Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования

**Иркутский национальный исследовательский   
технический университет**

|  |
| --- |
| Институт информационных технологий и анализа данных |
| наименование института |

|  |
| --- |
| Отчет по дисциплине |
| «Методы анализа данных» |
| по теме: |
| «Подготовка данных для анализа» |
|  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Выполнил студент группы |  | АСУб-20-1 |  |  |  | Устюжанин В. М. |
|  |  | Шифр группы |  | Подпись |  | И.О. Фамилия |
| Проверил преподаватель |  |  |  |  |  | Осипова Е. А. |
|  |  |  |  | Подпись |  | И.О. Фамилия |

Иркутск 2022 г.

**Содержание**

[Постановка задачи 3](#_Toc117630056)

[Выбор набора данных 4](#_Toc117630057)

[Подготовка данных для анализа 5](#_Toc117630058)

[Вывод 10](#_Toc117630059)

# Постановка задачи

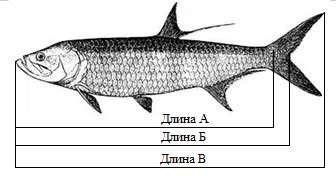
1. Выбрать среду программирования для языка Python. Используя в качестве источника данных список предложенных дата сетов выбрат044C набор данных.
2. С использованием Python выполнить подготовку данных для анализа. При этом необходимо:
   1. описать исходные данные;
   2. выполнить анализ типов шкал, в которых измерены данные (количественные, качественные (номинальные, дихотомические, порядковые));
   3. провести проверку логического соответствия данных;
   4. выявить наличие пропущенных значений.
   5. для выявления выбросов для каждой пары числовых признаков, построить диаграммы Кливленда (достаточно 4 признаков). Выделить наблюдения, похожие на выбросы;
   6. получить значения основных показателей описательной статистики для признаков, измеренных в количественной шкале;
   7. построить диаграмму размаха для каждого признака;
   8. выполнить нормализацию данных и повторно построить диаграмму размаха для каждого признака;
   9. для каждой пары числовых признаков построить диаграмму рассеяния.
3. Оформить отчет, в который включить все результаты анализа с описанием результатов.

# Выбор набора данных

Для анализа был выбран набор данных Fish.csv.

Данные, предоставленные в данном дата сете.

* Species – Порядковый номер вида, количественный тип данных.
* Weight – Вес рыбы, количественный тип данных.
* Lenght1 – Длина рыбы А, количественный тип данных.
* Lenght2 – Длина рыбы Б, количественный тип данных.
* Lenght3 – Длина рыбы В, количественный тип данных.
* Height – Высота рыбы, количественный тип данных.
* Width – Ширина рыбы, количественный тип данных.



# Подготовка данных для анализа

Подключаем библиотеки, необходимые для выполнения работы.

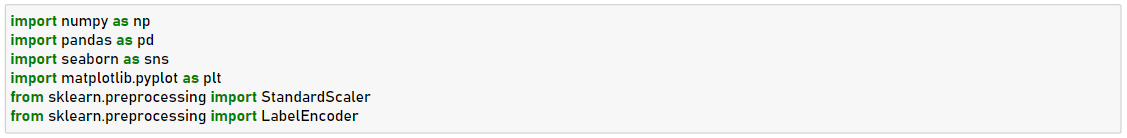


Рисунок 1 – Подключение библиотек в Python

Производим загрузку данных.

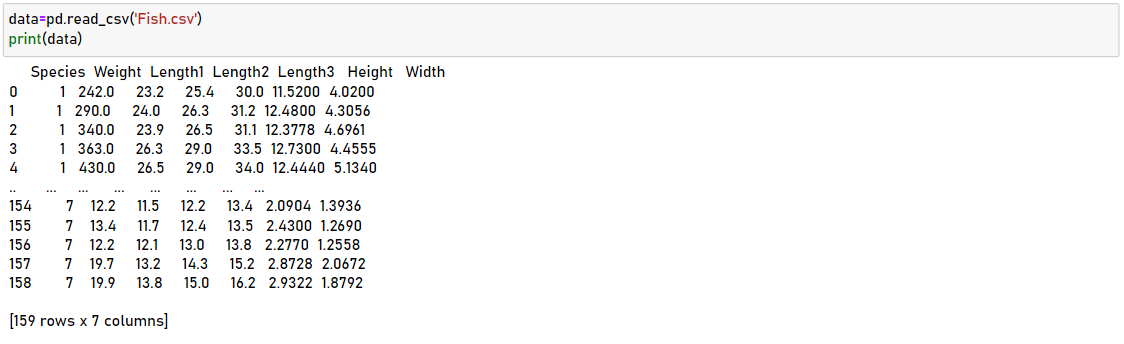


Рисунок 2 – Загрузка дата сета

Производим проверку на пропущенные значения.

