Выводим гистограмму.

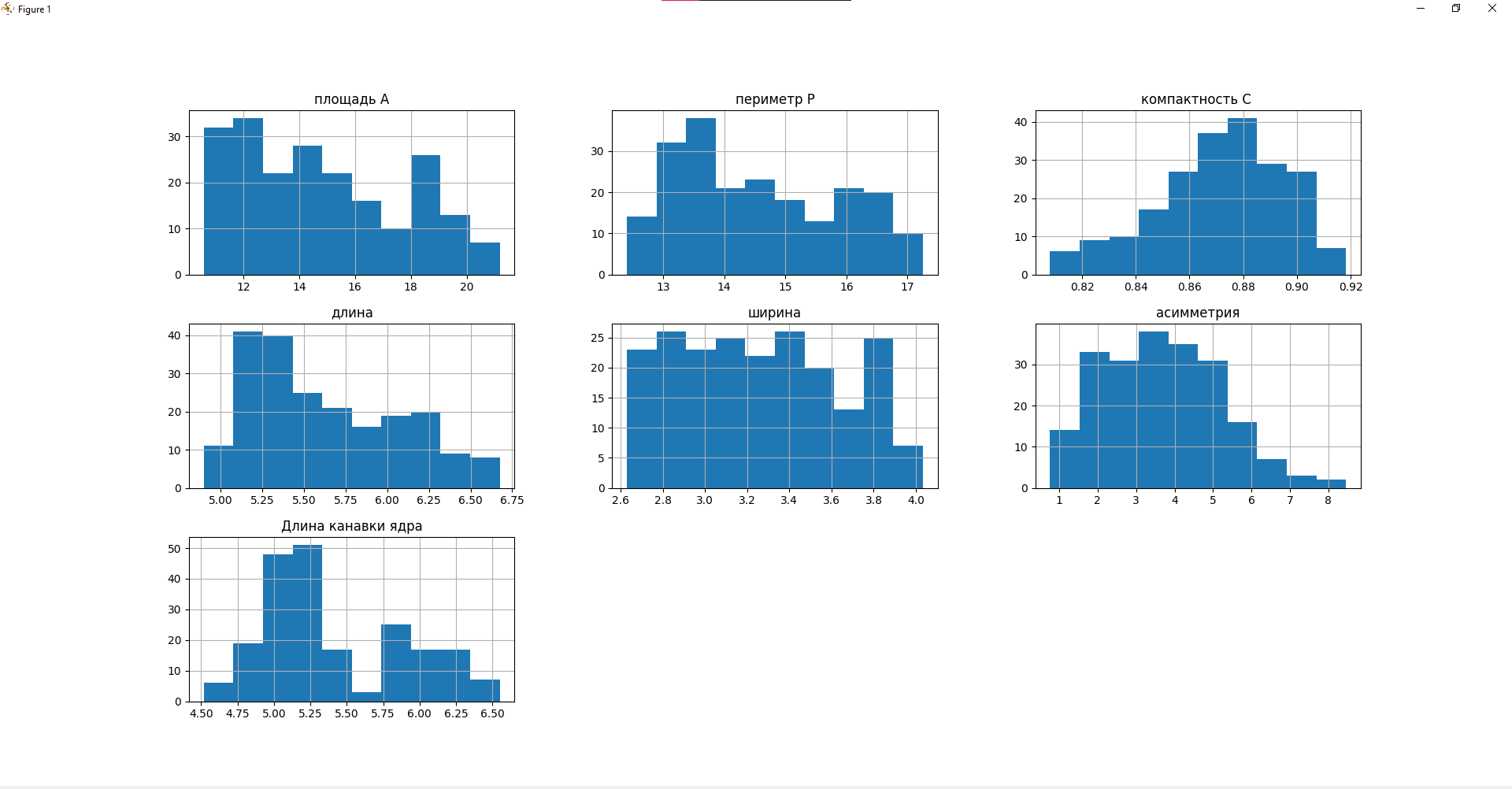


Рис. 1. Гистограмма по каждому признаку.

