

Выполнил: Тимошинов Егор Борисович

Группа: 16

Лабораторная работа 1

Перспективные информационные технологии

Цель работы

Изучение методов описательного анализа данных, построения графиков и расчета статистических характеристик с использованием языка программирования Python.

Теоретические сведения

Описательная статистика позволяет получить общее представление о данных. Основные характеристики включают меры центральной тенденции (среднее, медиана), меры вариации (дисперсия, стандартное отклонение), а также меры формы распределения (асимметрия, эксцесс).

Графические методы визуализации данных помогают наглядно представить структуру данных и выявить закономерности. К основным типам графиков относятся гистограммы, ящики с усами, диаграммы рассеяния.

Результаты выполнения задания

1. Исходные данные

Для анализа использованы данные о 24 наблюдениях. Переменные: Y (зависимая переменная), X_1, X_2, X_3, X_4 (независимые переменные).

Первые 20 наблюдений:

№	Y	X1	X2	X3	X4
1	76.4	117.7	81.6	10.3	28.5
2	77.6	123.8	73.2	11.4	28.7
3	88.2	126.9	75.3	12.2	29.1
4	87.3	134.1	71.3	11.5	29.2
5	82.5	123.1	77.3	11.2	29.1
6	79.4	126.7	76.1	10.5	29.2
7	80.3	130.4	76.6	9.4	29.3
8	80.1	129.3	84.7	9.5	29.2

9	105.2	145.4	92.4	9.3	29.1
10	102.5	163.8	80.3	9.2	28.7
11	108.7	164.8	82.6	9.4	28.4
12	104.5	165.3	70.9	9.7	27.8
13	103.7	164.1	89.9	8.2	27.7
14	117.8	183.7	81.3	8.4	27.6
15	115.8	195.8	83.7	8.2	27.5
16	117.8	219.4	76.1	8.1	27.5
17	118.4	209.8	80.4	7.8	27.4
18	120.4	223.3	78.1	7.2	27.5
19	123.8	223.6	79.8	8.2	27.6
20	134.9	236.6	82.1	7.5	27.7

2. Структура данных

Количество наблюдений: 24

Количество переменных: 5

Типы данных: все переменные являются числовыми (float64)

3. Гистограммы

Для построения гистограммы переменной Y было рассчитано оптимальное количество интервалов по формуле Стерджеса:

$$k = 1 + 3.322 \log_{10}(n) = 6$$

где $n = 24$ — количество наблюдений.

