

# Dyskretyzacja cech ciągłych, analiza składowych głównych (PCA) i skalowanie wielowymiarowe (MDS)

Stanisław Olek

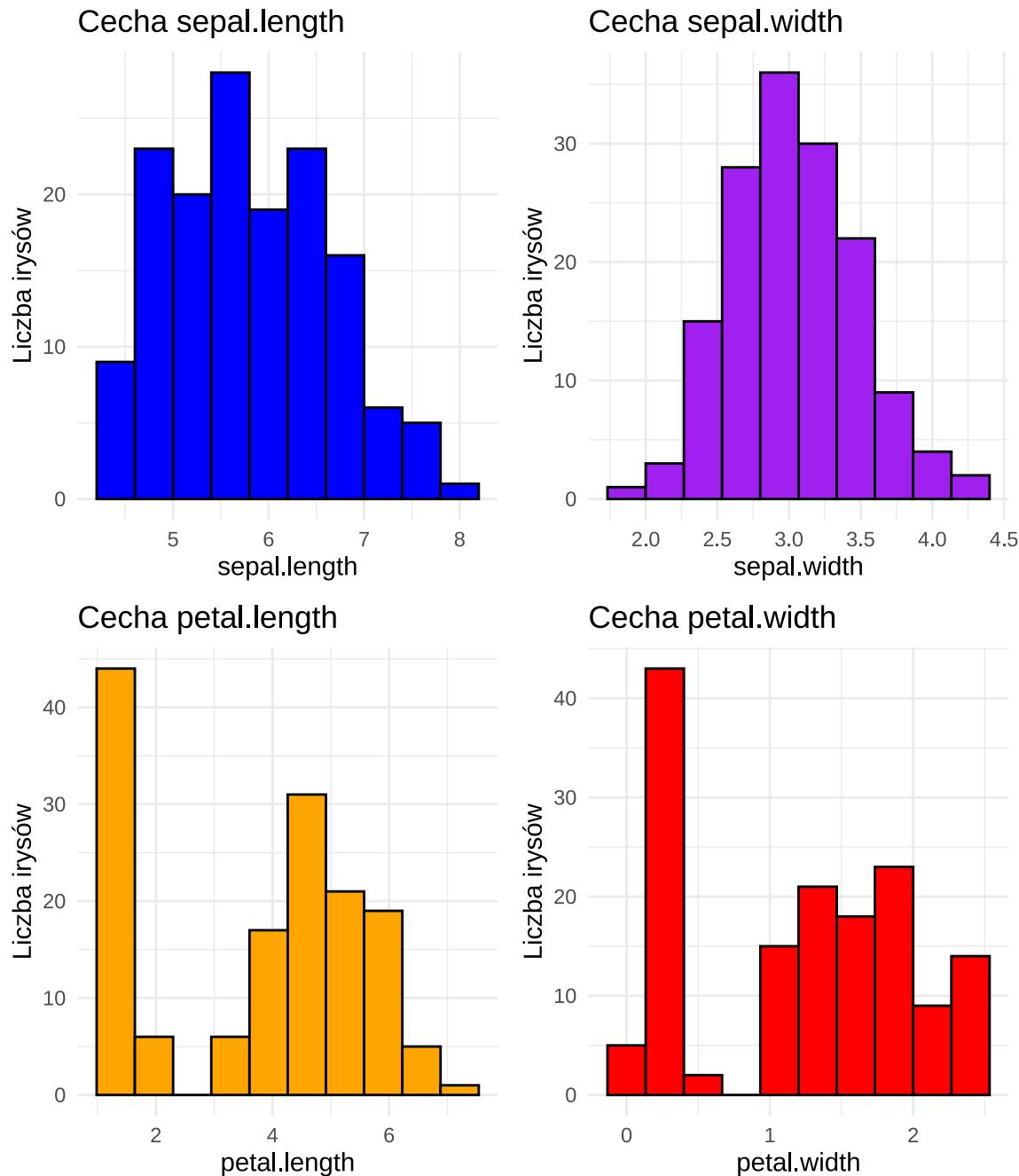
## Spis treści

<b>1 Dyskretyzacja (przedziałowanie) cech ciągłych</b>	<b>2</b>
1.1 Metoda oparta na równych częstościach . . . . .	5
1.2 Metoda oparta na przedziałach o jednakowej szerokości . . . . .	9
1.3 Metoda oparta na algorytmie grupowania (algorytm k-srednich) . . . . .	13
1.4 Dyskretyzacja z przedziałami zadanymi przez użytkownika . . . . .	17
1.5 Wnioski . . . . .	20
<b>2 Analiza składowych głównych (Principal Component Analysis (PCA))</b>	<b>21</b>
2.1 Przygotowanie danych . . . . .	21
2.2 Analiza składowych głównych . . . . .	24
2.3 Wizualizacja danych wielowymiarowych . . . . .	28
2.4 Korelacja zmiennych . . . . .	31
2.5 Wnioski . . . . .	34
<b>3 Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))</b>	<b>34</b>
3.1 Przygotowanie danych . . . . .	34
3.2 Redukcja wymiaru na bazie MDS . . . . .	35
3.3 Wizualizacja wyników MDS . . . . .	36
3.4 Wnioski: . . . . .	38

# 1 Dyskretyzacja (przedziałowanie) cech ciągłych

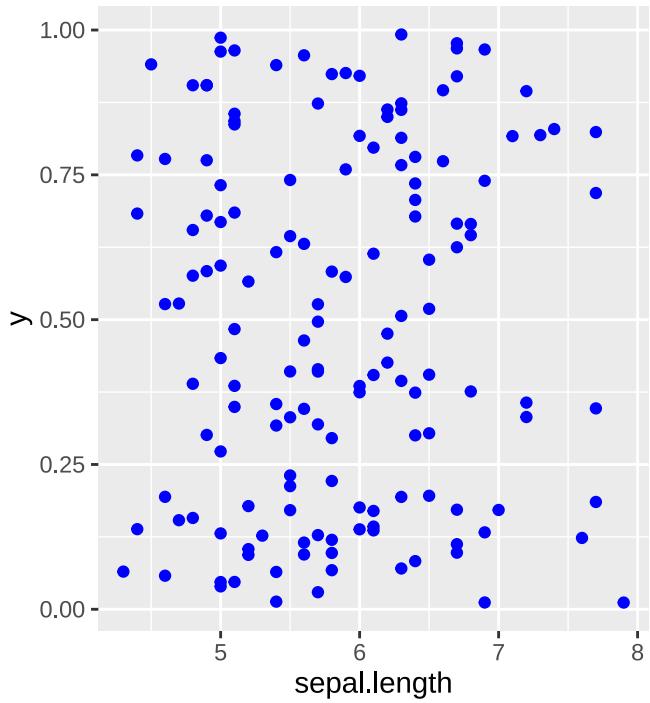
W tej sekcji zajmiemy się dyskretyzacją cech ciągłych danych ze zbioru `iris`. Zawiera on 3 gatunki irysów: **setosa**, **versicolor**, **virginica**.