Посмотрев на такие гистограммы, можно сказать, что распределение не похоже на нормальное. Попробуем сделать более наглядно. Строим таблицу с зависимостями между признаками и раскрашиваем точки в зависимости от сортов зерна:

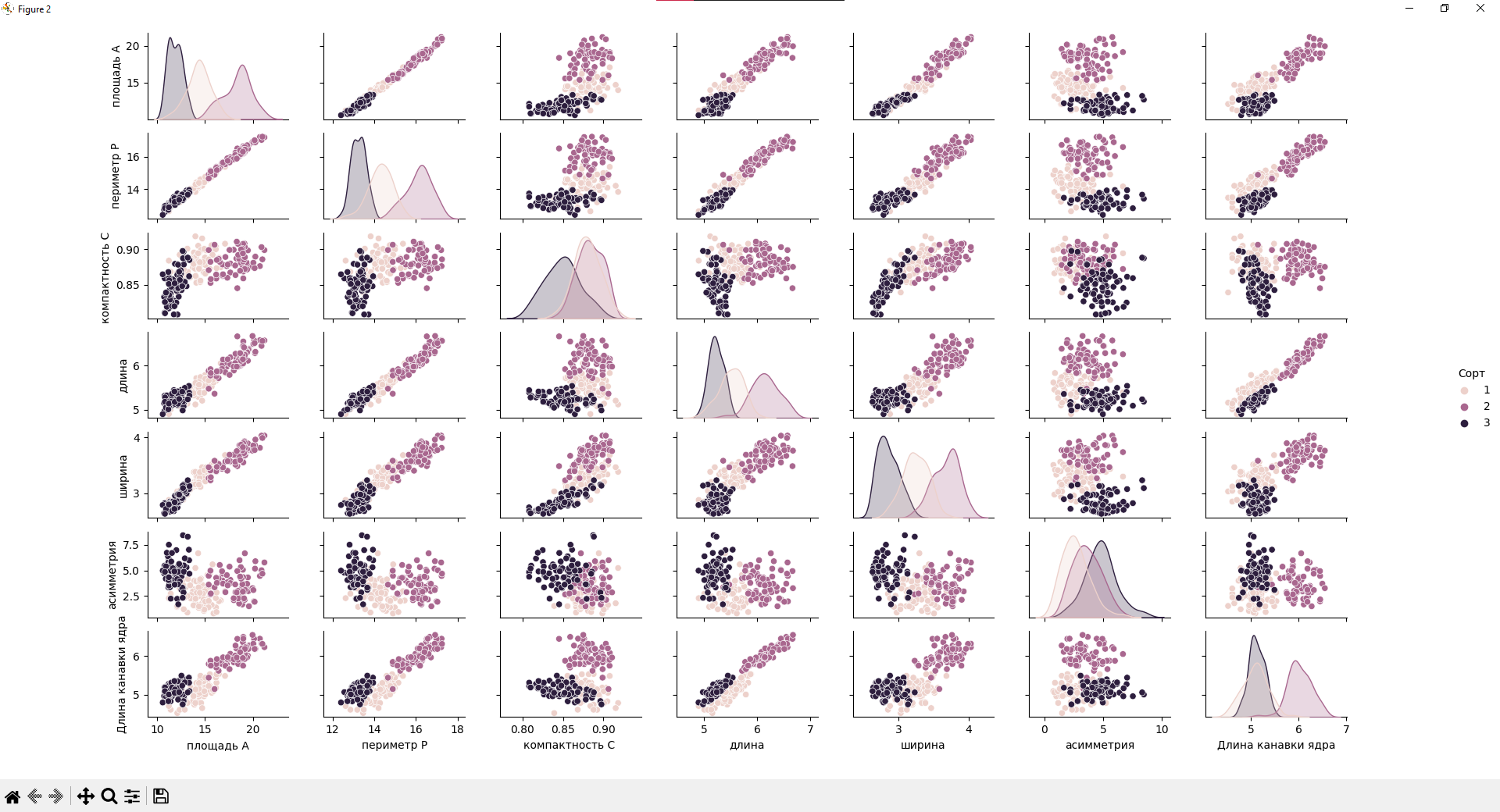


Рис. 2 Таблица зависимостей.

По таблице видно, что многие признаки имеют сильную зависимость. Их точки вытянуты вдоль одной линии. Для более наглядного вида необходимо построить тепловую карту признаков.