Ширина интервала: $h = 16.0500$

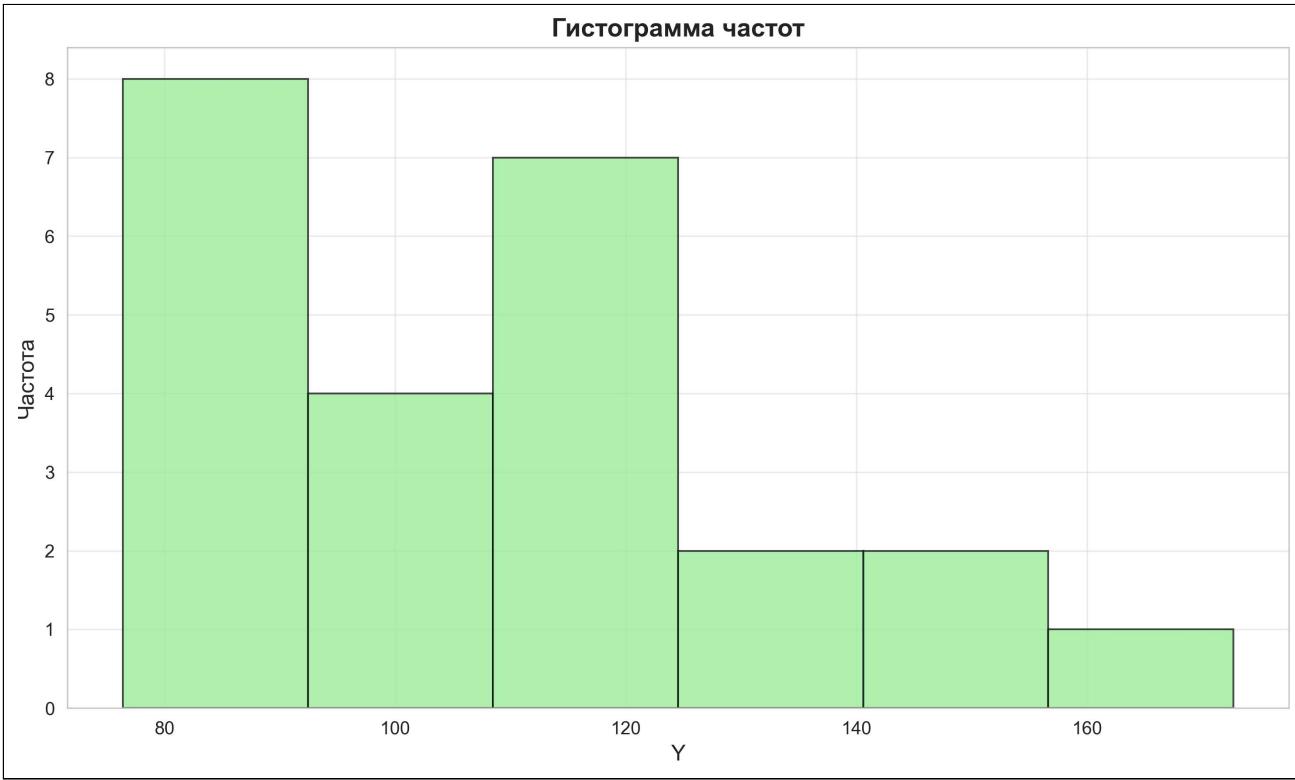


Рисунок 1. Гистограмма частот для переменной Y

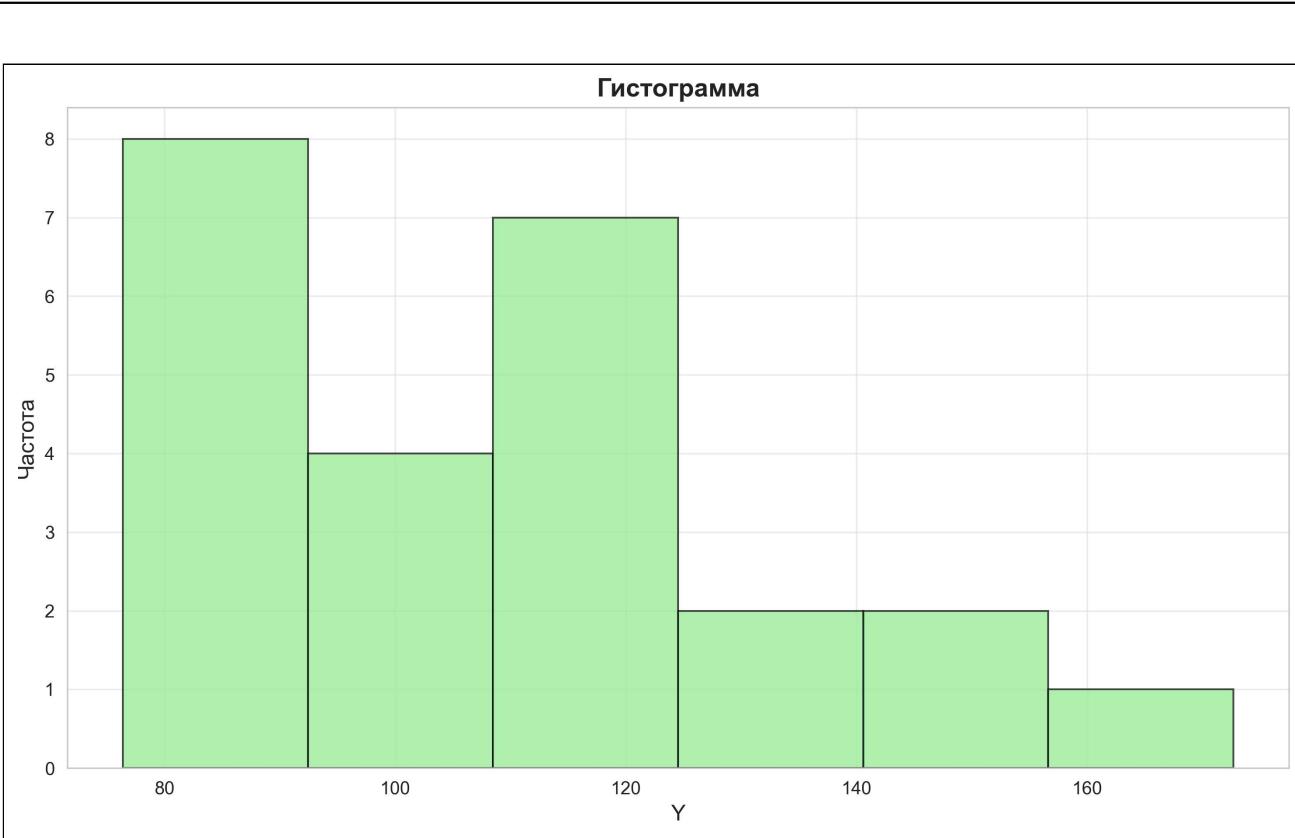


Рисунок 2. Гистограмма для переменной Y с заданными границами интервалов

4. Описательные статистики

Основные описательные статистики для всех переменных:

Переменная	Среднее	Медиана	Минимум	Максимум	Ст. отклонение	Дисперсия
Y	109.1500	106.9500	76.4000	172.7000	25.4030	645.3139
X1	179.1875	165.0500	117.7000	254.3000	48.3285	2335.6394
X2	80.2917	80.6000	70.9000	92.4000	5.3214	28.3173
X3	9.0333	8.8000	7.2000	12.2000	1.5104	2.2814
X4	28.3208	28.3500	27.4000	29.3000	0.6865	0.4713

Дополнительные характеристики для переменной Y:

- Среднее арифметическое: 109.1500
- Медиана: 106.9500
- Стандартное отклонение: 25.4030
- Дисперсия: 645.3139
- Размах: 96.3000
- Коэффициент вариации: 23.27%
- Коэффициент асимметрии: 0.6118
- Коэффициент эксцесса: 0.1154
- Стандартная ошибка среднего: 5.1854

Квантили переменной Y (с шагом 0.20):

Квантиль	Значение
0%	76.4000
20%	81.6200
40%	103.8600
60%	117.4000
80%	126.4800
100%	172.7000

99% доверительный интервал для среднего значения Y:

$$[94.5929, 123.7071]$$

5. Ящики с усами

Ящик с усами (box plot) позволяет визуализировать распределение данных, показать медиану, квартили и выбросы.

