|  |
| --- |
| Пензенский государственный университет  Факультет вычислительной техники  Кафедра «Вычислительная техника» |
| Отчет  по лабораторной работе №  по дисциплине «»  на тему «»  Вариант № 1 |
|  |
|  |
| Проверил:  Карамышева Н.С. |
| Пенза |

**Задание.**

1. Подготовить полученный набор данных в соответствии с этапами 3-й фазы методологии CRISP.

2. Описать использованные решения.

3. Организовать дополнительный сбор данных (если это необходимо и возможно).

4. Оформить отчёт, содержащий информацию о выполнении пунктов 1-3 задания.

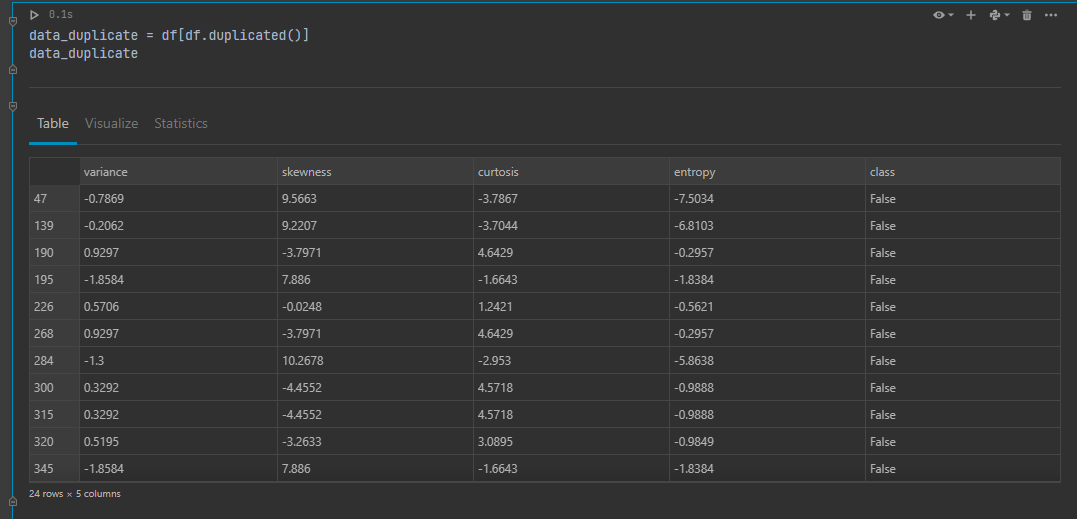
**Ход работы**

В ходе предыдущего анализа выявлены следующие особенности датасета:

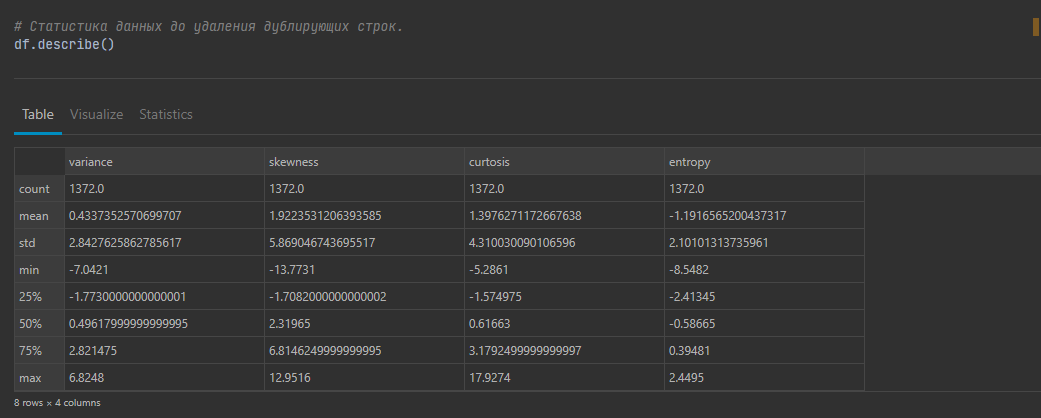
* Присутствие дубликатов.
* Отсутствие пропусков.
* Высокая корреляция по признакам curtosis и skewness.
* Отсутствие категориальных признаков.

