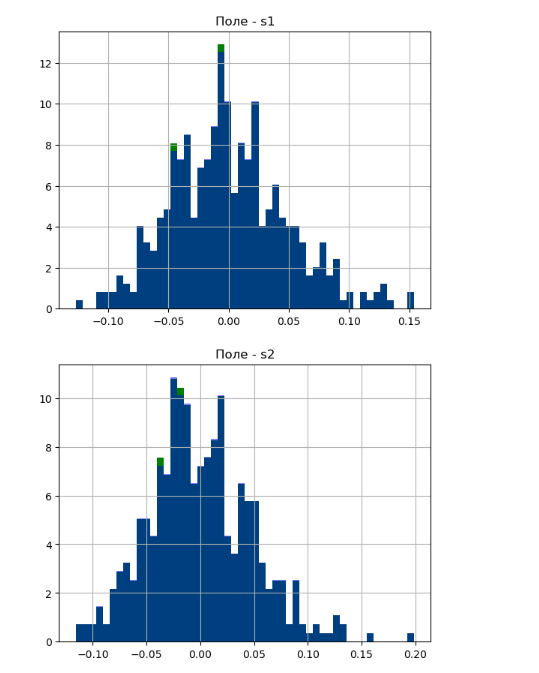
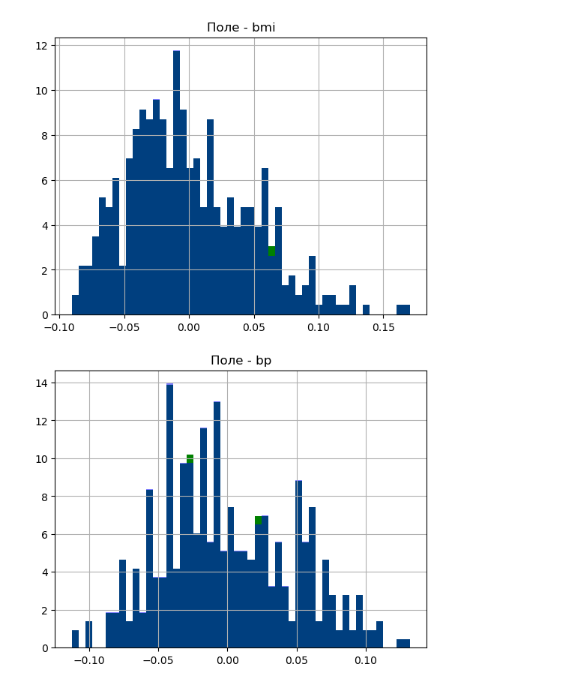
* Простота реализации.
* При случайном распределении пропусков сохраняются параметры распределения исходных данных.

**Недостатки:**

* Может быть удален большой фрагмент данных при неудачном распределении пропусков в нескольких столбцах.
* Если пропуски распределены не случайно, то можно удалить значимые данные.

Устранить различия гистограмм с противоположными значениями для указанных столбцов в разных наборах данных