Ящик с усами для Y

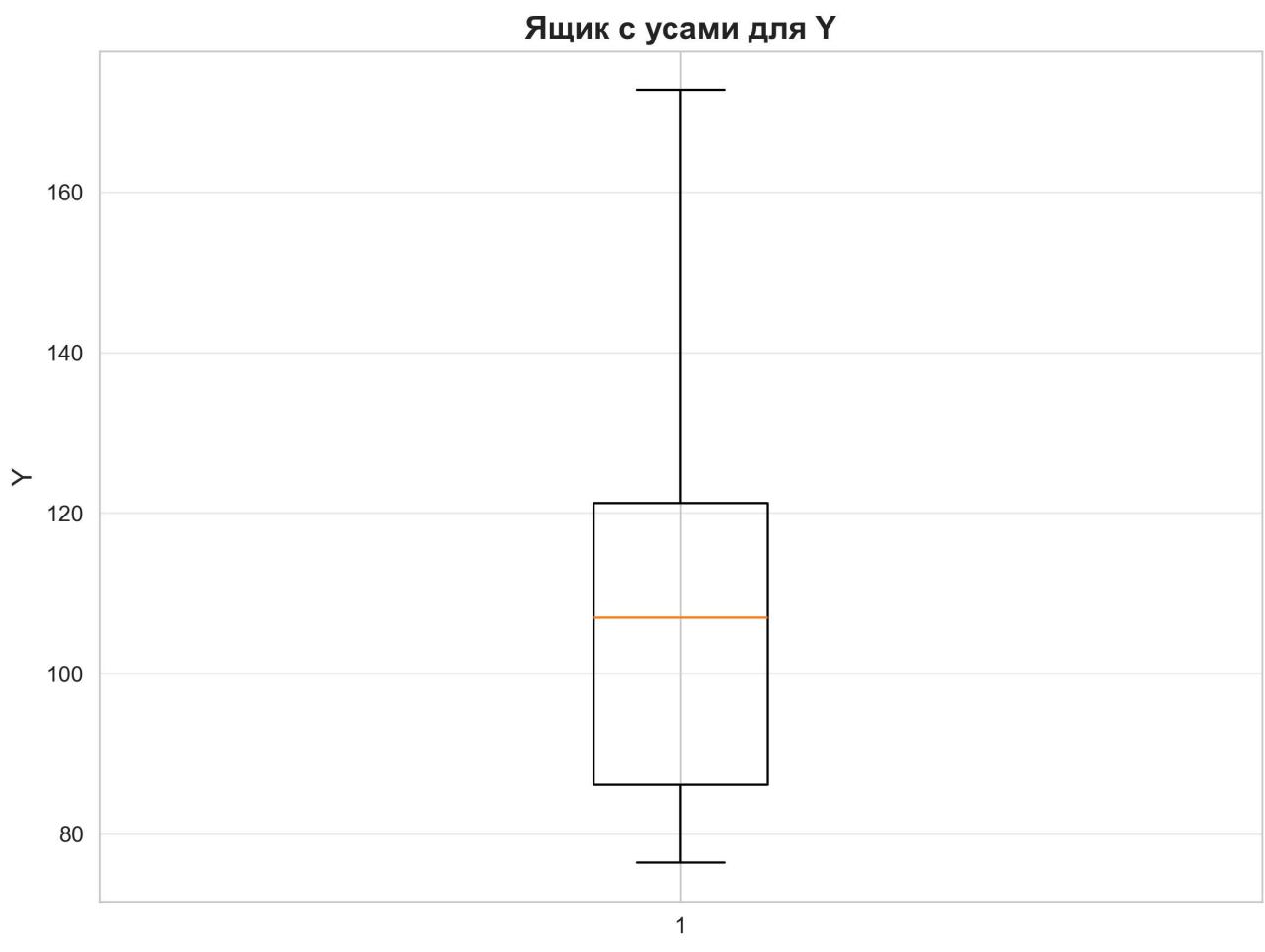


Рисунок 3. Ящик с усами для переменной Y

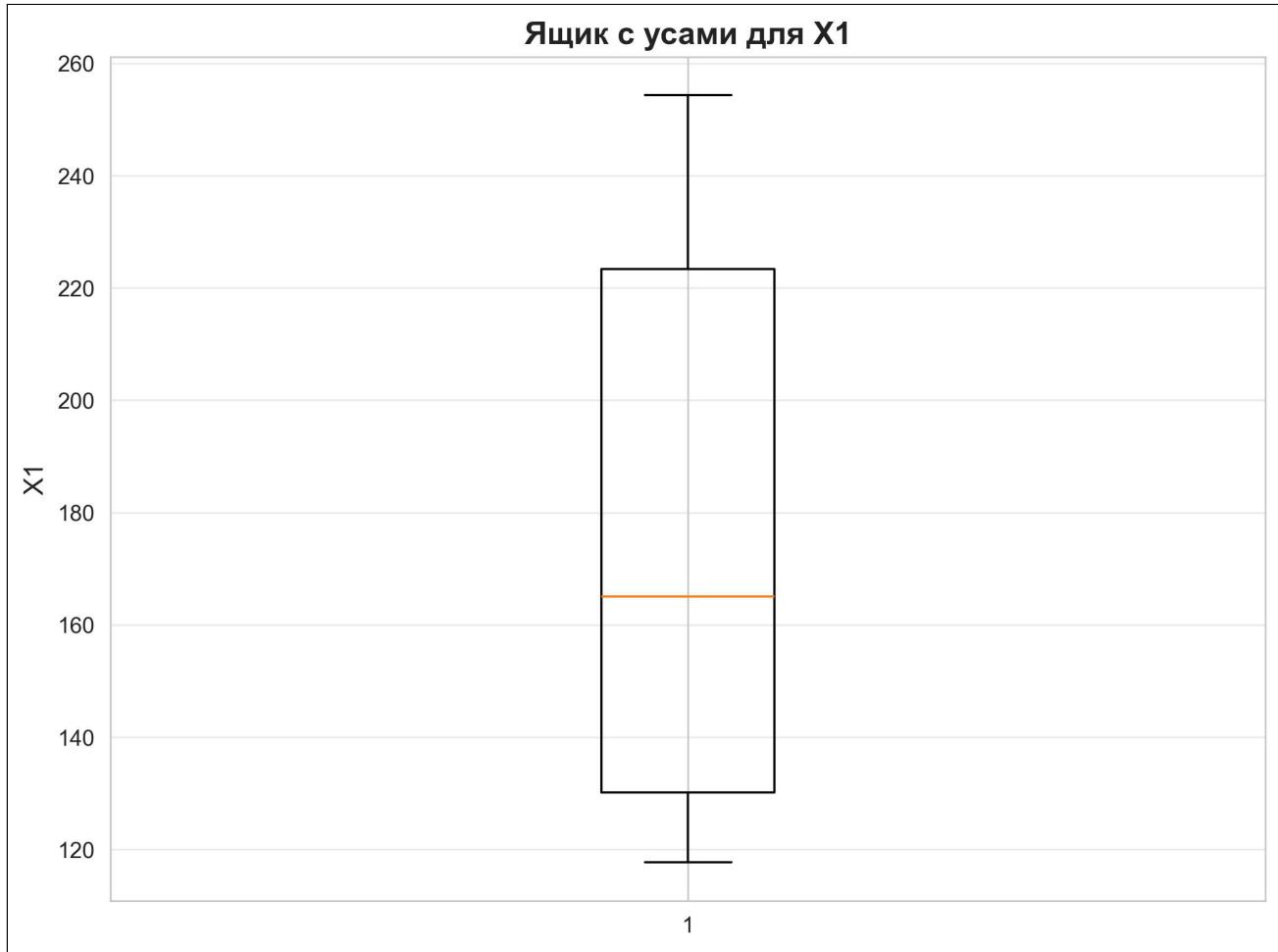


Рисунок 4. Ящик с усами для переменной X1