W tym celu wybierzymy jedną cechę o najlepszych zdolnościach dyskryminacyjnych, to znaczy taką, która zapewni nam najlepszą separację gatunków oraz jedną cechę o najgorszych takich zdolnościach.

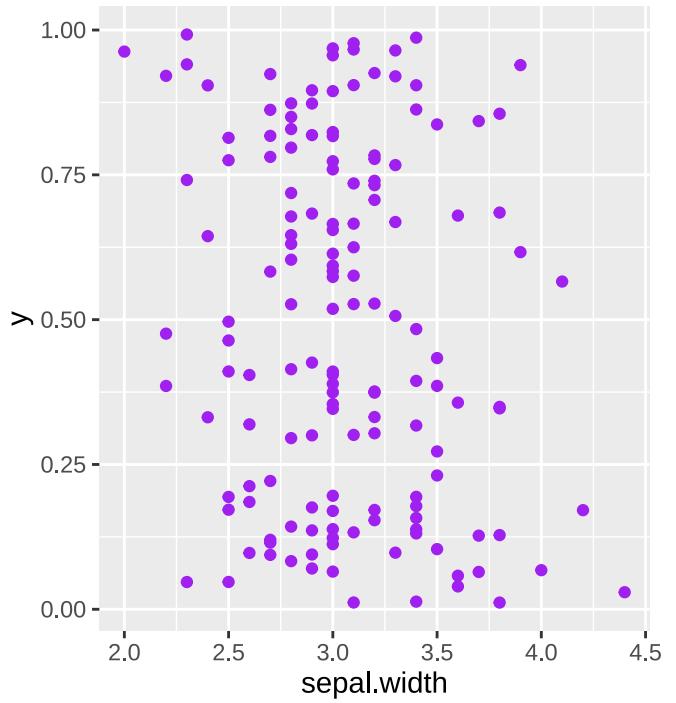


Rysunek 1: Histogramy poszczególnych cech

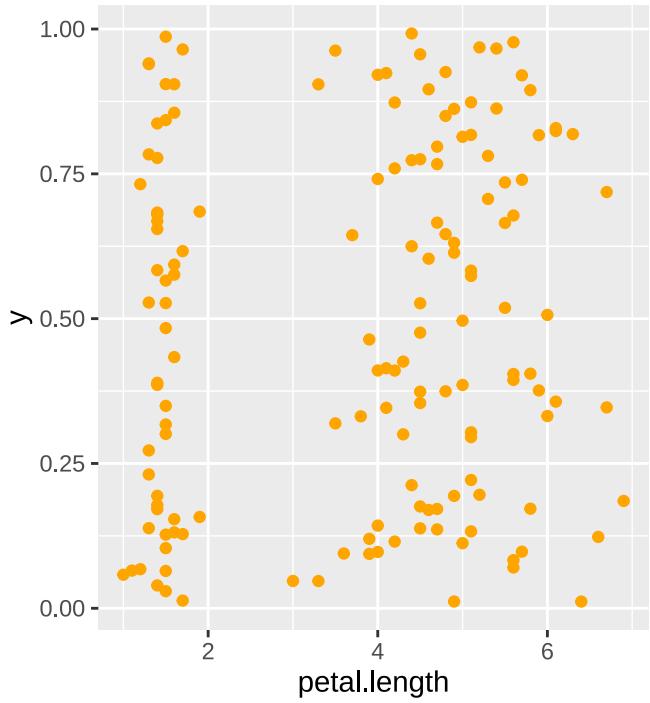
Cecha sepal.length



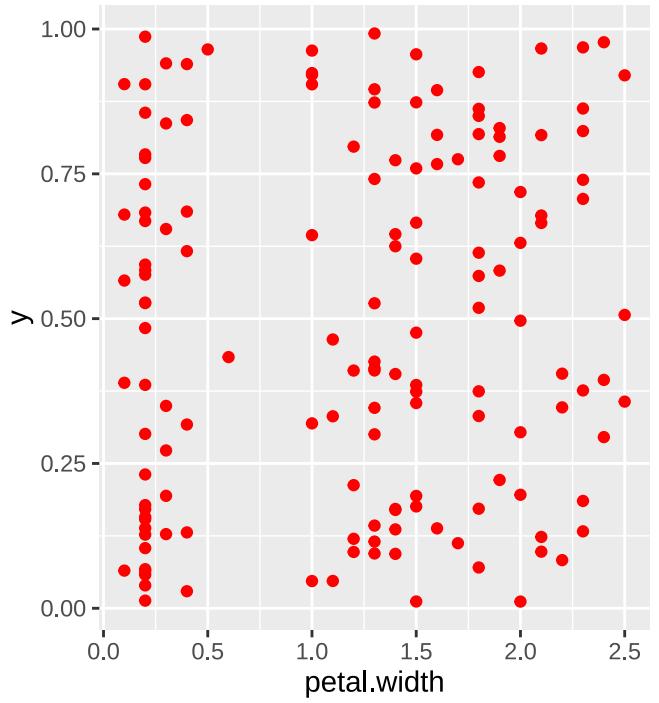
Cecha sepal.width



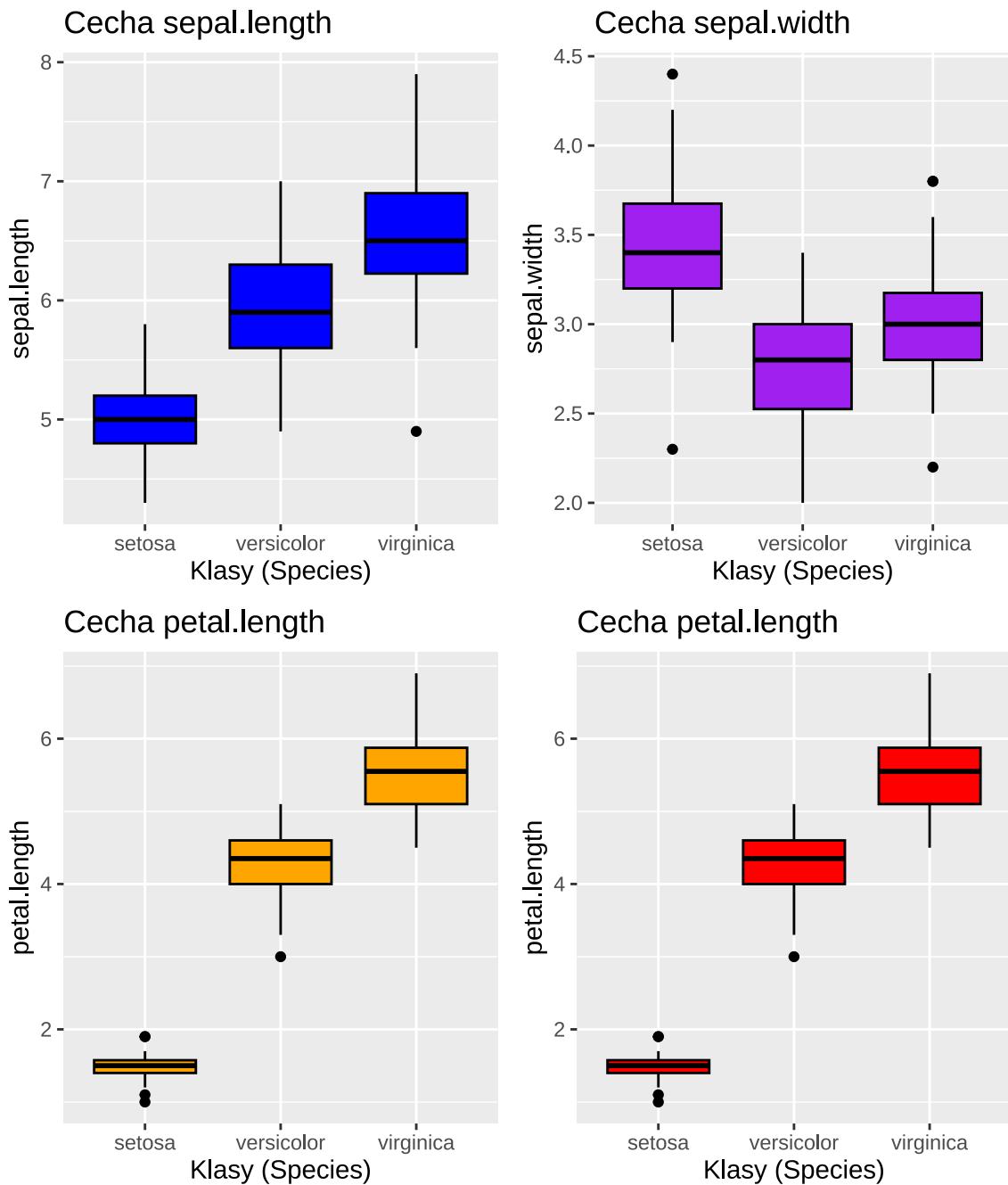
Cecha petal.length



Cecha petal.width