В ходе выполнения этого лабораторного задания нам нужно удалить дубликаты (желательно, хотя иногда дубликаты могут пригодится), а также нормализовать остальные значения. Возможно, удалить один из столбцов curtosis или skewness из-за высокой корреляции (0.78).

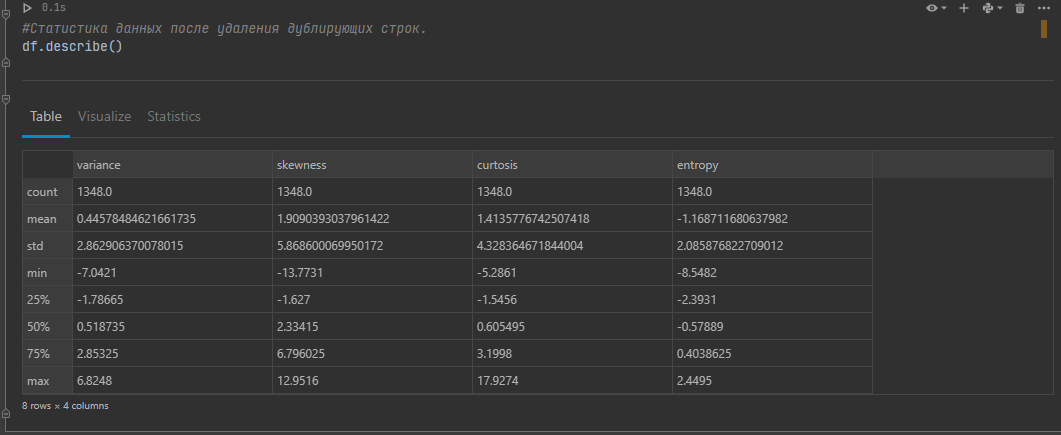
С помощью метода pd.duplicated определим дублирующие строки.



Удалим их с помощью метода drop\_duplicates.

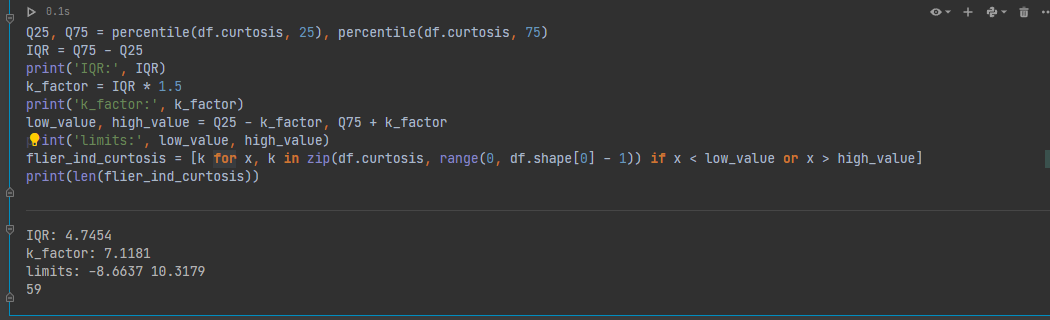
Статистические характеристики данных до преобразования. 

После

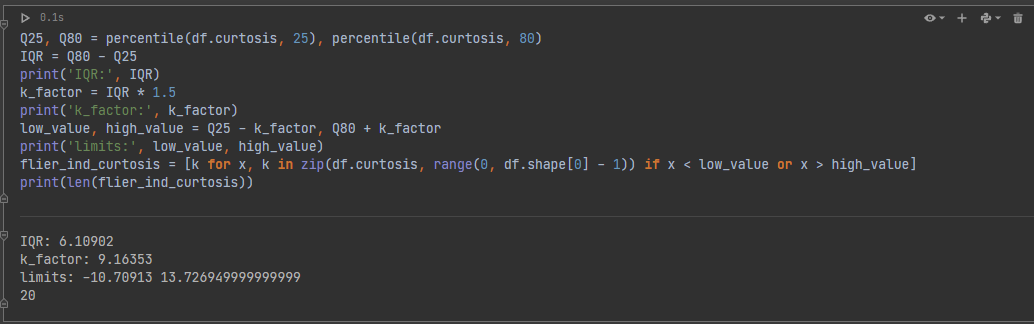


Произошли только незначительные изменения.