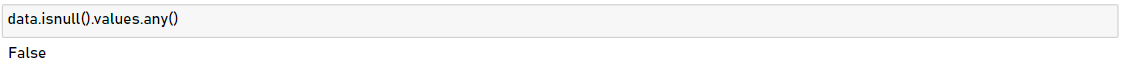


Рисунок 3 – Проверка на пропущенные данные

Пропущенные значения не обнаружены. Проверим значения, на соответствия типам данных.

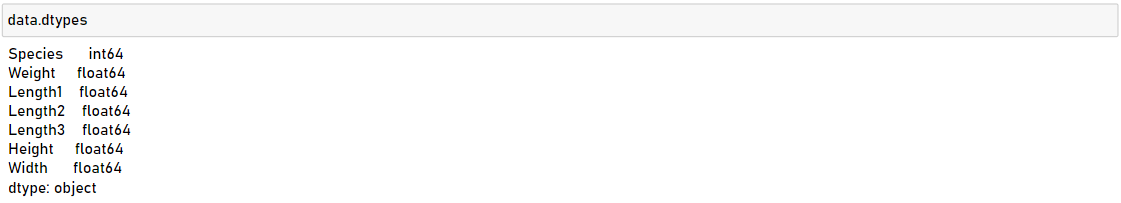


Рисунок 4 – Проверка типов данных

Типы данных соответствуют своим заявленным значениям. Получим значения основных показателей описательной статистики.

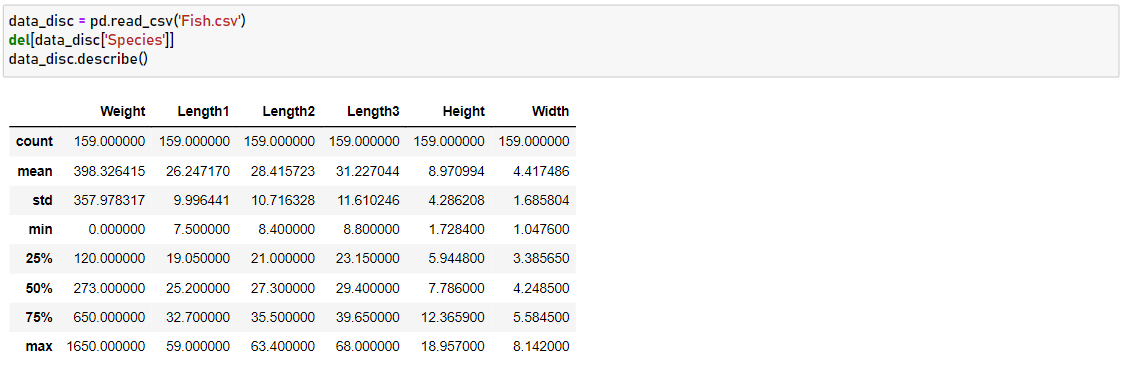


Рисунок 5 – Основные показатели описательной статистики

Произведем построение диаграмм Кливленда для трех пар количественных значений.

№1 – Weight – Lenght3;

№2 – Lenght3 – Width;

№3 – Weight – Width;