Rysunek 2: Wykresy rozrzutu dla poszczególnych cech



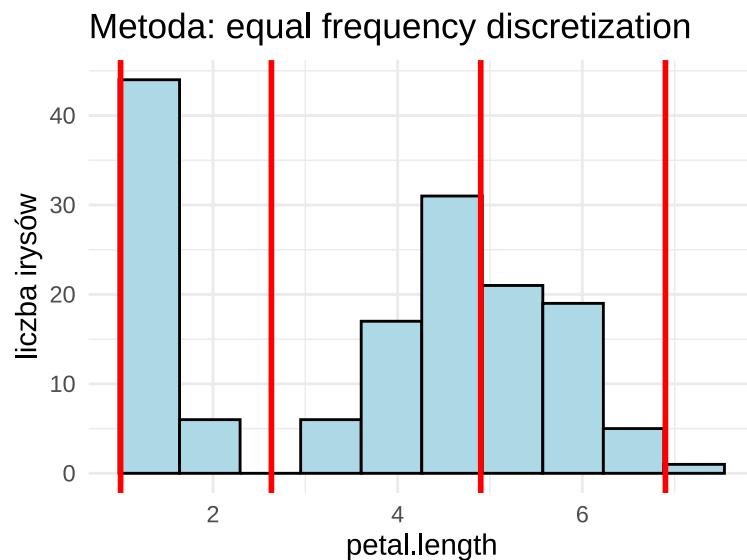
Rysunek 3: Wykresy pudełkowe dla poszczególnych cech

Na podstawie wykresów (Rysunki 1, 2, 3) możemy zauważyć, że cecha **petal.length** ma najlepsze zdolności dyskryminacyjne, ponieważ najlepiej oddziela klasy irysów. Natomiast cecha **sepal.width** ma najgorsze zdolności dyskryminacyjne, ponieważ nie oddziela klas irysów.

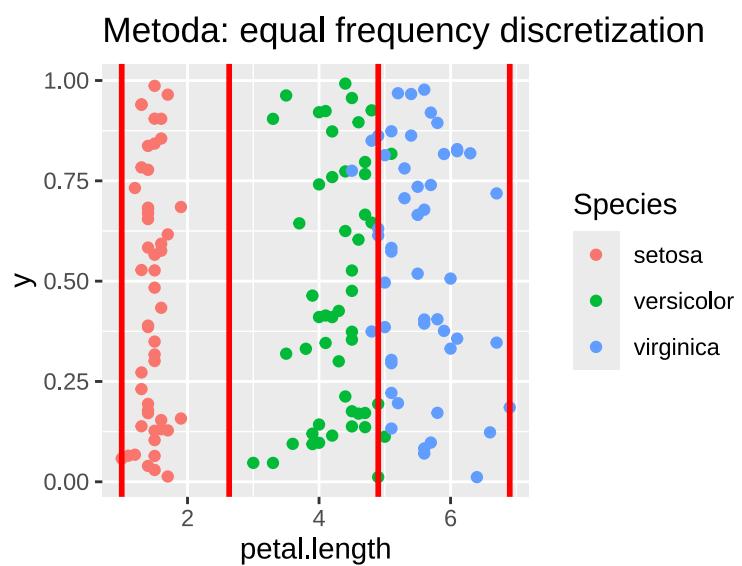
## 1.1 Metoda oparta na równych częstościach

