

# Sprawozdanie 2

## Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner

2025-04-23

## Spis treści

<b>1</b>	<b>ZADANIE 1 (Dyskretyzacja(przedziałowanie) cech ciągłych)</b>	<b>2</b>
1.1	a) Dane: iris (R-pakiet datasets). . . . .	2
1.2	b) Wybór cech . . . . .	3
1.3	c) Porównanie nienadzorowanych metod dyskretyzacji . . . . .	5
<b>2</b>	<b>ZADANIE 2 (Analizaskładowych głównych (Principal Component Analysis (PCA)))</b>	<b>5</b>
2.1	a) Dane: City Quality of Life Dataset (plik uaScoresDataFrame.csv, źródło: Kaggle/Teleport.org) . . . . .	5
2.2	b) Przygotowanie danych . . . . .	5
2.3	c) Wyznaczenie składowych głównych . . . . .	5
2.4	d) Zmienność odpowiadająca poszczególnym składowym . . . . .	5
2.5	e) Wizualizacja danych wielowymiarowych . . . . .	5
2.6	f) Korelacja zmiennych . . . . .	5
2.7	g) Końcowe wnioski . . . . .	5
<b>3</b>	<b>ZADANIE 3 (Skalowaniewielowymiarowe (Multidimensional Scaling (MDS)))</b>	<b>5</b>
3.1	a) Dane: titanic_train (R-pakiet titanic) . . . . .	5
3.2	b) Przygotowanie danych . . . . .	5
3.3	c) Redukcja wymiaru na bazie MDS . . . . .	5
3.4	d) Wizualizacja danych . . . . .	5
##	X UA_Name UA_Country UA_Continent Housing Cost.of.Living Startups	
## 1 0	Aarhus Denmark Europe 6.1315 4.015 2.8270	
## 2 1	Adelaide Australia Oceania 6.3095 4.692 3.1365	
## 3 2	Albuquerque New Mexico North America 7.2620 6.059 3.7720	
## 4 3	Almaty Kazakhstan Asia 9.2820 9.333 2.4585	
## 5 4	Amsterdam Netherlands Europe 3.0530 3.824 7.9715	
## 6 5	Anchorage Alaska North America 5.4335 3.141 2.7945	

```
## Venture.Capital Travel.Connectivity Commute Business.Freedom Safety
## 1 2.512 3.5360 6.31175 9.940000 9.6165
## 2 2.640 1.7765 5.33625 9.399667 7.9260
## 3 1.493 1.4555 5.05575 8.671000 1.3435
## 4 0.000 4.5920 5.87125 5.568000 7.3090
## 5 6.107 8.3245 6.11850 8.836667 8.5035
## 6 0.000 1.7380 4.71525 8.671000 3.4705
## Healthcare Education Environmental.Quality Economy Taxation Internet.Access
## 1 8.704333 5.3665 7.63300 4.8865 5.0680 8.3730
## 2 7.936667 5.1420 8.33075 6.0695 4.5885 4.3410
## 3 6.430000 4.1520 7.31950 6.5145 4.3460 5.3960
## 4 4.545667 2.2830 3.85675 5.2690 8.5220 2.8860
## 5 7.907333 6.1800 7.59725 5.0530 4.9550 4.5230
## 6 6.060333 3.6245 9.27200 6.5145 4.7720 4.9645
## Leisure...Culture Tolerance Outdoors
## 1 3.1870 9.7385 4.1300
## 2 4.3285 7.8220 5.5310
## 3 4.8900 7.0285 3.5155
## 4 2.9370 6.5395 5.5000
## 5 8.8740 8.3680 5.3070
## 6 3.2660 7.0930 5.3580

## Warning: pakiet 'dplyr' został zbudowany w wersji R 4.4.2
## Warning: pakiet 'kableExtra' został zbudowany w wersji R 4.4.3
## Warning: pakiet 'patchwork' został zbudowany w wersji R 4.4.2
## Warning: pakiet 'ggplot2' został zbudowany w wersji R 4.4.2
```