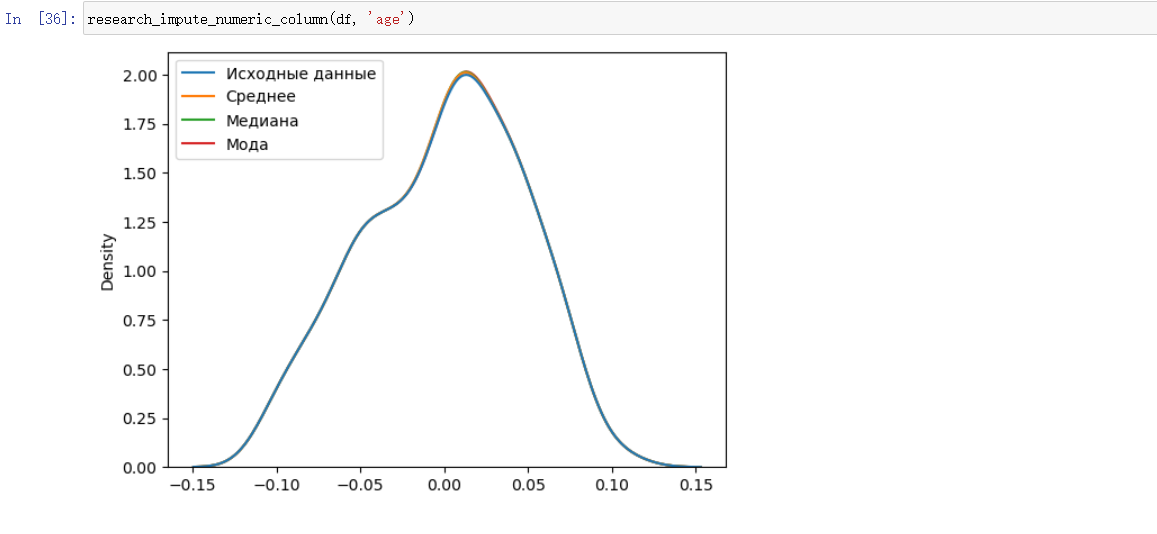
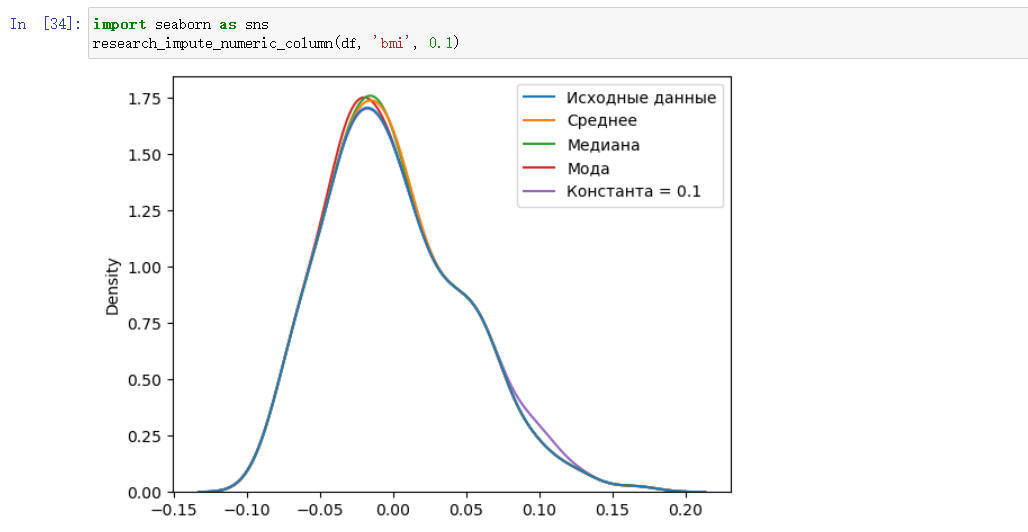


## 

## Заполнение значений для одного признака

### Заполнение [показателями центра распределения](https://ru.wikipedia.org/wiki/%D0%9F%D0%BE%D0%BA%D0%B0%D0%B7%D0%B0%D1%82%D0%B5%D0%BB%D0%B8_%D1%86%D0%B5%D0%BD%D1%82%D1%80%D0%B0_%D1%80%D0%B0%D1%81%D0%BF%D1%80%D0%B5%D0%B4%D0%B5%D0%BB%D0%B5%D0%BD%D0%B8%D1%8F) и константой

Используя четыре стратегии заполнения: «среднее», «медиана», «наиболее часто встречающаяся» и «константа», нарисуйте диаграмму оценки плотности ядра для каждого столбца данных, чтобы показать влияние различных стратегий заполнения на распределение данных.