### 1.1.1 Cecha petal.length

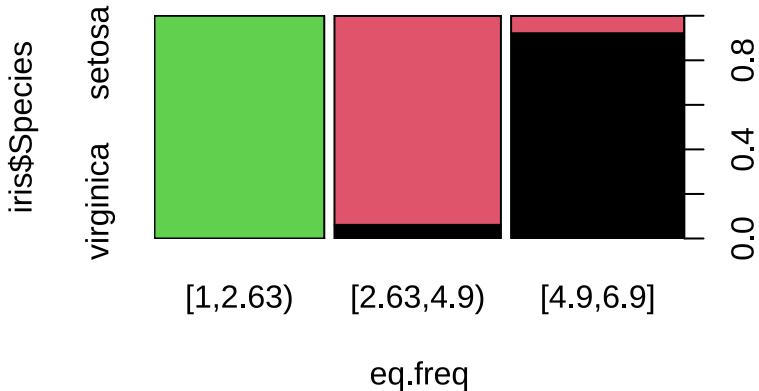
```
## eq.freq  
## [1,2.63) [2.63,4.9) [4.9,6.9]  
##      50       49       51
```



Rysunek 4: Dyskretyzacja - metoda: equal frequency discretization - porównanie z rzeczywistymi klasami



Rysunek 5: Dyskretyzacja - metoda: equal frequency discretization - wykres rozrzutu



Rysunek 6: Dyskretyzacja - metoda: equal frequency discretization - wykres mozaikowy

Tabela 1: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

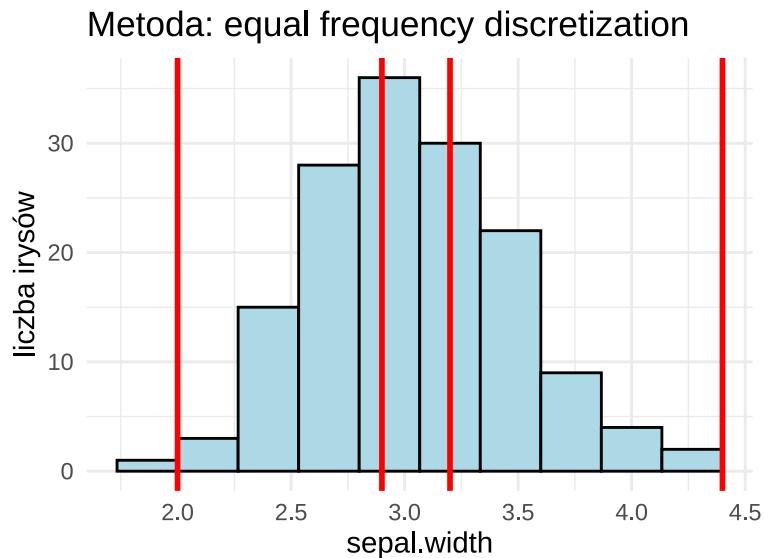
	setosa	versicolor	virginica
[1,2.63)	50	0	0
[2.63,4.9)	0	46	3
[4.9,6.9]	0	4	47

```
## Cases in matched pairs: 95.33 %
##      [1,2.63)   [2.63,4.9)   [4.9,6.9]
## "setosa" "versicolor" "virginica"
```

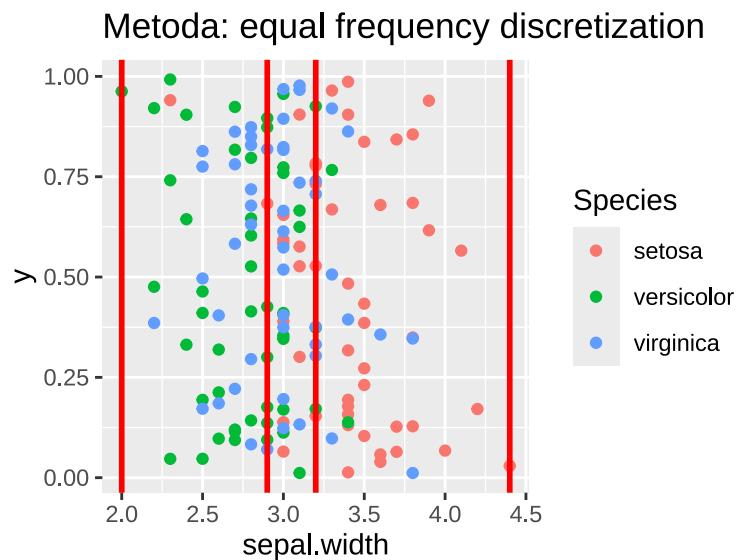
Wyniki dyskretyzacji metodą równych częstości dla cechy **petal.length** przedstawiono na Rysunku 4, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 5. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 6) oraz tabela kontyngencji (Tabela 1). Ponadto wynik funkcji **matchClasses()** wskazuje na 95.33 % zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

### 1.1.2 Cecha sepal.width

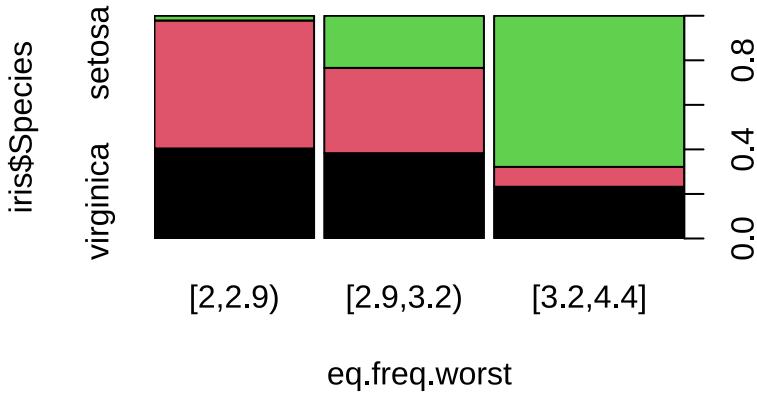
```
## eq.freq.worst  
## [2,2.9) [2.9,3.2) [3.2,4.4]  
##      47          47          56
```