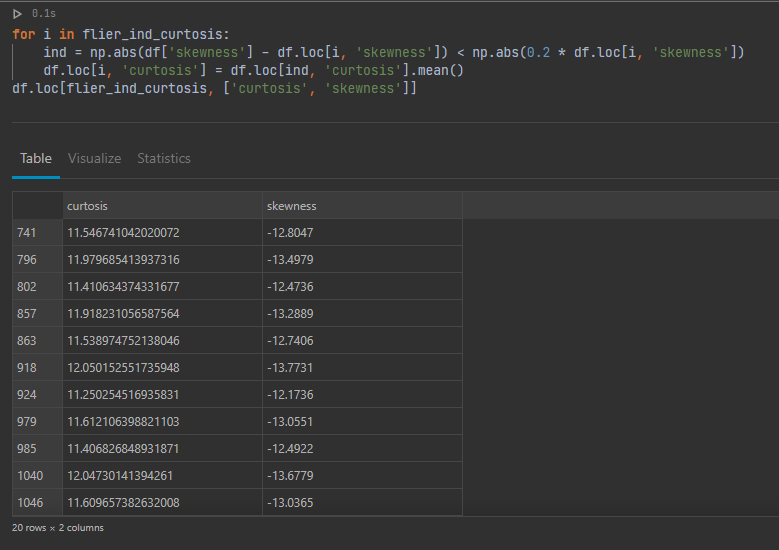
Теперь избавимся от выбросов перед этим проанализировав их количество.



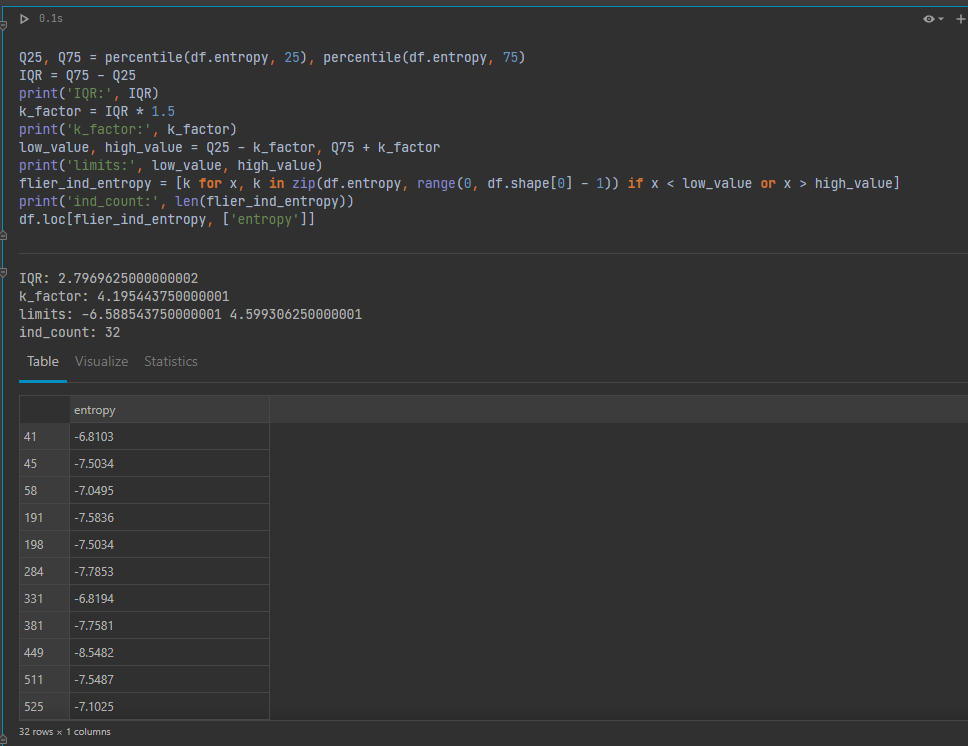
Всего 59 строк, которые находятся за пределами значения 10 по столбцу curtosis. Мы приняли решение немного поднять верхнюю перцентиль.