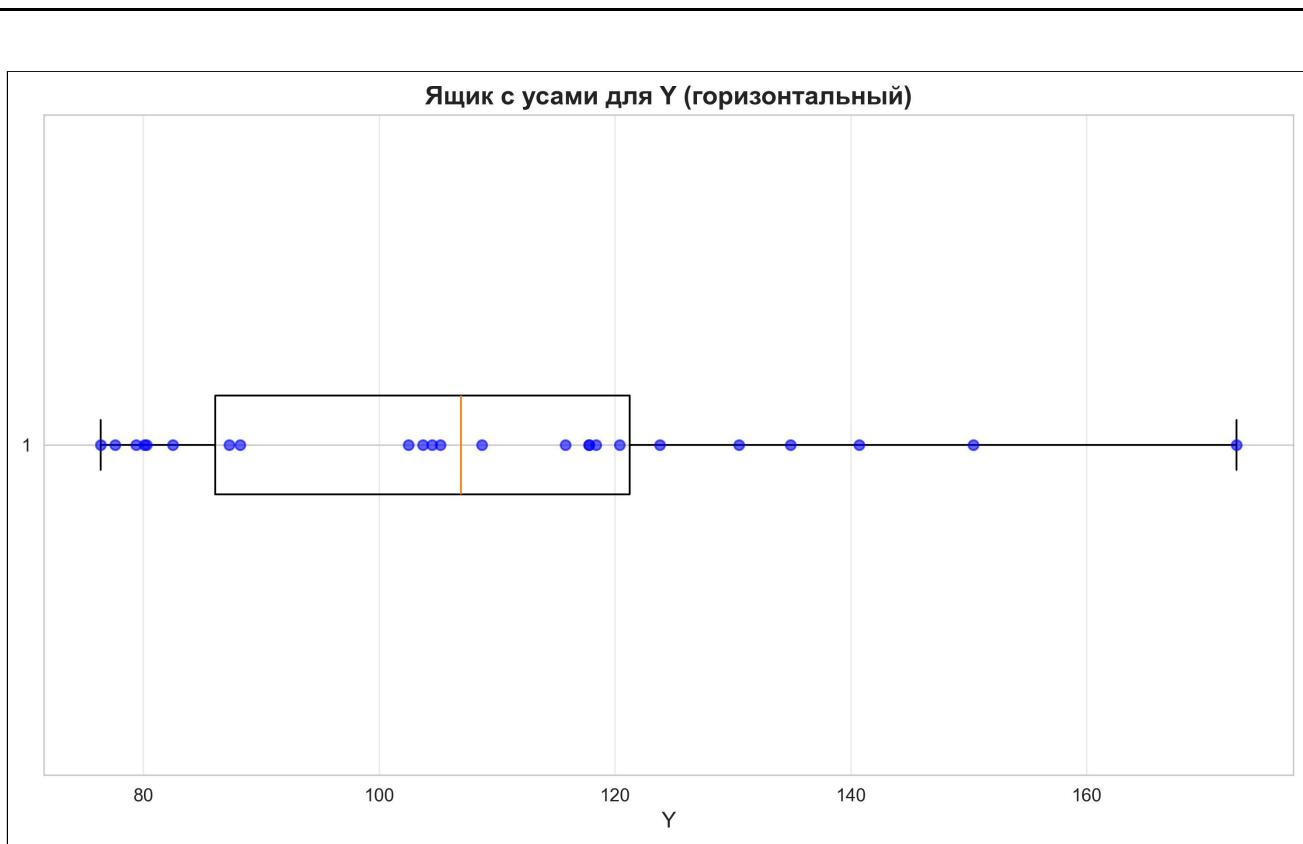


Рисунок 5. Горизонтальный ящик с усами для переменной Y с отображением точек

6. Диаграммы рассеяния

Диаграммы рассеяния позволяют визуализировать связь между переменными.