Rysunek 7: Dyskretyzacja - metoda: equal frequency discretization - porównanie z rzeczywistymi klasami



Rysunek 8: Dyskretyzacja - metoda: equal frequency discretization - wykres rozrzutu



Rysunek 9: Dyskretyzacja - metoda: equal frequency discretization - wykres mozaikowy

Tabela 2: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

	setosa	versicolor	virginica
<code>[2,2.9)</code>	1	27	19
<code>[2.9,3.2)</code>	11	18	18
<code>[3.2,4.4]</code>	38	5	13

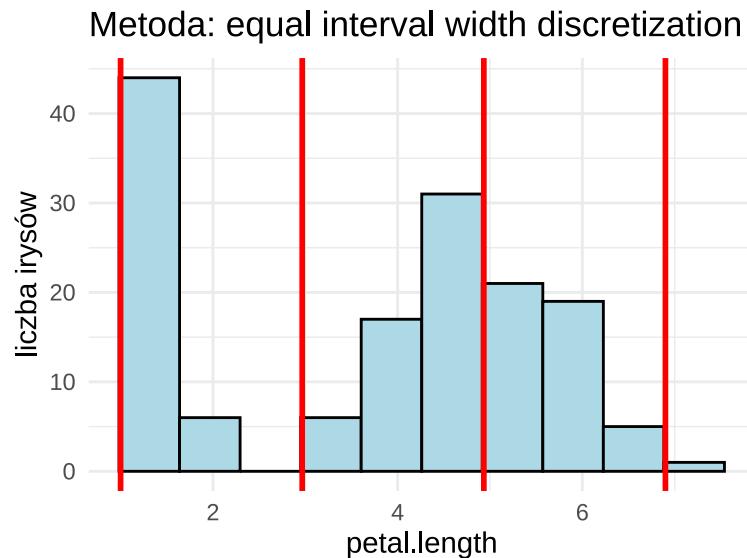
```
## Cases in matched pairs: 55.33 %
##      [2,2.9)    [2.9,3.2)    [3.2,4.4]
## "versicolor" "versicolor"   "setosa"
```

Wyniki dyskretyzacji metodą równych częstości dla cechy `sepal.width` przedstawiono na Rysunku 7, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 8. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 9) oraz tabela kontyngencji (Tabela 2). Ponadto wynik funkcji `matchClasses()` wskazuje na 55.33 % zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

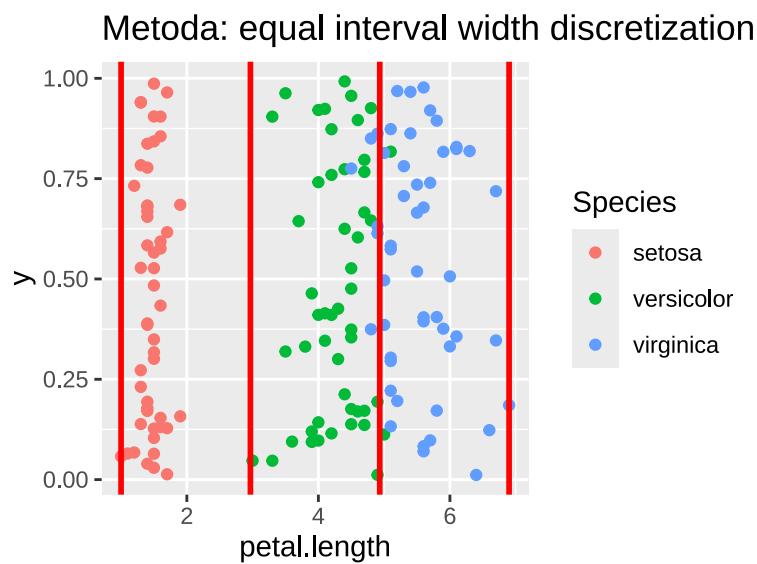
## 1.2 Metoda oparta na przedziałach o jednakowej szerokości

### 1.2.1 Cecha petal.length

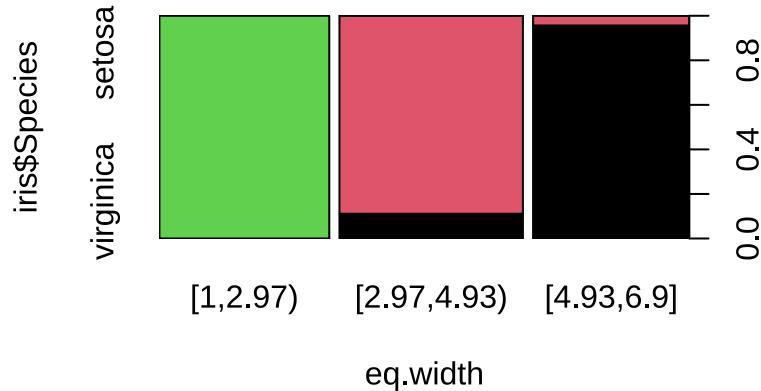
```
## eq.width  
## [1,2.97) [2.97,4.93) [4.93,6.9]  
##      50          54          46
```



Rysunek 10: Dyskretyzacja - metoda: equal interval width - porównanie z rzeczywistymi klasami



Rysunek 11: Dyskretyzacja - metoda: equal interval length - wykres rozrzutu



Rysunek 12: Dyskretyzacja - metoda: equal interval width discretization - wykres mozaikowy

Tabela 3: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

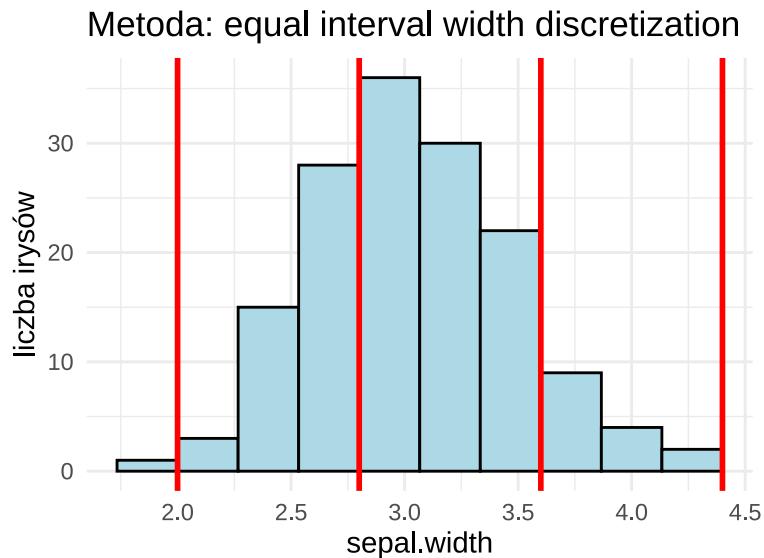
	setosa	versicolor	virginica
[1,2.97)	50	0	0
[2.97,4.93)	0	48	6
[4.93,6.9]	0	2	44

```
## Cases in matched pairs: 94.67 %
##      [1,2.97)  [2.97,4.93)  [4.93,6.9]
## "setosa" "versicolor" "virginica"
```

