Katedra Systemów Informacyjnych
Algorytmy analizy danych
Kierunek Informatyka, WE, sem. VII
Laboratorium nr 4

Grzegorz Cichosz

Celem zajęć jest zapoznanie studentów z problemem grupowania danych.
Aby zaliczyć laboratorium przygotuj sprawozdanie,
które będzie stanowić rozwiązania do poleceń zawartych w następujących plikach:
1. grupowanie_cz. 1_dane-test.xls
2. grupowanie_cz. 1_dane-iris.xls
~~3. grupowanie_cz. 1_dane-wine.xls~~
(mieliśmy zrobić pierwsze 2 w excelu a potem pythonie)

oraz
a. Zapoznaj się ze skryptem języka Phyton zawartym w p4.txt
Co oznaczają poszczególne fragmenty skryptu, za co są odpowiedzialne?
Sprawdź jak działają poszczególne funkcje i instrukcje.

#!/usr/bin/env python3
skrypt ma być uruchamiany za pomocą interpretera Python 3

-*- coding: utf-8 -*-
ustawia kodowanie UTF-8, co umożliwia użycie polskich znaków

import pandas as pd
importuje biblioteke pandas do wczytywania i przetwarzania danych tabelarycznych

import seaborn as sns
importuje bibliotekę seaborn, która służy do tworzenia wykresów statystycznych

import scipy.cluster.hierarchy as shc
importuje moduł do grupowania hierarchicznego i tworzenia dendrogramów

import matplotlib.pyplot as plt
importuje bibliotekę matplotlib do wyświetlania i opisywania wykresów

df = pd.read_csv('dane_test.csv')
wczytuje dane z pliku csv do obiektu DataFrame

print('liczba rekordów we wczytamym pliku wynosi:', len(df))
wyświetla liczbę wierszy w zbiorze danych

print(df.head())
wyświetla pierwsze kilka wierszy danych w celu ich podglądu

print(df.shape)
wyświetla liczbę wierszy i kolumn w zbiorze danych

sns.relplot(data=df, x="x", y="y")
tworzy wykres rozrzutu punktów dla zmiennych x oraz y

plt.show()
wyświetla wykres na ekranie

sns.pairplot(data=df[["x", "y"]])
tworzy macierz wykresów rozrzutu typu „każdy z każdym” dla wybranych zmiennych

plt.show()
wyświetla wykresy macierzowe

dendrogram = shc.dendrogram(shc.linkage(df, method='ward'))
wykonuje grupowanie hierarchiczne metodą Warda i tworzy dendrogram

plt.title('Dendrogram')
ustawia tytuł wykresu dendrogramu

plt.xlabel('nr instancji (wiersza w naszych danych)')
opisuje oś x jako numer instancji w zbiorze danych

plt.ylabel('Euclidean distances')
opsiuje oś y jako odległość euklidesową pomiędzy klastrami

plt.show()
wyświetla dendrogram

Odpowiedzi umieść w niezależnym pliku zwanym sprawozdaniem
b. Otwórz nowy projekt w środowisku PyCharm i utwórz nowy plik programu Python (nadaj mu nazwę np. p4 –czyli plik będzie się nazywał p4.py)
c. Skopiuj skrypt do p4.py zawartość z pliku p4.txt
d. Uruchom skrypt na danych (kolejno): test, iris i wine.
(nie korzystam z pycharma po prostu odpalę to w shellu. Nagłówek w sumie sugeruje że może to być tak uruchamiane)
Zestaw komend po kolei użytych do zrobienia środowiska I uruchomienia skryptu

#zmiana nazwy
mv p4.txt p4.py

#nadaj uprawnienia do uruchomienia
chmod +x p4.py

#robimy virtualne środowisko
python3 -m venv venv

#aktywujemy środowisko
source venv/bin/activate

#instalujemy zależności
pip install pandas seaborn matplotlib scipy

#odpalamy

```
(venv) talandar@gentoo ~/workspace/aad/lab4 $ ./p4.py
liczba rekordów we wczytamym pliku wynosi: 38
    x   y
0  19  65
1  22  74
2  27  72
3  28  76
4  24  58
(38, 2)
/home/talandar/workspace/aad/lab4/./p4.py:21: UserWarning: FigureCanvasAgg is non-interactive, and thus cannot be shown
  plt.show()
/home/talandar/workspace/aad/lab4/./p4.py:24: UserWarning: FigureCanvasAgg is non-interactive, and thus cannot be shown
  plt.show()
/home/talandar/workspace/aad/lab4/./p4.py:32: UserWarning: FigureCanvasAgg is non-interactive, and thus cannot be shown
  plt.show()  # show the dendrogram
(venv) talandar@gentoo ~/workspace/aad/lab4 $
```