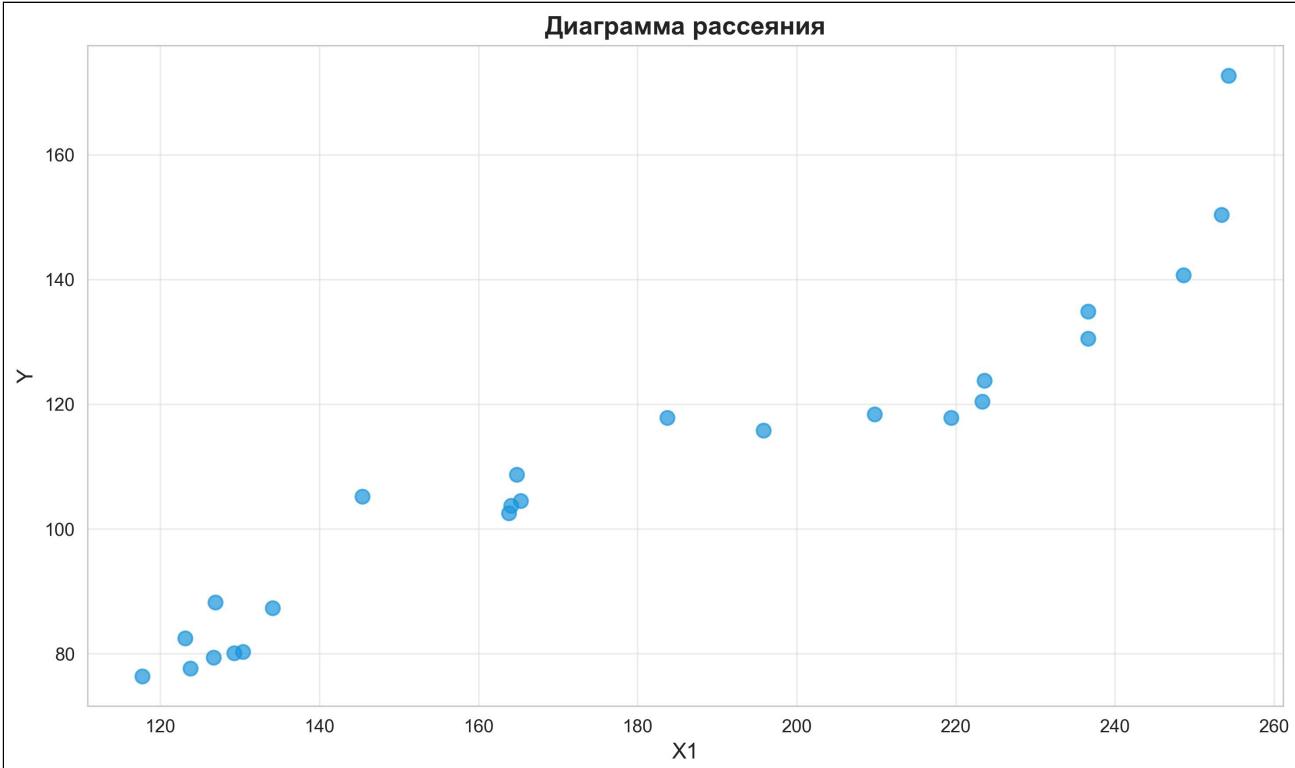


Рисунок 6. Диаграмма рассеяния для переменных X1 и Y

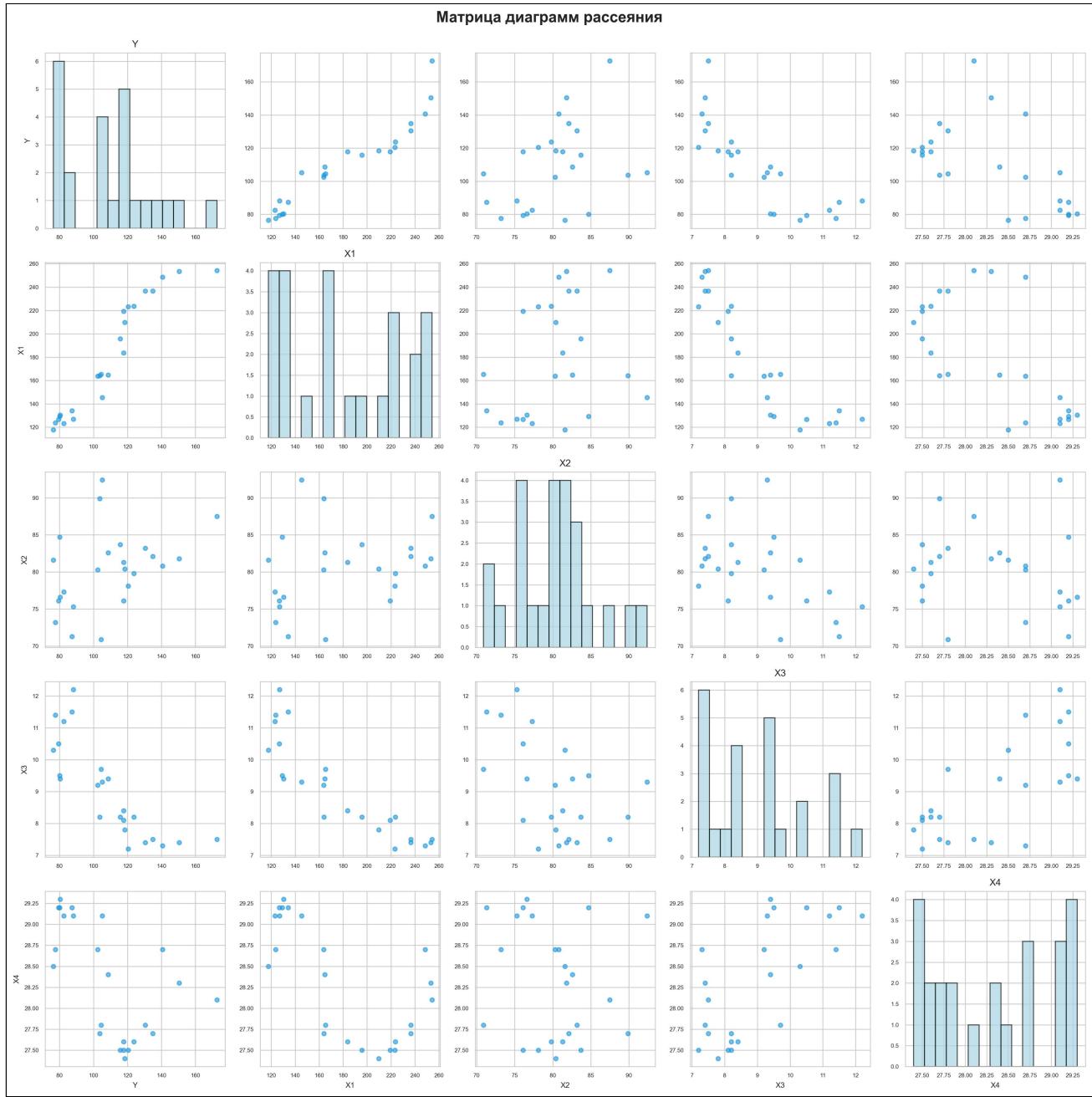


Рисунок 7. Матрица диаграмм рассеяния для всех переменных

7. Ковариация и корреляция

Ковариация между переменными Y и X_1 :

$$\text{cov}(Y, X_1) = 1165.2176$$

Коэффициент корреляции между переменными Y и X_1 :

$$r_{Y,X_1} = 0.9491$$

Корреляционная матрица для всех переменных:

	Y	X1	X2	X3	X4
--	---	----	----	----	----

Y	1.0000	0.9491	0.3913	-0.8158	-0.5760
X1	0.9491	1.0000	0.2716	-0.8859	-0.6739
X2	0.3913	0.2716	1.0000	-0.4999	-0.1613
X3	-0.8158	-0.8859	-0.4999	1.0000	0.6979
X4	-0.5760	-0.6739	-0.1613	0.6979	1.0000