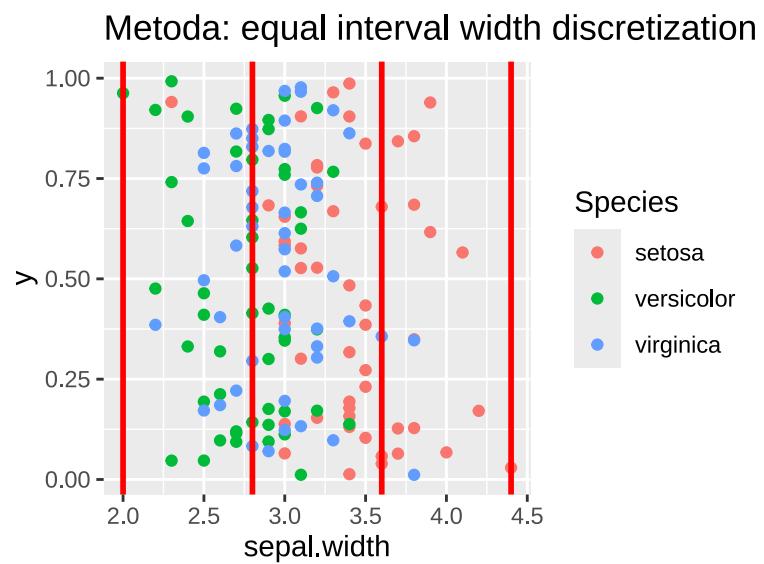
Wyniki dyskretyzacji metodą opartą na przedziałach o jednakowej szerokości dla cechy **petal.length** przedstawiono na Rysunku 7, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 11. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 12) oraz tabela kontyngencji (Tabela 3). Ponadto wynik funkcji **matchClasses()** wskazuje na 94.67 % zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

### 1.2.2 Cecha sepal.width

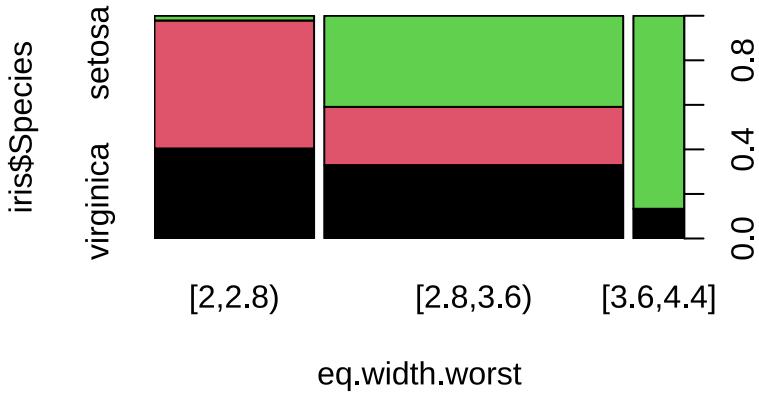
```
## eq.width.worst  
## [2,2.8) [2.8,3.6) [3.6,4.4]  
##      47          88         15
```



Rysunek 13: Dyskretyzacja - metoda: equal interval width - porównanie z rzeczywistymi klasami



Rysunek 14: Dyskretyzacja - metoda: equal interval width - wykres rozrzutu



Rysunek 15: Dyskretyzacja - metoda: equal interval width discretization - wykres mozaikowy

Tabela 4: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

	setosa	versicolor	virginica
[2,2.8)	1	27	19
[2.8,3.6)	36	23	29
[3.6,4.4]	13	0	2

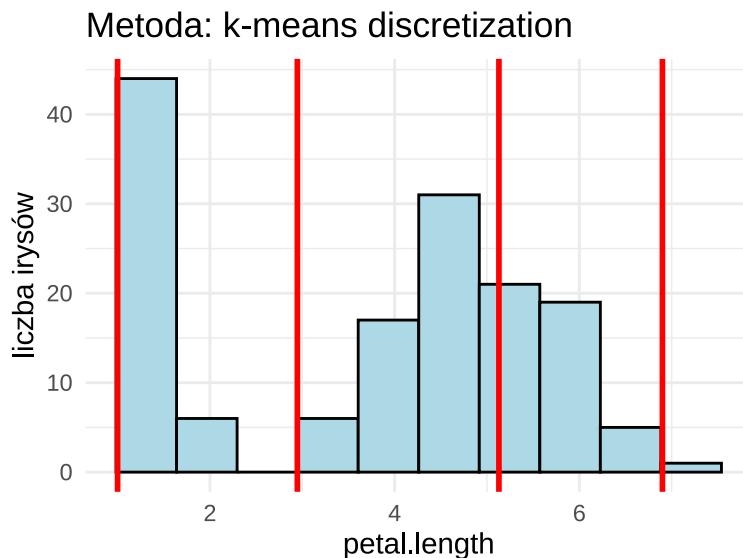
```
## Cases in matched pairs: 50.67 %
##      [2,2.8)    [2.8,3.6)    [3.6,4.4]
## "versicolor"   "setosa"     "setosa"
```

Wyniki dyskretyzacji metodą opartą na przedziałach o jednakowej szerokości dla cechy **sepal.width** przedstawiono na Rysunku 13, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 14. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 15) oraz tabela kontyngencji (Tabela 4). Ponadto wynik funkcji **matchClasses()** wskazuje na 50.67 % % zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