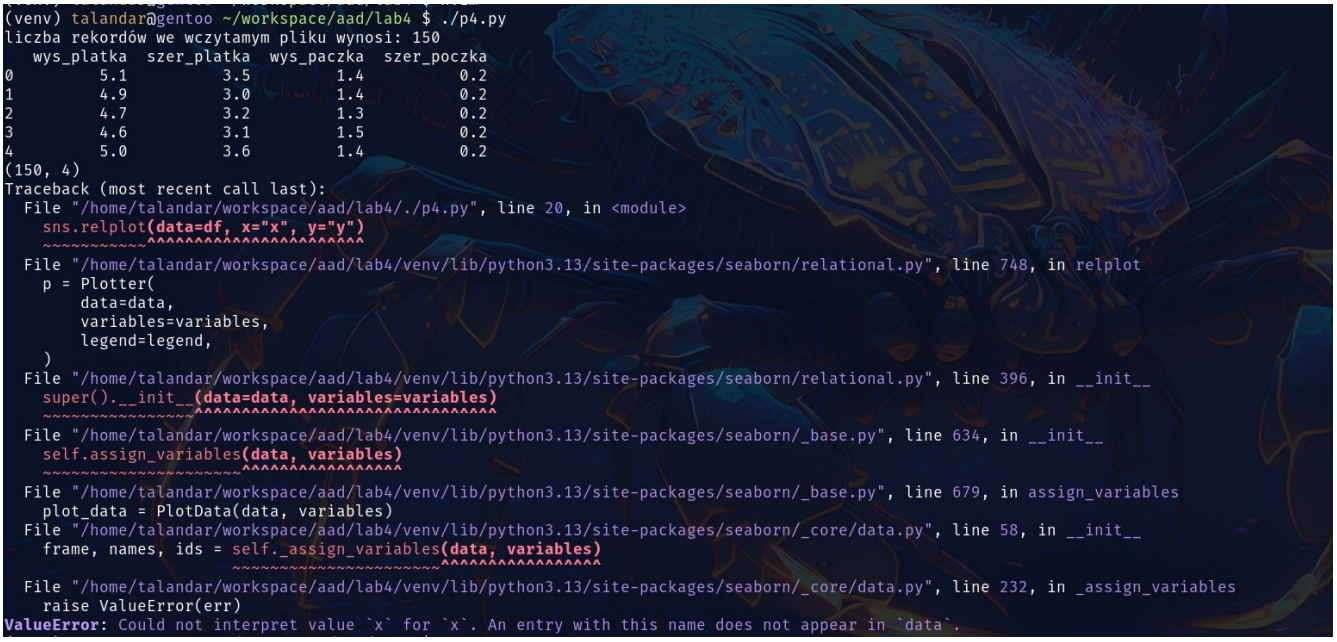
#ponieważ korzystamy z terminala musimy zapisać je jako obraz żeby jakkolwiek je zobaczyć.

Na ile musiałeś zmodyfikować skrypt, aby go uruchomić na tych danych (opisz to w sprawozdaniu)?

```
print(df.shape)

# wykres rozrzutu punktów
sns.relplot(data=df, x="x", y="y")
plt.savefig("scatter.png", dpi=150, bbox_inches="tight")
plt.close()

# wykres macierzowy - kazdy z kazdym
sns.pairplot(data=df[["x", "y"]])
plt.savefig("pairplot.png", dpi=150, bbox_inches="tight")
plt.close("all")

# dendogram
# finding the optimal number of clusters using dendrogram
dendrogram = shc.dendrogram(shc.linkage(df, method='ward'))
plt.title('Dendrogram')
plt.xlabel('nr instancji (wiersza w naszych danych)')
plt.ylabel('Euclidean distances')
plt.savefig("dendrogram.png", dpi=150, bbox_inches="tight")
plt.close()
```

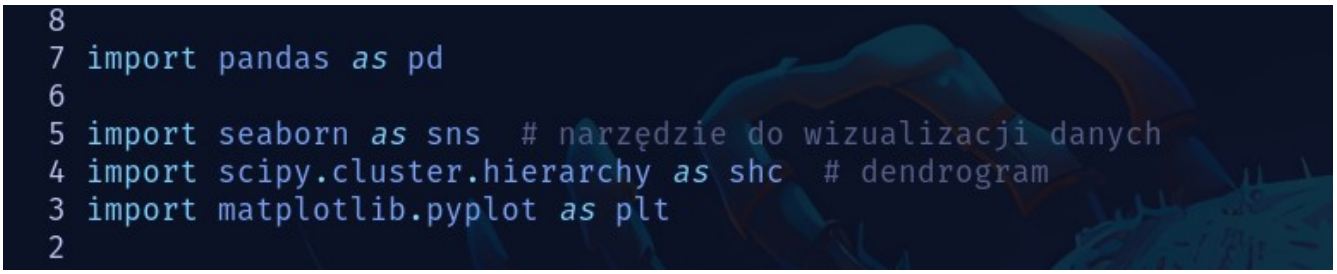
Na potrzeby czytania I wklejania wykresów do sprawka zapisałem je do plików