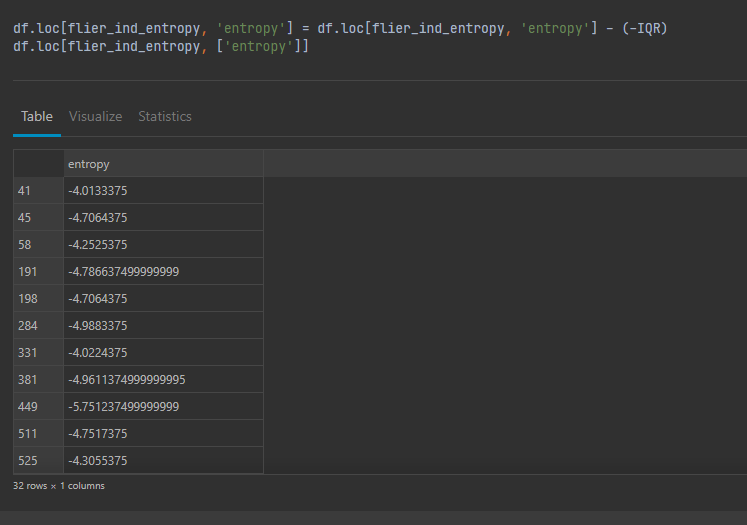
Заполним значениями в среднем отличающиеся на 20% по столбцу skewness, т.к между ними есть отрицательная корреляция.



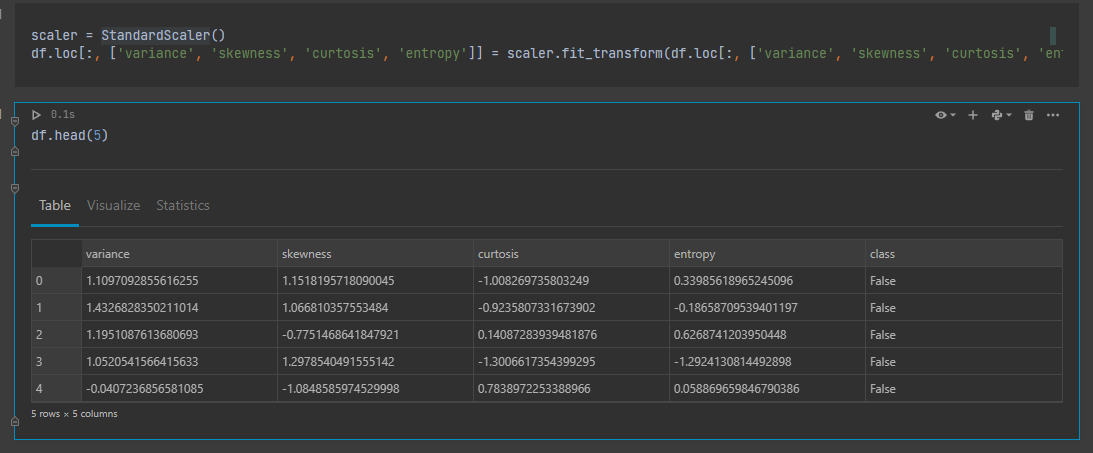
Теперь отследим выбросы в столбце entropy.



Просто вычтем из каждого выброса значение -IQR = -2. Таким образом, усечем значения.



Последним этапом нормализуем значения с помощью методы StandardScaler().



Сохраняем таблицу в csv.