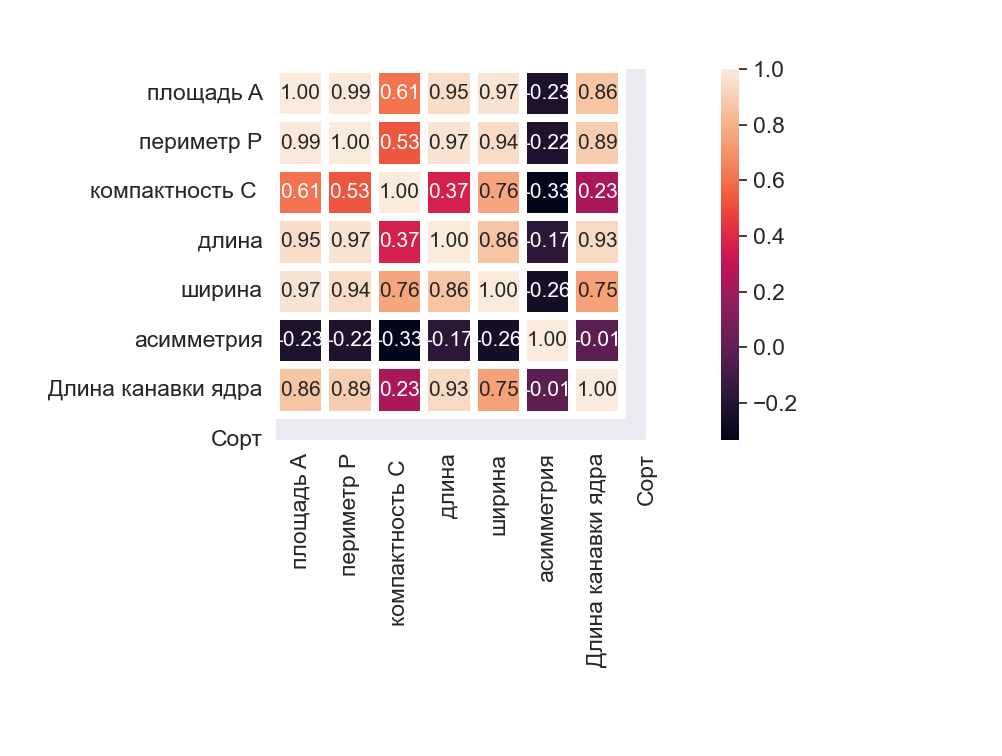


Рис. 3. Тепловая карта

Значения коэффициента корреляции интерпретируются следующим образом:

* До 0,2 — очень слабая корреляция
* До 0,5 — слабая
* До 0,7 — средняя
* До 0,9 — высокая
* Больше 0,9 — очень высокая

Самая слабая корреляция оказалась у «асимметрии», в основном у многих высокая корреляция.

**Код программы**

import numpy as np

import pandas as pd

import matplotlib

import matplotlib.pyplot as plt

import seaborn as sns

url = "C:/Users/fox/Desktop/ii/1lb/seeds\_dataset.data"

names = [

    "площадь A",

    "периметр P",

    "компактность C ",

    "длина",

    "ширина",

    "асимметрия",

    "Длина канавки ядра",

    "Сорт"

] #название атрибутов

dataset = pd.read\_csv(url, names=names) #чтение файла с данными

array = dataset.values

#( Кама , Роза , Канадка ) обозначены как числовые переменные 1, 2 и 3 соответственно. Семь исходных переменных

# ВЫВОД ГИСТОГРАММЫ

dataset[names[:7]].hist()

#ВЫВОД ЗАВИСИМОСТЕЙ МЕЖДУ ПРИЗНАКАМИ

sns.pairplot(dataset[names], height=1, hue = "Сорт")

plt.show()

#ВЫВОД ТЕПЛОВОЙ КАРТЫ ЗАВИСИМОСТИ ПРИЗНАКОВ.

corr = np.corrcoef(dataset[names[:7]].values.T)

sns.set(font\_scale=1.5)

ax = sns.heatmap(

    corr,

    cbar=True,

    annot=True,

    square=True,

    linewidths=5.55,

    fmt=".2f",

    annot\_kws={"size": 15},

    yticklabels=names,

    xticklabels=names,

) #3 диаграмма

plt.show()