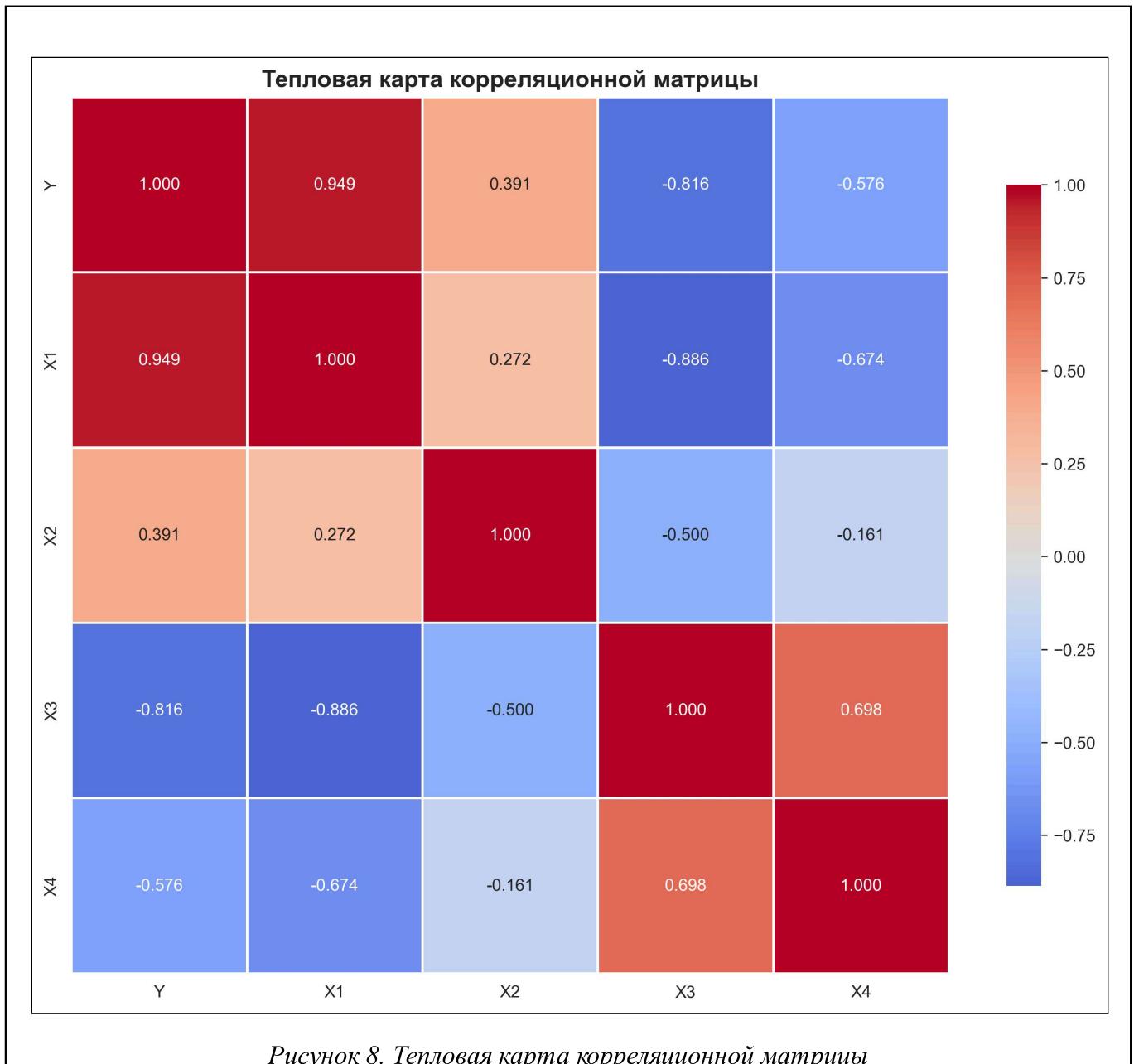


Рисунок 8. Термальная карта корреляционной матрицы

Выводы

В ходе выполнения лабораторной работы был проведен описательный анализ данных. Построены гистограммы, ящики с усами и диаграммы рассеяния для визуализации распределений и связей между переменными. Рассчитаны основные описательные статистики, включая меры центральной тенденции, вариации и формы распределения. Вычислена корреляционная матрица, которая

показывает степень линейной связи между переменными. Коэффициент корреляции между Y и X_1 составляет 0.9491, что указывает на сильную положительную связь между этими переменными.