Sprawozdanie 2

Eksploracja danych

Kacper Szmigielski, 282255 i Mateusz Wizner

2025-04-22

Spis treści


```
##   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
```

1 ZADANIE 1 (Dyskretyzacja(przedziałowanie) cech ciągłych)

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

3 Pierwsze wiersze z pakietu iris

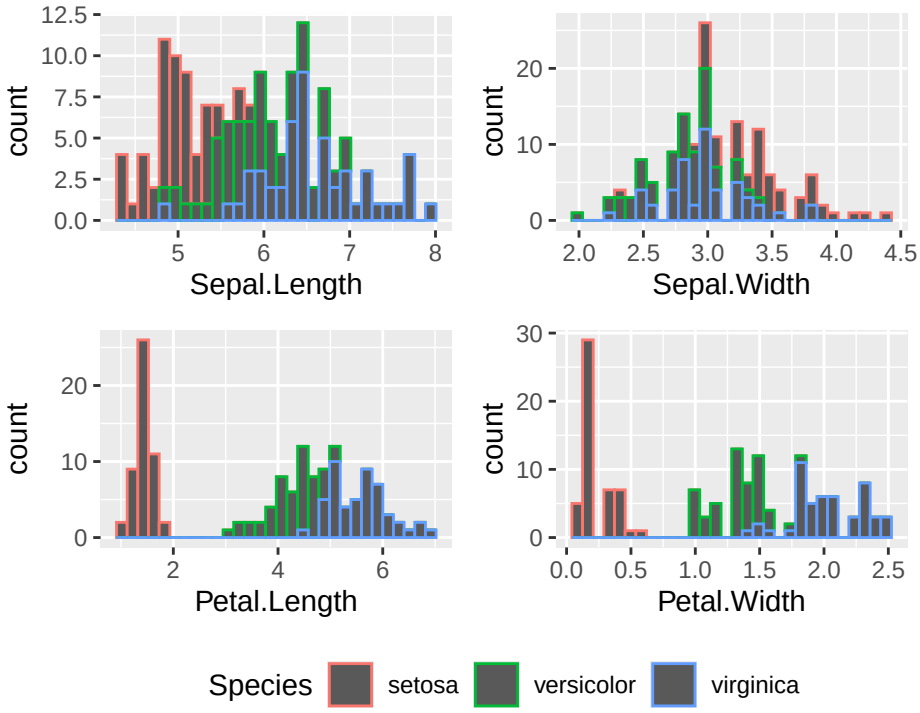
Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa

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 utworzeniu wykresów zależności długości i szerokości taki zmiennych jak od gatunku, jasno widać ,że

warto jest zwrócić uwagę na takie cechy jak: Petal.Length i Petal.Width

Z wykresów widać,że Petal.Legnth jak i Petal.Width charakteryzuje się 3 zagęszczeniami wyników z czego każde należy do innej grupy.

Widać również,że Sepal nie jest tak zróżnicowany na tle gatunkowym kwiatów jak Petal.

Wariancje:

Dla Length: 3.1162779

Dla Width: 0.5810063

Pomimo znacznie większej wariancji dla Petal.Length , widać że o wiele lepiej można wnioskować gatunek za pomocą Petal.Width, gdzie

dla

setosy jest on w przedziale (0,2)

versicolor jest on w przedziale (1,1.7)

a dla virginica znaczna większość znajduje się w przedziale (1.5,2.5)

Natomiast gdyby rozróżniać wedłóg Petal.Length mogły by wystąpić sprzeczność idla versicolor i setosy, które

w znaczniej więszkośći wykazują skołonność do posiadania od 4 do 6 Petal.Length