### Заполнение "хвостом распределения"

# Заполнение наиболее распространенным значением категории

### 

# Введение отдельного значения категории для пропущенных значений

### 

# Добавления флага пропусков

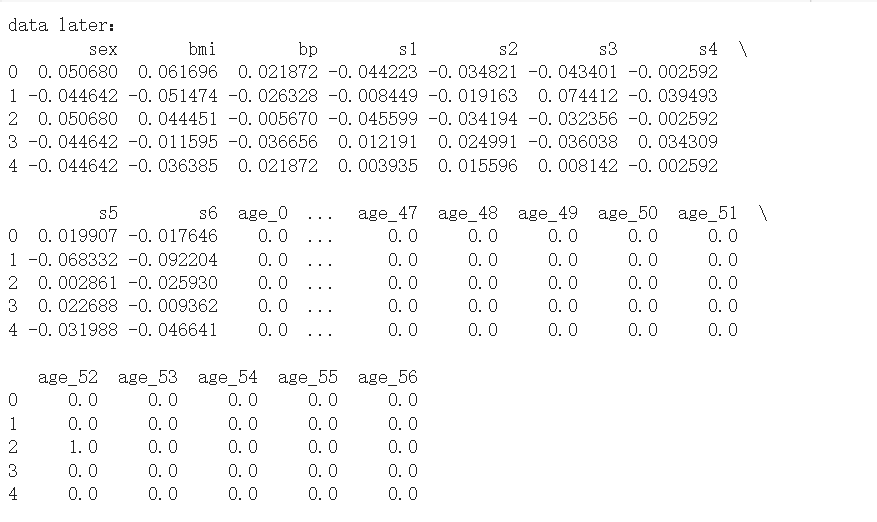
### 

# Заполнение значений для нескольких признаков Импьютация с использованием метода ближайших соседей

## Метод MICE (multivariate Imputation of Chained Equations)

### Кодирование категориальных признаков

Для набора данных по обнаружению диабета все функции являются числовыми и не имеют очевидной последовательной связи, поэтому более подходящим является One-Hot Encoding.

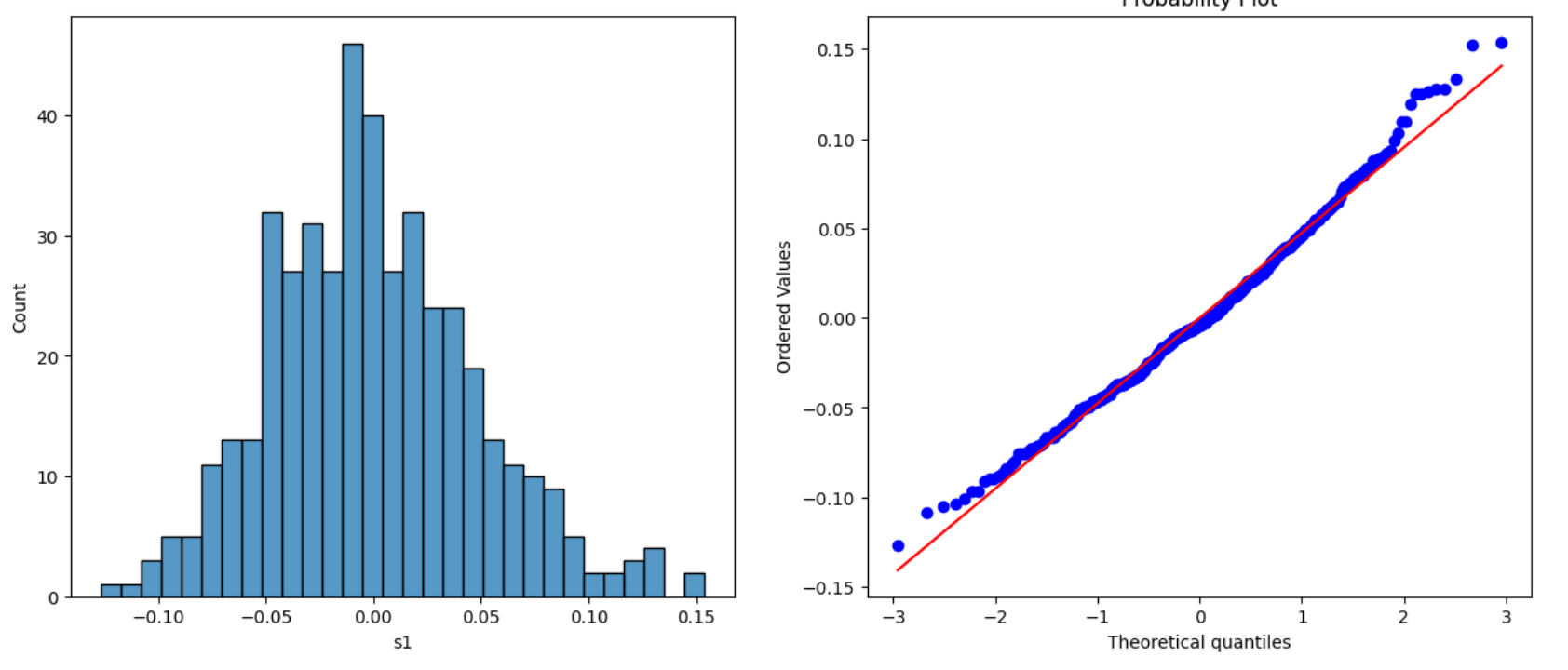
### Нормализация числовых признаков

Нормализация числовых признаков Нормализация числового признака предполагает что на основе существующего признака мы создаем новый признак, который в идеале имеет нормальное распределение.