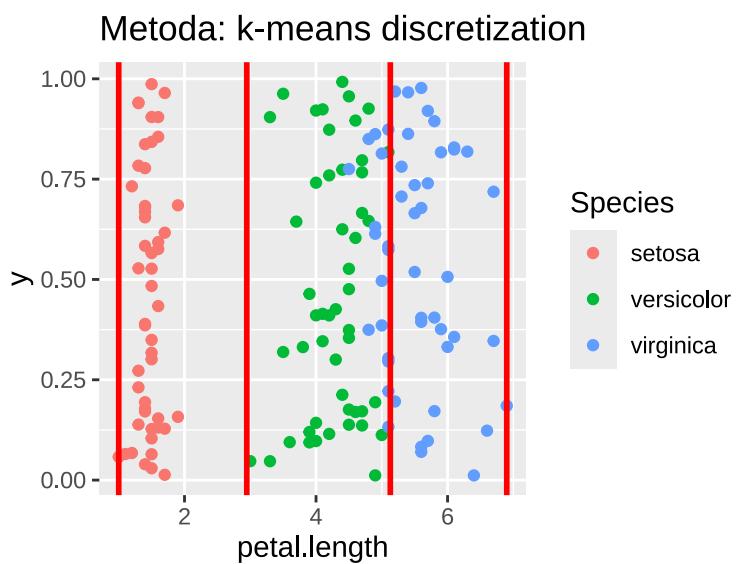
## 1.3 Metoda oparta na algorytmie grupowania (algorytm k-srednich)

### 1.3.1 Cecha petal.length

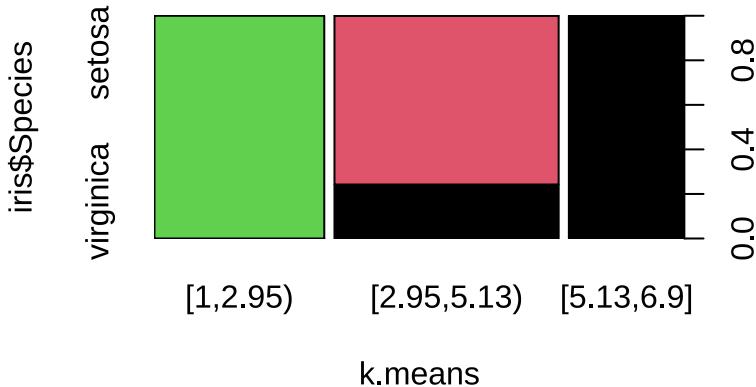
```
## k.means  
##      [1,2.95) [2.95,5.13)  [5.13,6.9]  
##          50           66           34
```



Rysunek 16: Dyskretyzacja - metoda: k-means - porównanie z rzeczywistymi klasami



Rysunek 17: Dyskretyzacja - metoda: k-means - wykres rozrzutu



Rysunek 18: Dyskretyzacja - metoda: k-means discretization - wykres mozaikowy

Tabela 5: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

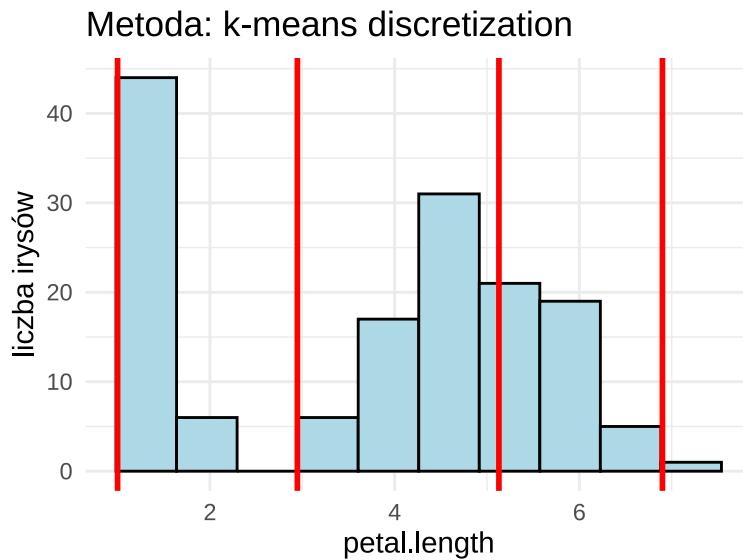
	setosa	versicolor	virginica
[1,2.95)	50	0	0
[2.95,5.13)	0	50	16
[5.13,6.9]	0	0	34

```
## Cases in matched pairs: 89.33 %
##      [1,2.95)  [2.95,5.13)  [5.13,6.9]
## "setosa" "versicolor" "virginica"
```

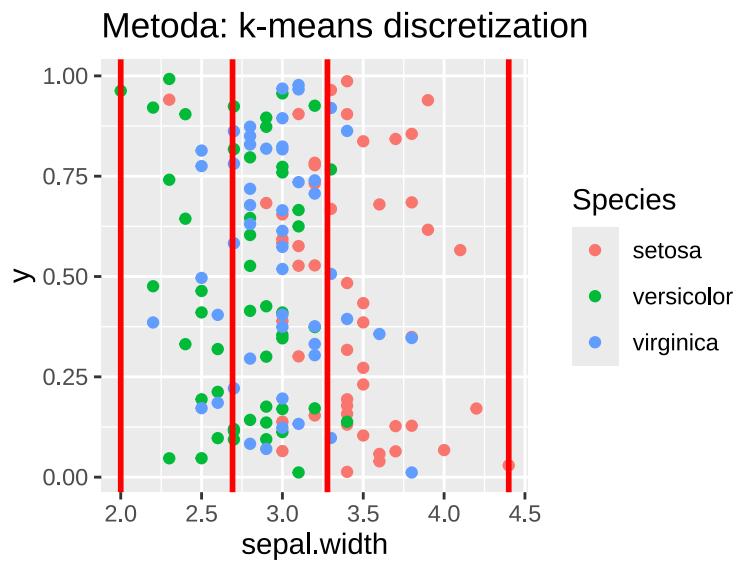
Wyniki dyskretyzacji metodą opartą na algorytmie grupowania k-średnich dla cechy **petal.length** przedstawiono na Rysunku 16, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 17. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 18) oraz tabela kontyngencji (Tabela 5). Ponadto wynik funkcji **matchClasses()** wskazuje na około 90 % (w zależności od wybranych początkowych środków) zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

