|  |
| --- |
| Пензенский государственный университет  Факультет вычислительной техники  Кафедра «Вычислительная техника» |
| Отчет  по лабораторной работе №  по дисциплине «Программные средства кибертехнических систем»  на тему «Подготовка набора данных»  Вариант № 6 |
|  |
|  |
| выполнили: ст -ты гр. 22ВВП1  Захаров А. С.  Беляев Д. И.  Демин М. С  Проверили:  Карамышева Н.С.  Зинкин С.А. |
| 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.

Вариант 6. Задача кредитного скоринга (15-й столбец набора данных {1 – кредит выдан, 2 – кредит не выдан}) по 14 признакам.

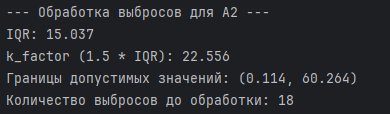
**Ход работы**

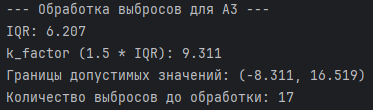
В ходе предыдущего анализа выявлены следующие особенности датасета:

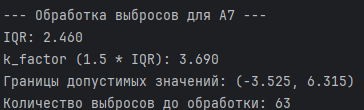
* Отсутствие пропусков.
* Низкая корреляция числовых признаков.
* Отсутствие категориальных признаков.
* Выбросы в числовых признаках

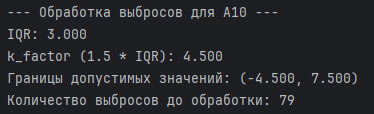
В ходе выполнения этого лабораторного задания нам нужно нормализовать числовые значения. Возможно, удалить один из столбцов A2, A3, A7, A10 из-за низкой корреляции (<0.4).

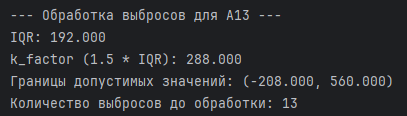
Избавимся от выбросов перед этим проанализировав их количество.

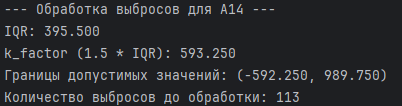






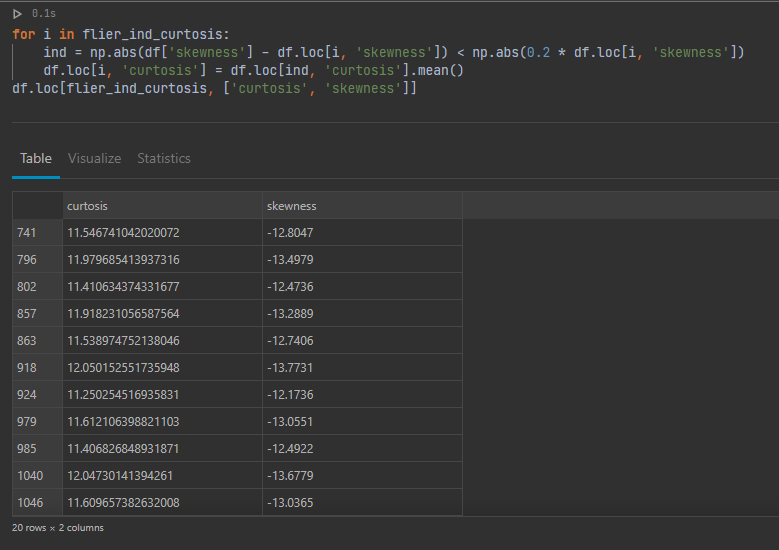




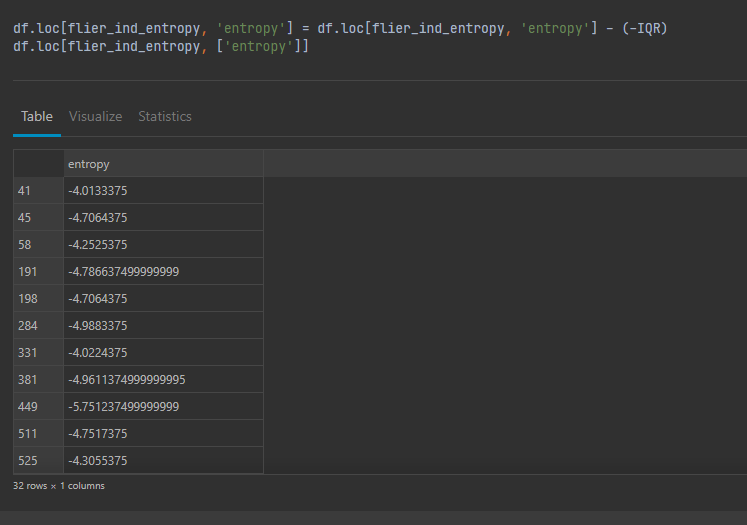


Мы решили сразу проанализировать все столбцы, в которых были выбросы.

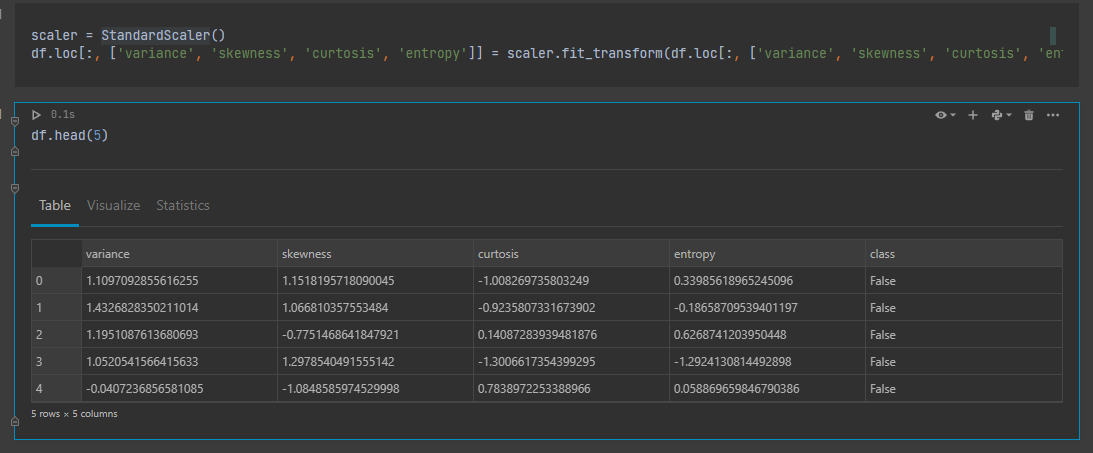
Заполним значениями в среднем отличающиеся на 10% по столбцам , т.к между ними есть отрицательная корреляция.



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



Последним этапом нормализуем значения с помощью методы StandardScaler().



Сохраняем таблицу в csv.