# 1 ZADANIE 1 (Dyskretyzacja(przedziałowanie) cech ciągłych)

## 1.1 a) Dane: iris (R-pakiet datasets).

3 Pierwsze wiersze z pakietu iris

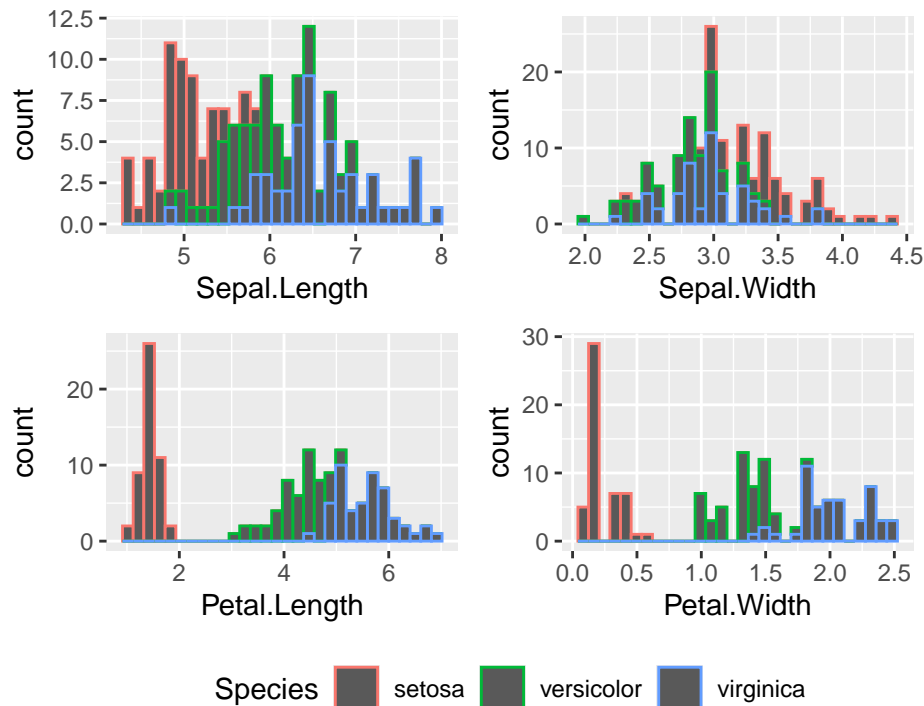
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa

Zbiór danych zawiera wyniki pomiarów uzyskanych dla **trzech gatunków irysów** (tj. setosa, versicolor i virginica) i został **udostępniony przez Ronalda Fishera w roku 1936**.

– **Pomiary** dotyczą **długości oraz szerokości** dwóch różnych części kwiatu– działki **kielicha** (ang. sepal) oraz **płatka** (ang. petal).

## 1.2 b) Wybór cech

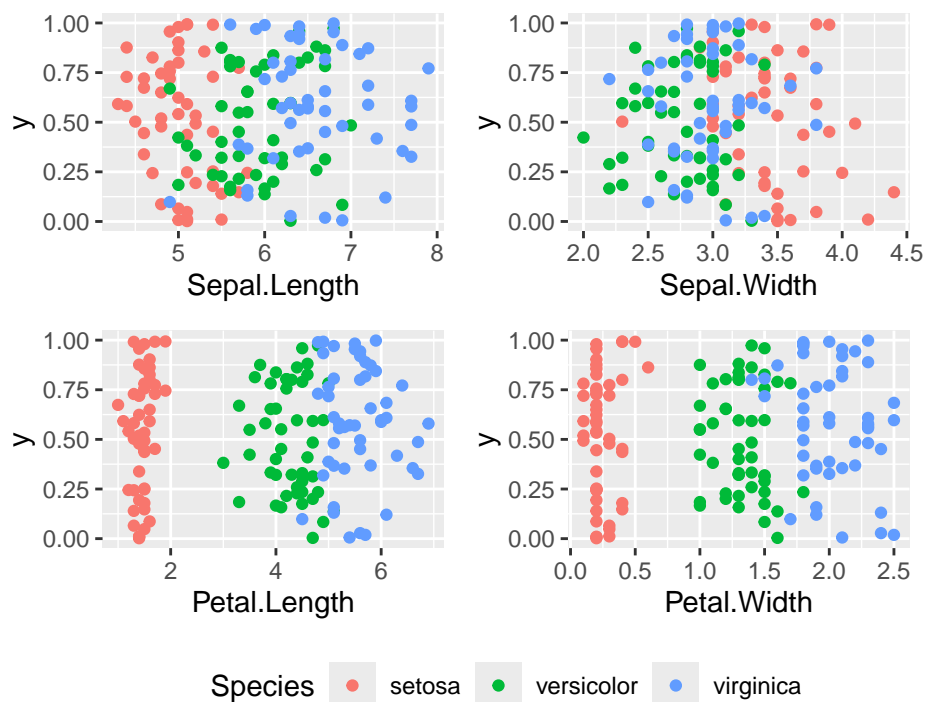
Cechy, inaczej właściwie możemy to rozstrzygać jako kolumny, które charakteryzują się **największym zróżnicowaniem** w stosunku do rodzaju gatunku