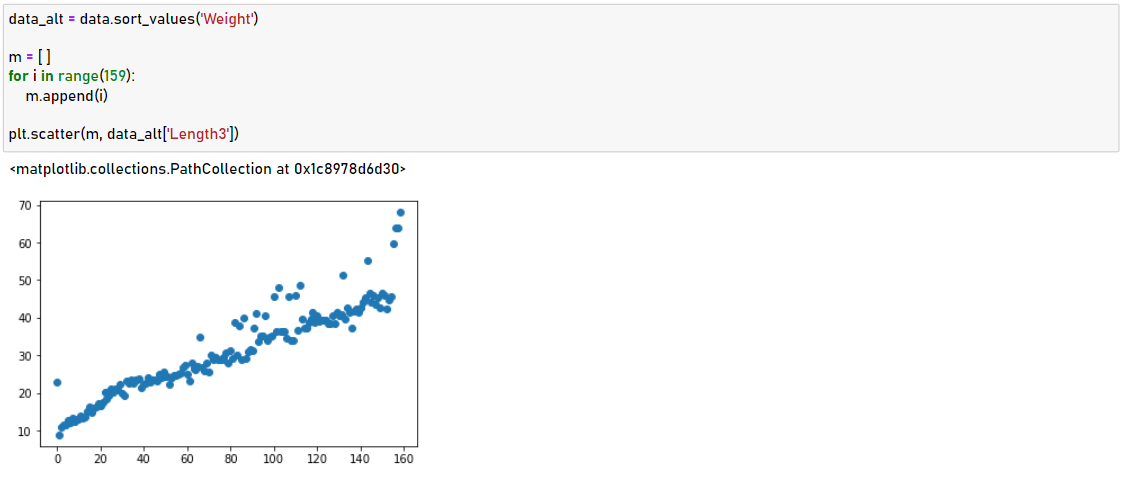


Рисунок 6 – Диаграмма Кливленда для пары признаков №1

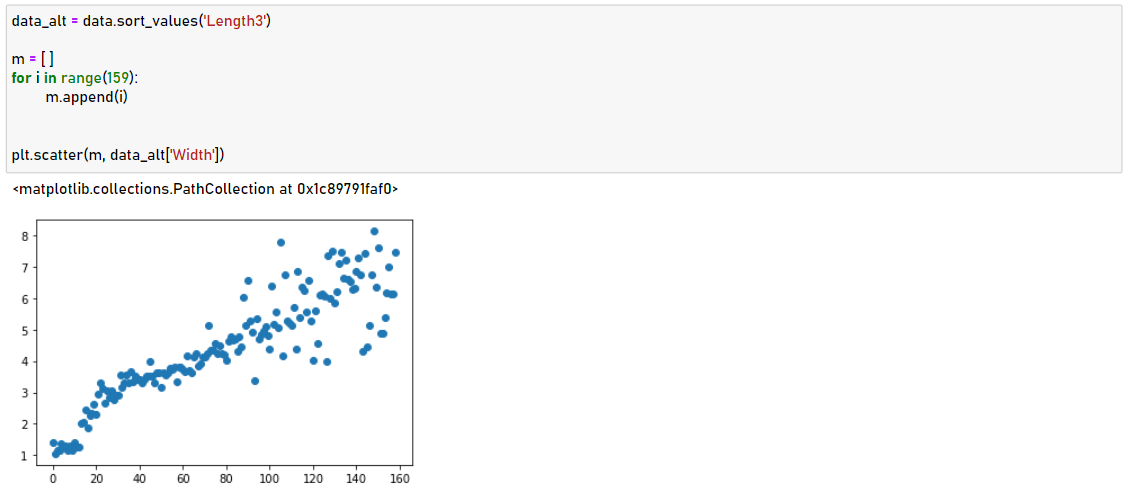


Рисунок 7 – Диаграмма Кливленда для пары признаков №2

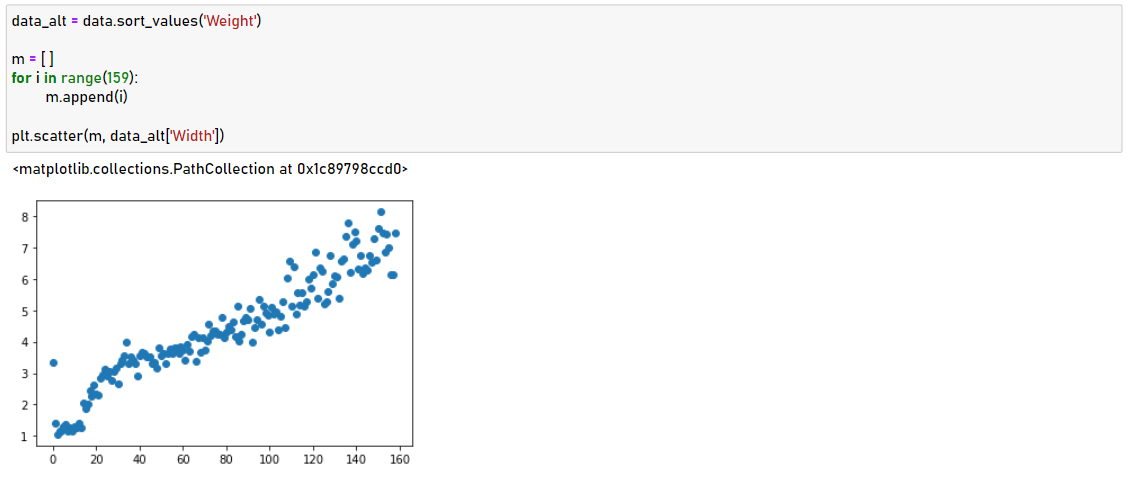


Рисунок 8 – Диаграмма Кливленда для пары признаков №3

Произведем построение гистограмм частот.