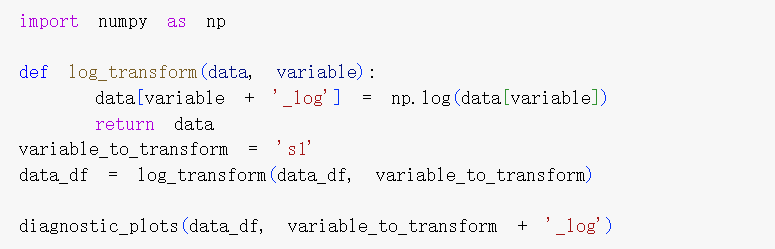
Многие модели машинного обучения лучше работают с нормализованными признаками, особенно линейные модели.

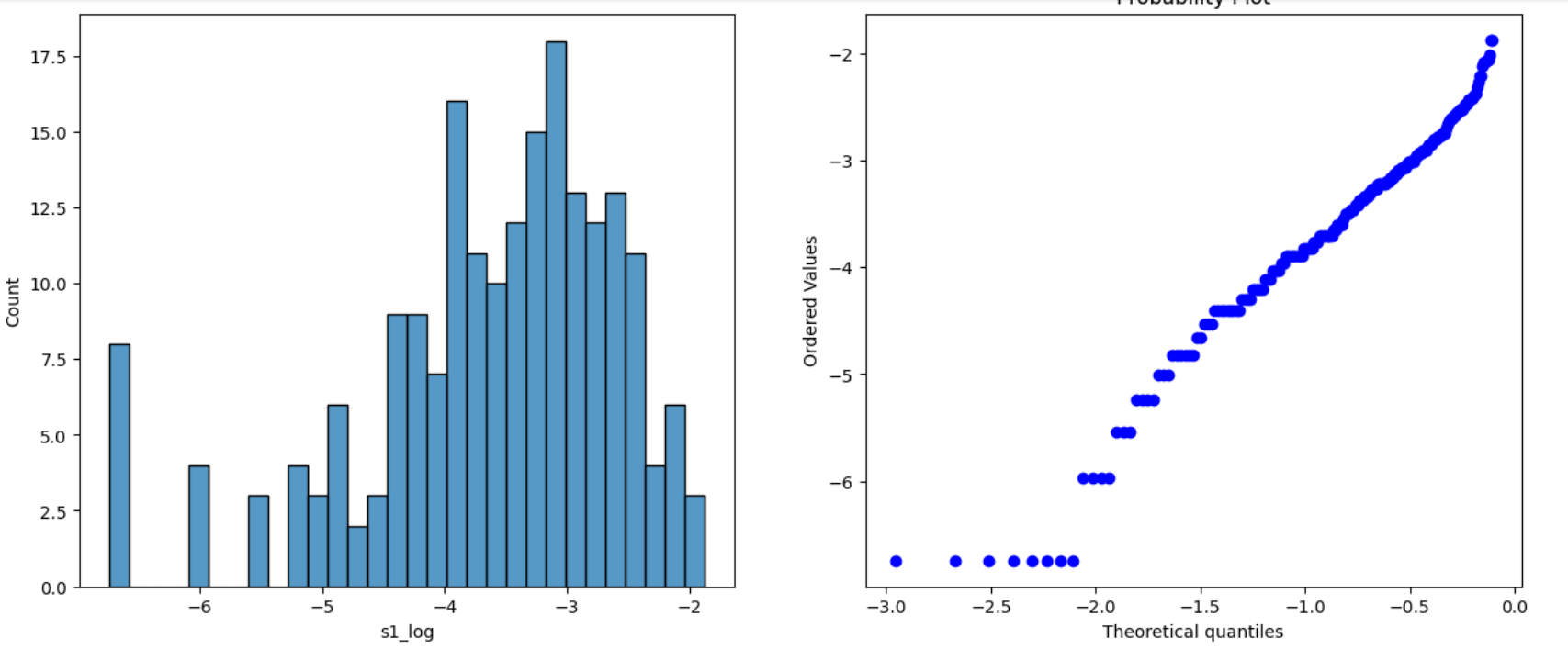
Какие функции чаще всего используются для нормализации:

логарифм - np.log(X) обратная зависимость - 1 / X квадратный корень возведение в степень преобразование Бокса-Кокса (Box-Cox transformation) преобразование Йео-Джонсона (Yeo-Johnson transformation) - расширение метода Бокса-Кокса для работы с произвольными величинами

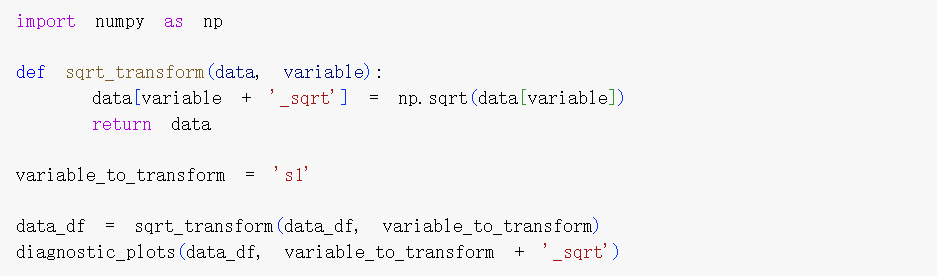
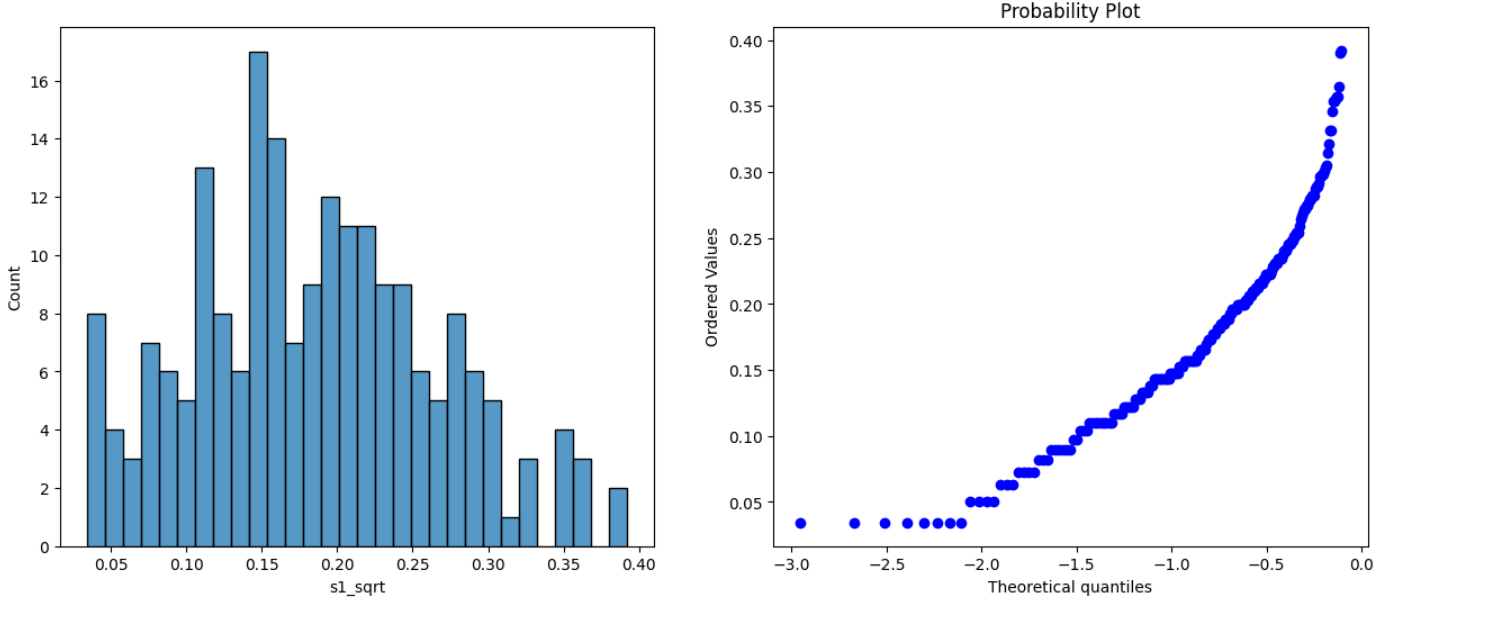
 

Логарифмическое преобразование. Не очень хороший результат.