przy próbie uruchomienia iris trafiłem na błąd

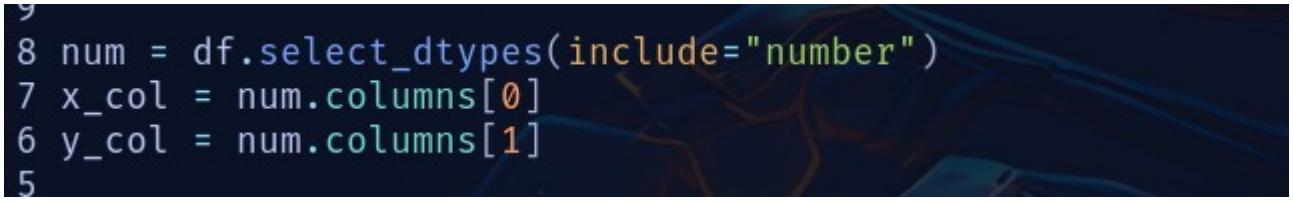


```
(venv) talandar@gentoo ~/workspace/aad/lab4 $ ./p4.py
liczba rekordów we wczytamym pliku wynosi: 150
   wys_platka  szer_platka  wys_paczka  szer_poczka
0         5.1          3.5         1.4          0.2
1         4.9          3.0         1.4          0.2
2         4.7          3.2         1.3          0.2
3         4.6          3.1         1.5          0.2
4         5.0          3.6         1.4          0.2
(150, 4)
Traceback (most recent call last):
  File "/home/talandar/workspace/aad/lab4/./p4.py", line 20, in <module>
    sns.relplot(data=df, x="x", y="y")
    ~~~~~~~~~~~^^^^^^^^^^^^^^^^^^^^^^^
  File "/home/talandar/workspace/aad/lab4/venv/lib/python3.13/site-packages/seaborn/relational.py", line 748, in relplot
    p = Plotter(
        data=data,
        variables=variables,
        legend=legend,
    )
  File "/home/talandar/workspace/aad/lab4/venv/lib/python3.13/site-packages/seaborn/relational.py", line 396, in __init__
    super().__init__(data=data, variables=variables)
    ~~~~~~~~~~~~~~~~^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^^
  File "/home/talandar/workspace/aad/lab4/venv/lib/python3.13/site-packages/seaborn/_base.py", line 634, in __init__
    self.assign_variables(data, variables)
    ~~~~~~~~~~~~~~~~~~~~~^^^^^^^^^^^^^^^^^
  File "/home/talandar/workspace/aad/lab4/venv/lib/python3.13/site-packages/seaborn/_base.py", line 679, in assign_variables
    plot_data = PlotData(data, variables)
  File "/home/talandar/workspace/aad/lab4/venv/lib/python3.13/site-packages/seaborn/_core/data.py", line 58, in __init__
    frame, names, ids = self._assign_variables(data, variables)
                        ~~~~~~~~~~~~~~~~~~~~~~^^^^^^^^^^^^^^^^^
  File "/home/talandar/workspace/aad/lab4/venv/lib/python3.13/site-packages/seaborn/_core/data.py", line 232, in _assign_variables
    raise ValueError(err)
ValueError: Could not interpret value `x` for `x`. An entry with this name does not appear in `data`.
```

Zmienione zostało: wybór danych do wizualizacji i dendrogramu na kolumny liczbowe (bez kolumn tekstowych), a do wykresu rozrzutu użyto pierwszych dwóch cech liczbowych



```
import pandas as pd

import seaborn as sns  # narzędzie do wizualizacji danych
import scipy.cluster.hierarchy as shc  # dendrogram
import matplotlib.pyplot as plt
```



```
num = df.select_dtypes(include="number")
x_col = num.columns[0]
y_col = num.columns[1]
```