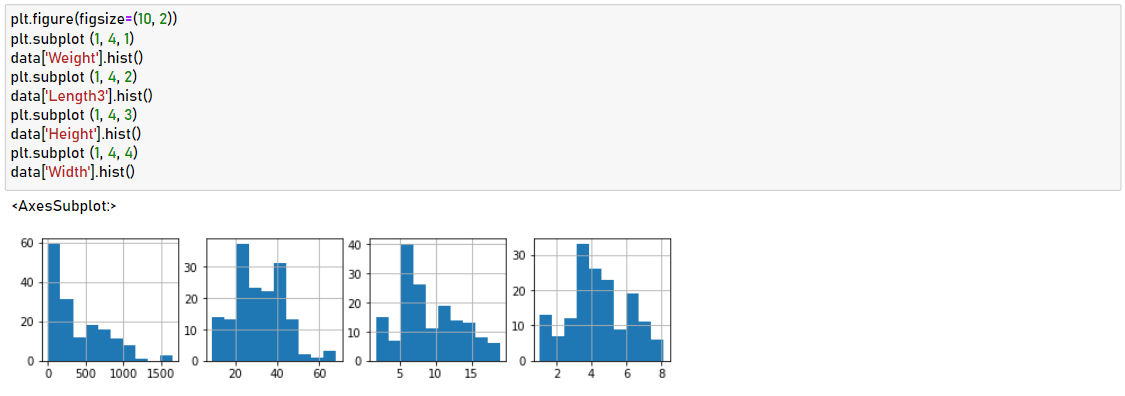


Рисунок 9 – Гистограммы частот

Построим диаграмму размаха для всех столбцов.



Рисунок 10 – Диаграмма размаха

Выполним стандартизацию и повторно построим диаграмму размаха.

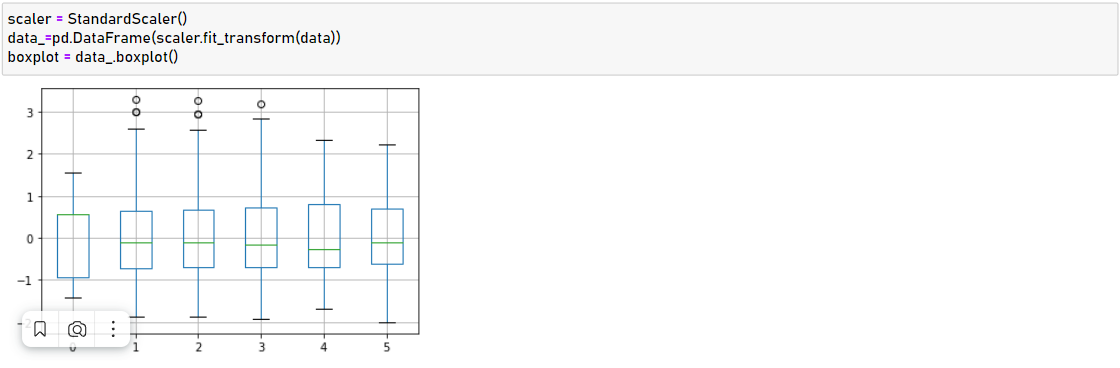


Рисунок 11 – Стандартизированная диаграмма размаха

Произведем удаление выбросов из столбцов и повторно построим диаграмму размаха.