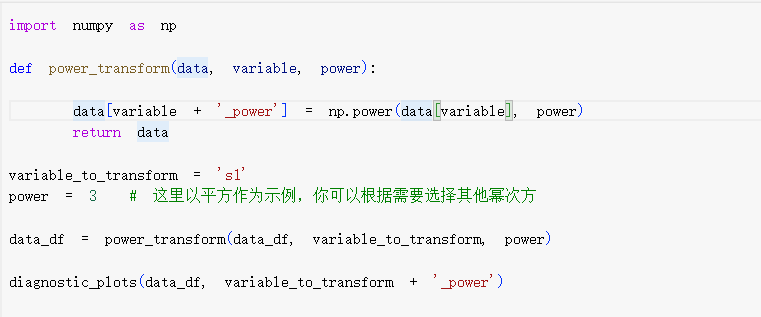


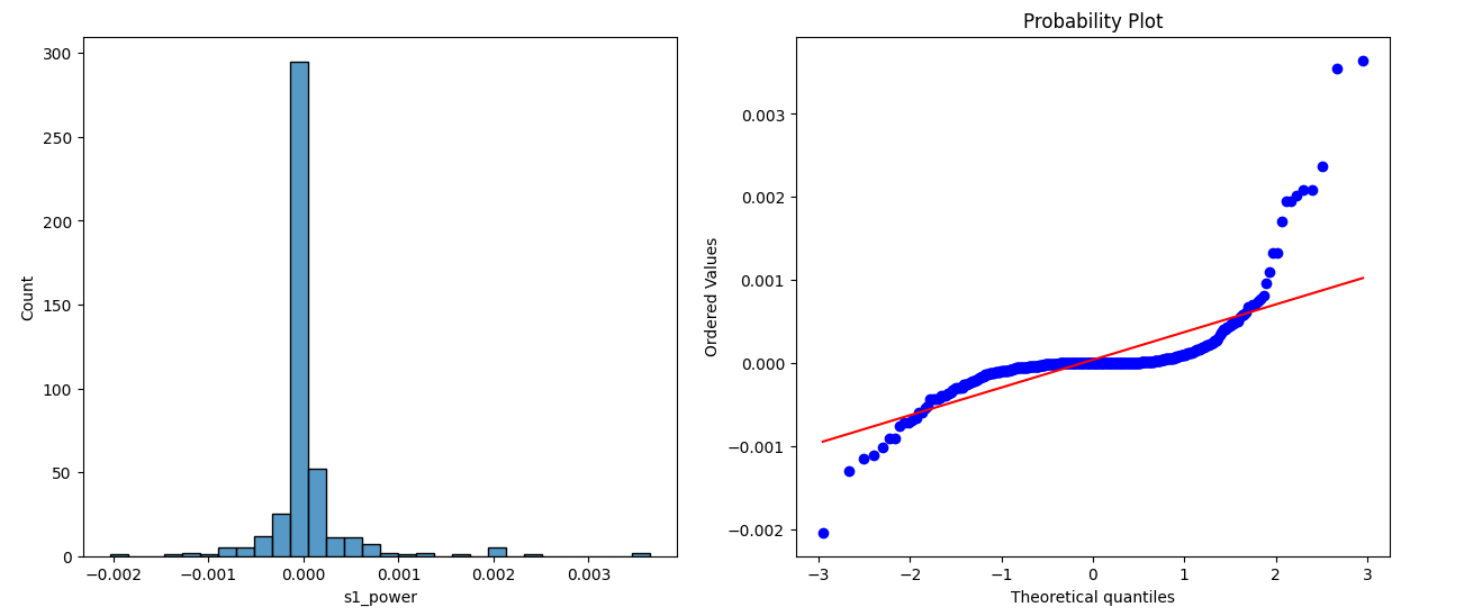


Квадратный корень

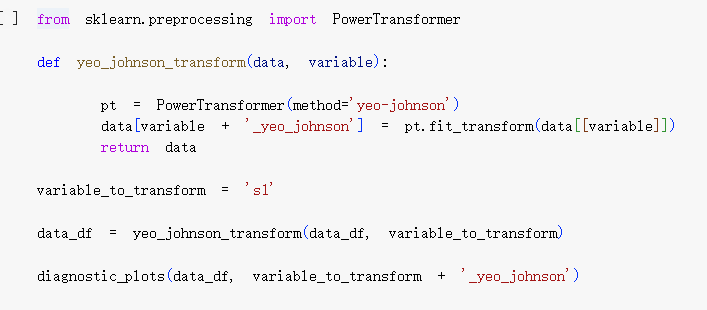
 

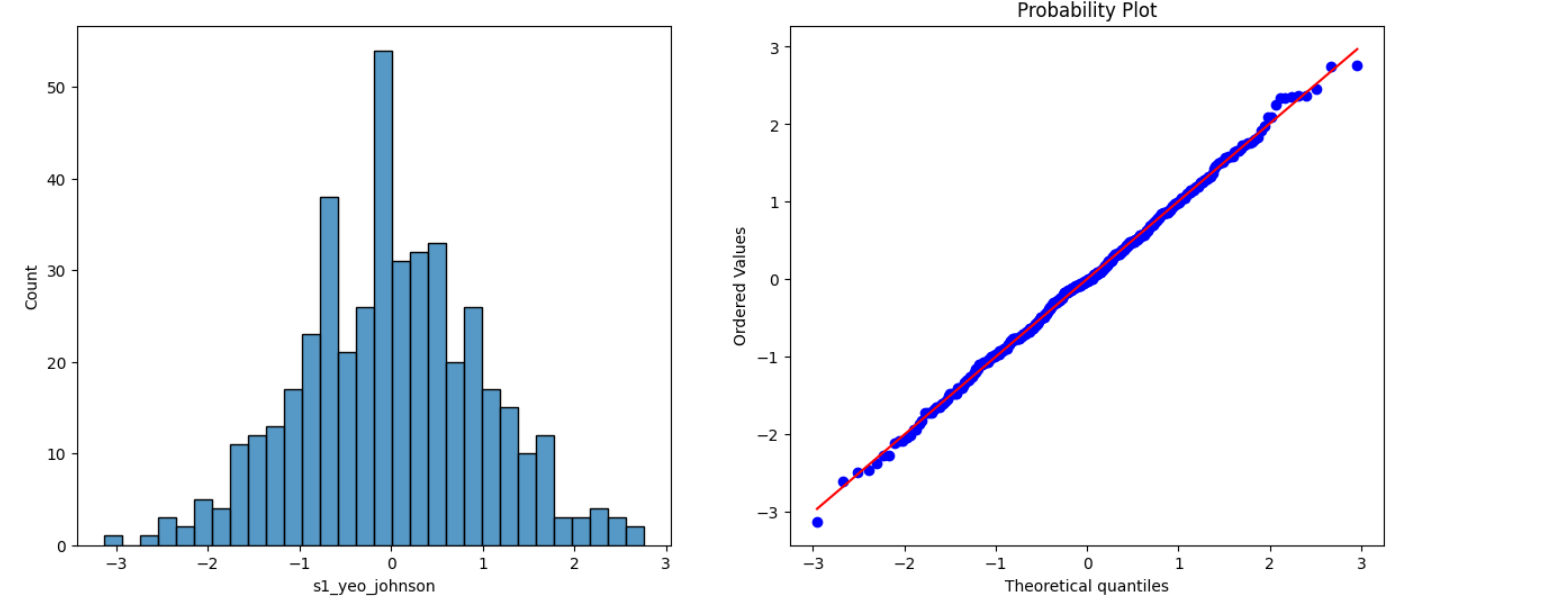
Возведение в степень





Преобразование Йео-Джонсона.Очень хороший результат.





# Вывод

При обработке набора данных о диабете в Scikit-learn мы реализовали ряд шагов предварительной обработки данных. Во-первых, мы устраняем пробелы между данными и повышаем стабильность модели, нормализуя числовые характеристики, чтобы гарантировать, что они имеют одинаковые масштабы. Затем мы закодировали категориальные признаки в числовую форму, приемлемую для модели для эффективной обработки. Наконец, мы стандартизировали числовые характеристики и скорректировали их среднее значение до 0, а дисперсию до 1, чтобы устранить различия в размерах между различными функциями и дополнительно оптимизировать производительность модели. Эти этапы предварительной обработки данных закладывают прочную основу для последующей работы по моделированию и повышают надежность и способность модели к обобщению.