Министерство науки и высшего образования Российской Федерации  
Пензенский государственный университет  
Кафедра вычислительная техника

**ОТЧЕТ**

по лабораторной работе №6

по дисциплине «Программные средства кибертехнических систем»

на тему «Подготовка набора данных

в соответствии с третьей фазой методологии

CRISP-DM: обработка выбросов и пропусков,

масштабирование данных, преобразование

категориальных переменных»

Выполнили:

студенты группы 22ВВП1

Хоссейни Нежад С.А.С.М.

Амиров И.Р.

Сергунов М. Р.

Приняли:

Зинкин С.А.

Карамышева Н.С.

Пенза 2025

**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**

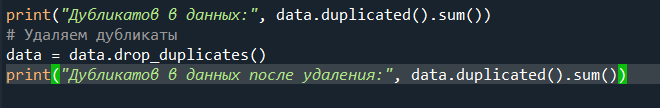
В ходе предыдущего анализа выявлены следующие особенности датасета:

* Присутствие дубликатов.
* Отсутствие пропусков.
* Отсутствие категориальных признаков.

В ходе выполнения этого лабораторного задания нам нужно удалить дубликаты (желательно, хотя иногда дубликаты могут пригодится), а также нормализовать остальные значения.

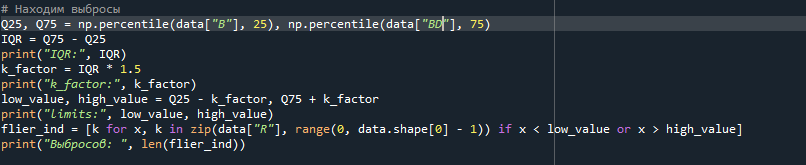
С помощью метода data.duplicated определим дублирующие строки.

Удалили дубликаты





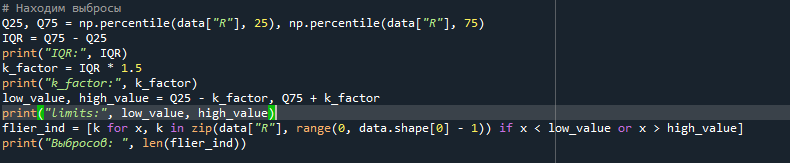
Теперь избавимся от выбросов перед этим проанализировав их количество.

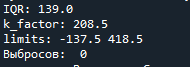




Выбросов нету в столбце “B”,

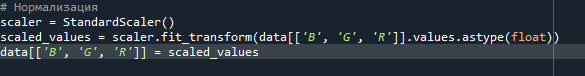
Теперь отследим выбросы в столбце “R”.

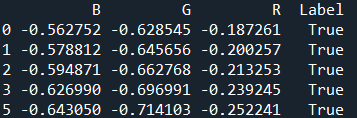




Также видим, что в столбце “R” выбросов нет.

Последним этапом нормализуем значения с помощью методы StandardScaler().





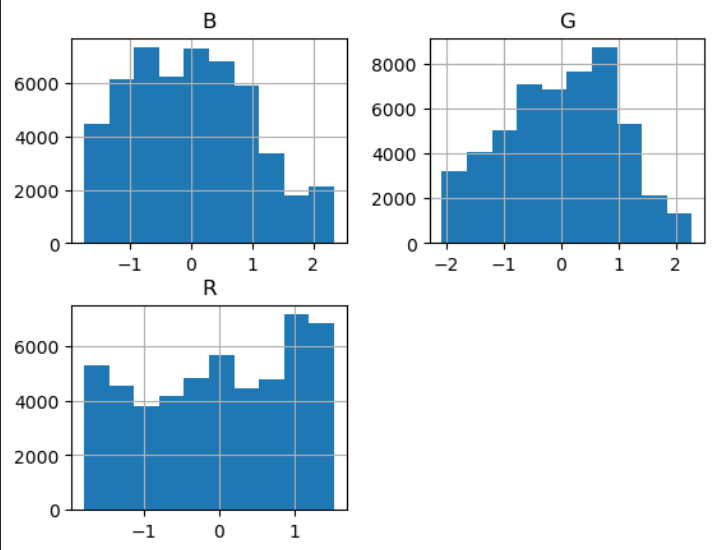
Сохраняем таблицу в csv.





Вывели гистограмму:





**Вывод**

В результате выполнения лабораторной работы изучили как удалять дубликаты, выявлять выбросы, а также нормализовать значения.