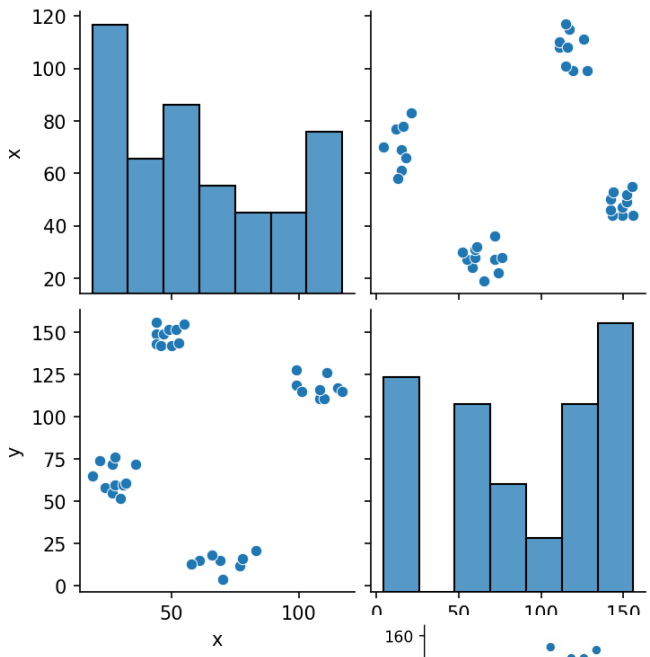
W pliku sprawozdania umieść wykresy, jakie wygenerowałeś przy użyciu skryptu wraz z komentarzem

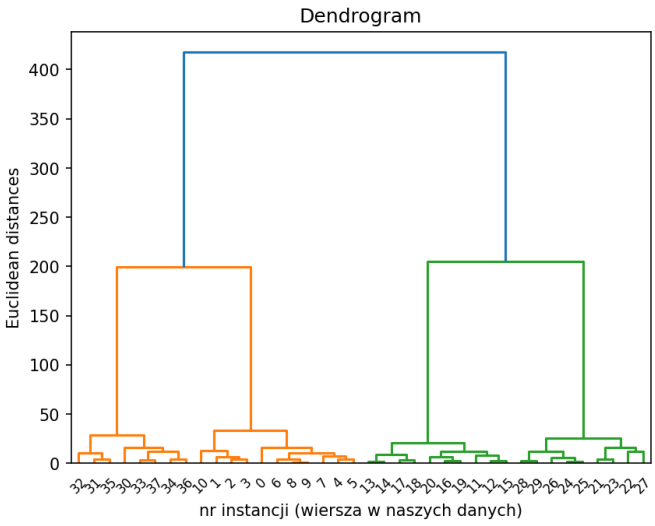
dane_test.csv

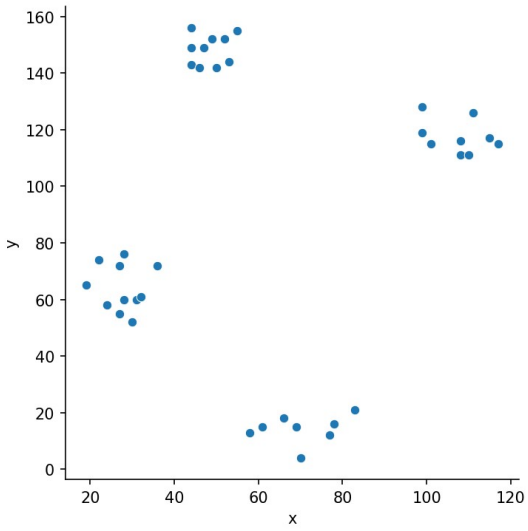
liczba rekordów we wczytamym pliku wynosi: 38

	x	y
0	19	65
1	22	74
2	27	72
3	28	76
4	24	58

(38, 2)