Po przeanalizowaniu histogramów, widać ,że warto zwrócić uwagę na takie cechy jak **Petal.Length** i **Petal.Width**, ponieważ

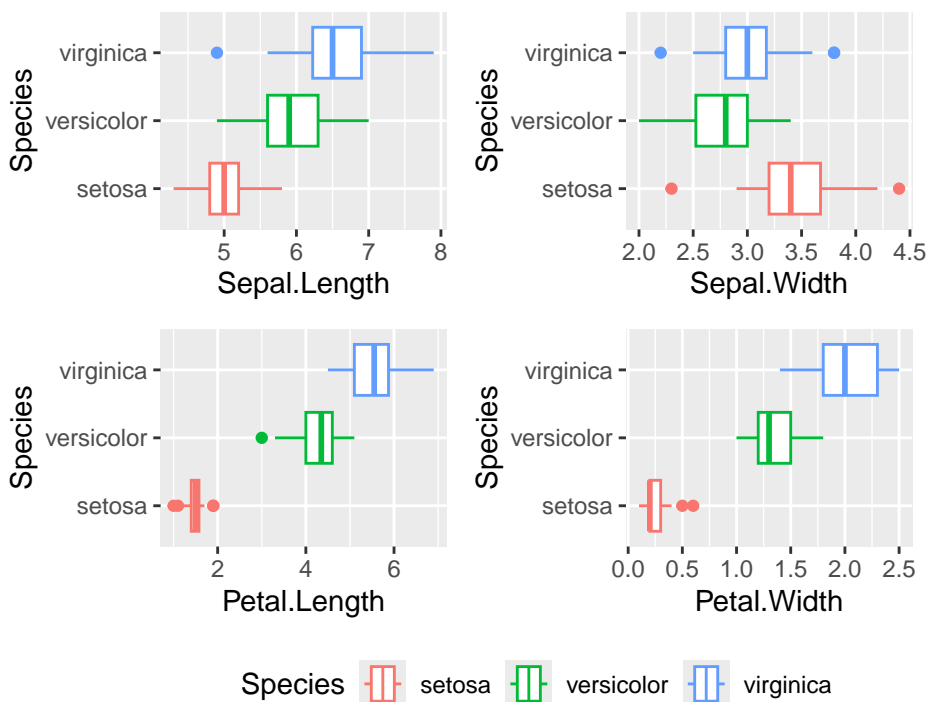
widać dobrze zaznaczone przedziały w których występuje większość kwiatków danego gatunku.

Dalej warto jest też spojrzeć na to jak nasze *obserwacje* teoretycznie rozkładają się w przestrzeni 2D, aby to zrobić dodajemy jedną dodatkową kolumnę y, wypełnioną losowymi liczbami od 0 do 1 (rozkład jednostajny)



Wykresy typu scatter-plot potwierdzają, że **Petal.Length** i **Petal.Width** są bardzo dobrym wyborem cech, które mogłyby być wyznacznikami gatunków roślin.

Musimy jednak wybrać wartości najlepsze i najgorsze, aby to zrobić przeanalizujemy jeszcze boxploty.



Na ich podstawie możemy uznać, że **Petal.Width** może stanowić najlepszy wyznacznik gatunku roślin. Najgorszym natomiast jest **Sepal.Width**, tutaj duża część gatunków dzieli te

same wartości tej cechy.

### **1.3 c) Porównanie nienadzorowanych metod dyskretyzacji**

## **2 ZADANIE 2 (Analizaskładowych głównych (Principal Component Analysis (PCA)))**

**2.1 a) Dane: City Quality of Life Dataset (plik uaScoresDataFrame.csv, źródło: Kaggle/Teleport.org)**

**2.2 b) Przygotowanie danych**

**2.3 c) Wyznaczenie składowych głównych**

**2.4 d) Zmienność odpowiadająca poszczególnym składowym**

**2.5 e) Wizualizacja danych wielowymiarowych**

**2.6 f) Korelacja zmiennych**

**2.7 g) Końcowe wnioski**

## **3 ZADANIE 3 (Skalowaniewielowymiarowe (Multidimensional Scaling (MDS)))**

**3.1 a) Dane: titanic\_train (R-pakiet titanic)**

**3.2 b) Przygotowanie danych**

**3.3 c) Redukcja wymiaru na bazie MDS**

**3.4 d) Wizualizacja danych**