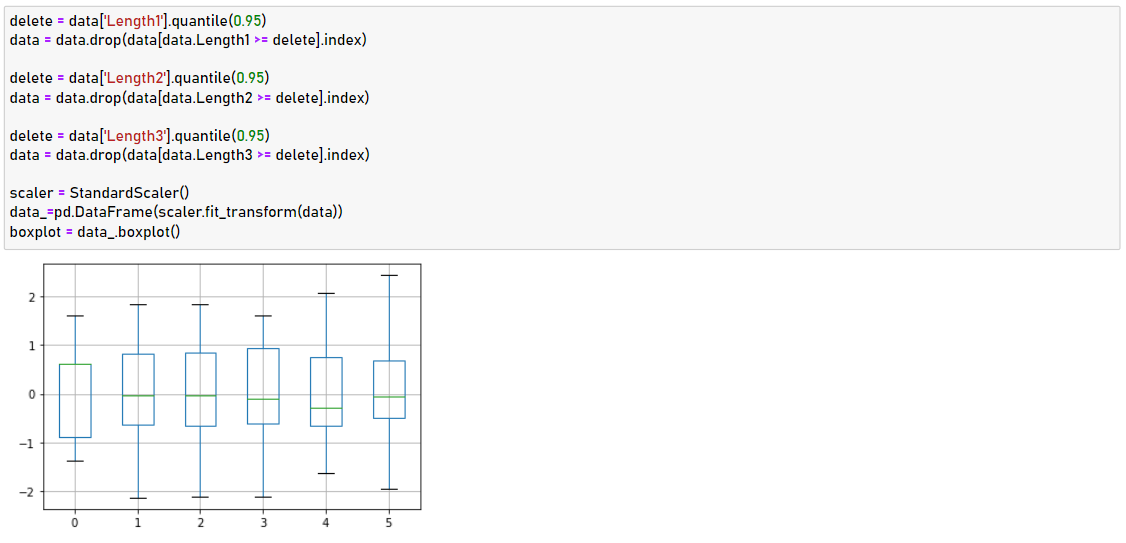


Рисунок 12 – Диаграмма размаха с удаленными выбросами

Для каждой пары количественных признаков построим диаграмму рассеяния.

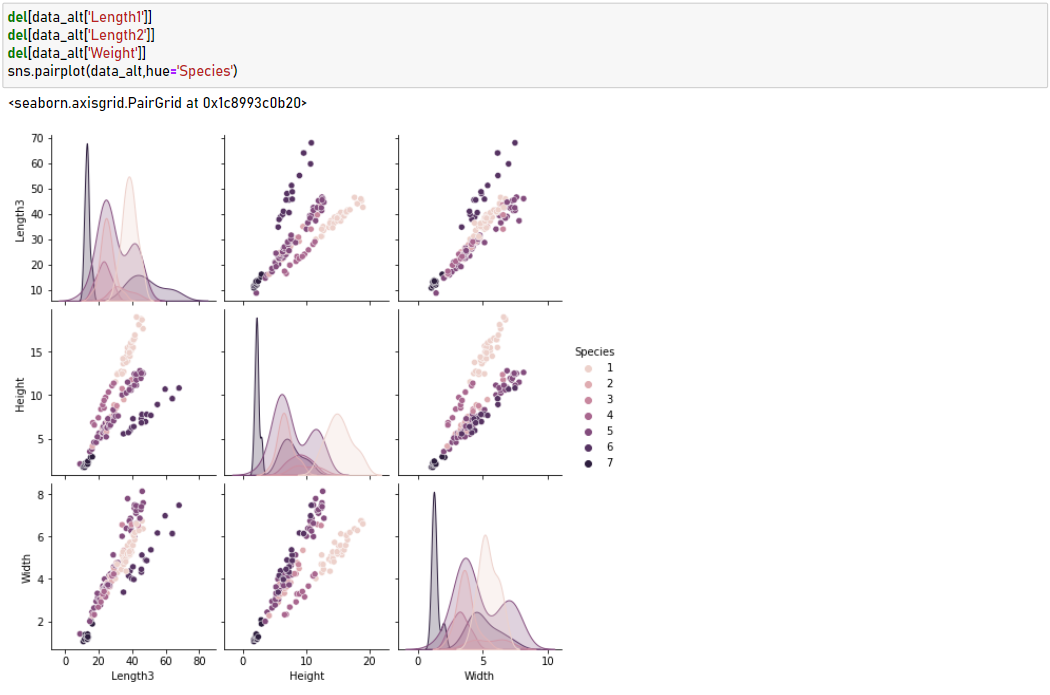


Рисунок 13 – Диаграмма рассеяния

# Вывод

В процессе выполнения лабораторной работы была рассмотрена выборка о значениях веса, трех длин, высоты, ширины для множества различных видов рыб. Построив диаграмму Кливленда, мы обнаружили выбросы для пар признаков, затем избавившись от ненужных данных, построили диаграмму размаха для всех столбцов, нормализовали ее, построили диаграмму рассеяния, которая позволила оценить меру и вид зависимости между признаками, а также выделить подмножества.