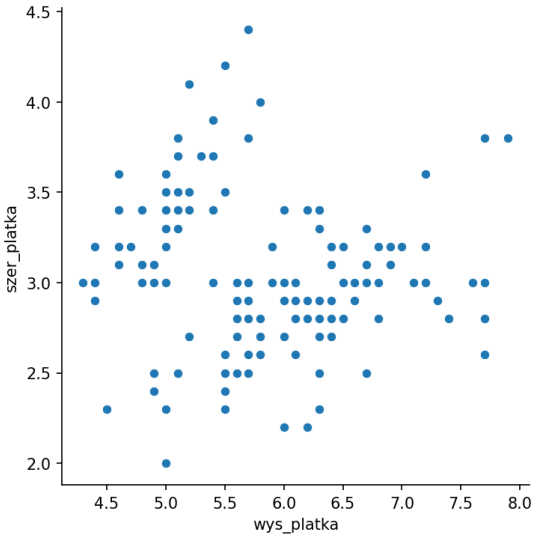


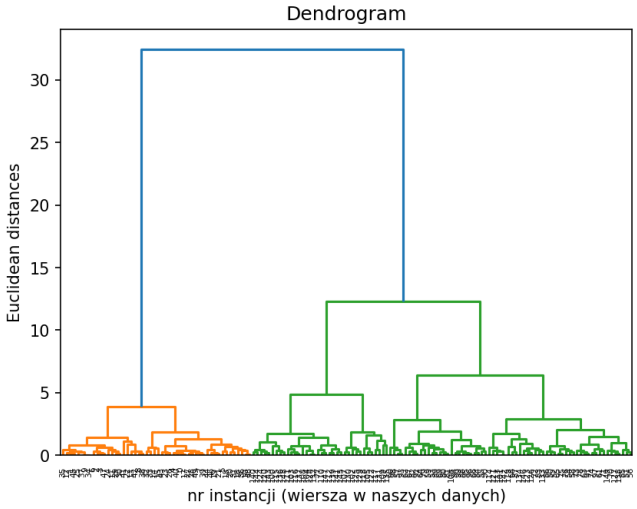
dane_iris.csv

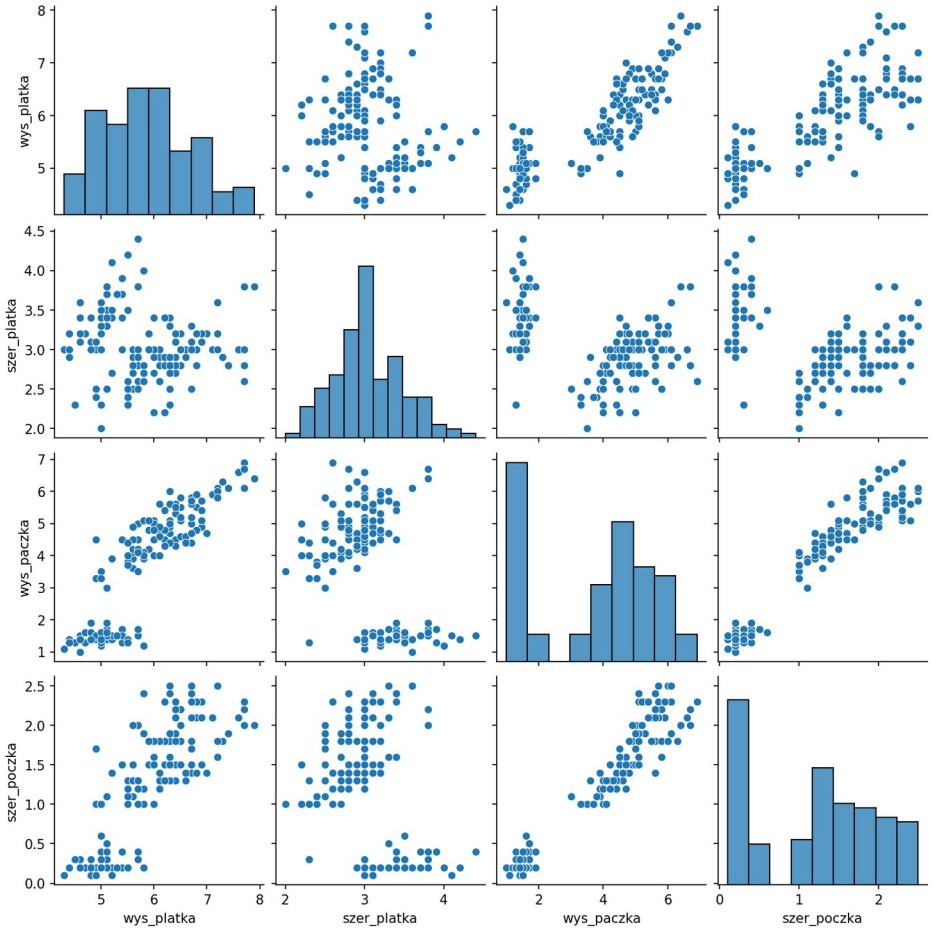
liczba rekordów we wczytamym pliku wynosi: 150

	wys_platka	szer_platka	wys_paczka	szer_poczka
0	**5.1**	**3.5**	**1.4**	**0.2**
1	**4.9**	**3.0**	**1.4**	**0.2**
2	**4.7**	**3.2**	**1.3**	**0.2**
3	**4.6**	**3.1**	**1.5**	**0.2**
4	**5.0**	**3.6**	**1.4**	**0.2**

(150, 4)







Jako sprawozdanie (w iliasie) prześlij pliki xls z rozwiązaniami oraz plik z odpowiedziami do pkt. a-d