### 1.3.2 Cecha sepal.width

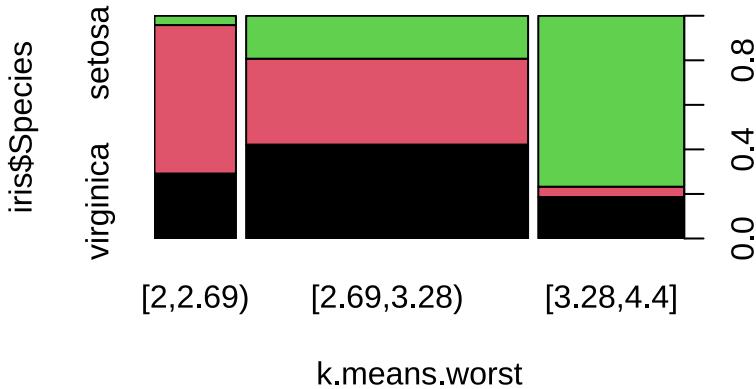
```
## k.means.worst  
##      [2,2.69) [2.69,3.28)  [3.28,4.4]  
##          24           83           43
```



Rysunek 19: Dyskretyzacja - metoda: k-means - porównanie z rzeczywistymi klasami



Rysunek 20: Dyskretyzacja - metoda: k-means - wykres rozrzutu



Rysunek 21: Dyskretyzacja - metoda: k-means discretization - wykres mozaikowy

Tabela 6: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

	setosa	versicolor	virginica
[2,2.69)	1	16	7
[2.69,3.28)	16	32	35
[3.28,4.4]	33	2	8

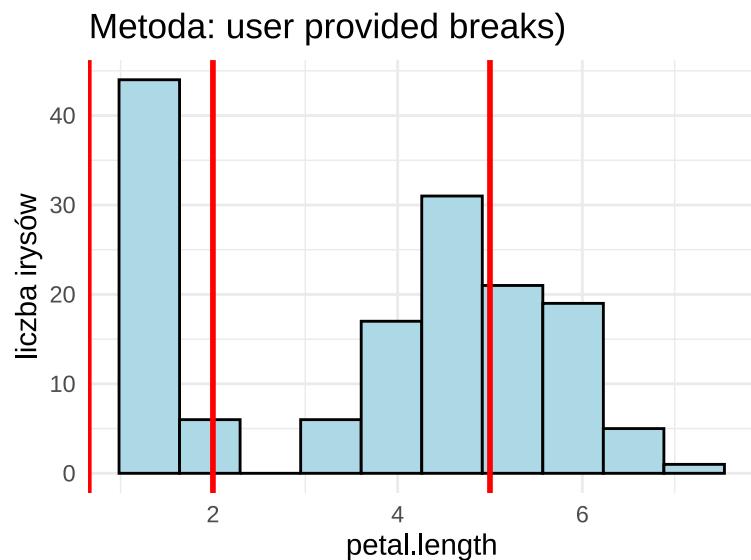
```
## Cases in matched pairs: 56 %
##      [2,2.69)  [2.69,3.28)  [3.28,4.4]
## "versicolor" "virginica"   "setosa"
```

Wyniki dyskretyzacji metodą opartą na algorytmie grupowania k-srednich dla cechy **sepal.width** przedstawiono na Rysunku 19, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 20. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 21) oraz tabela kontyngencji (Tabela 6). Ponadto wynik funkcji **matchClasses()** wskazuje na około 50 % (w zależności od wybranych początkowych środków) zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

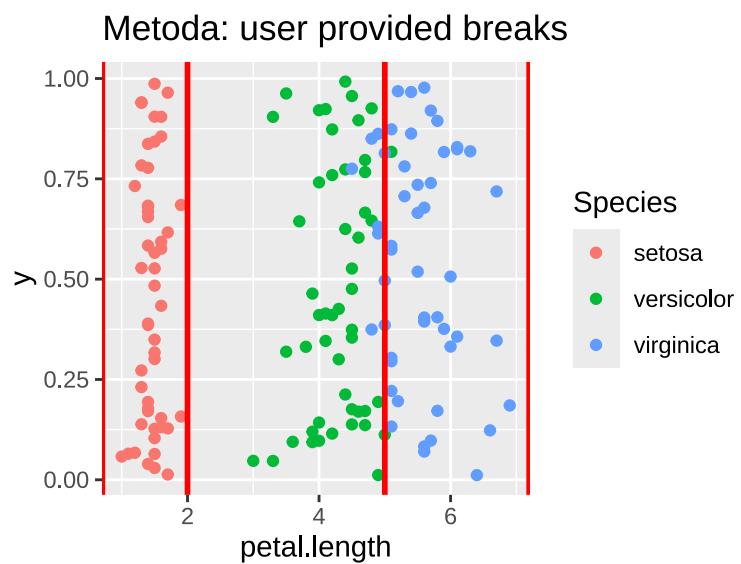
## 1.4 Dyskretyzacja z przedziałami zadanymi przez użytkownika

### 1.4.1 Cecha petal.length

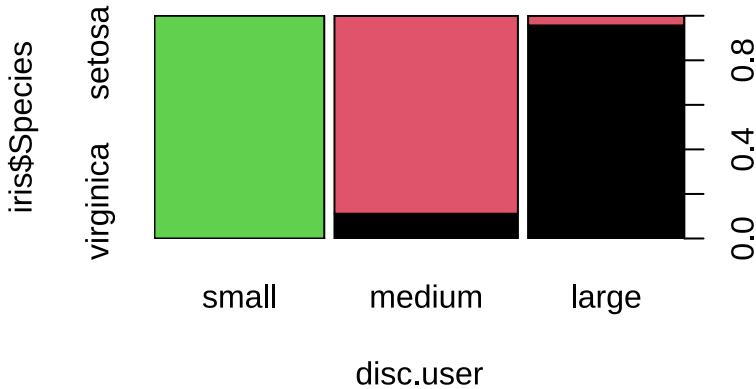
```
## disc.user  
##   small   medium    large  
##     50      54       46
```



Rysunek 22: Dyskretyzacja z przedziałami zadanymi przez użytkownika - porównanie z rzeczywistymi klasami



Rysunek 23: Dyskretyzacja z przedziałami zadanymi przez użytkownika - wykres rozrzutu



Rysunek 24: Dyskretyzacja z przedziałami zadanymi przez użytkownika - wykres mozaikowy

Tabela 7: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

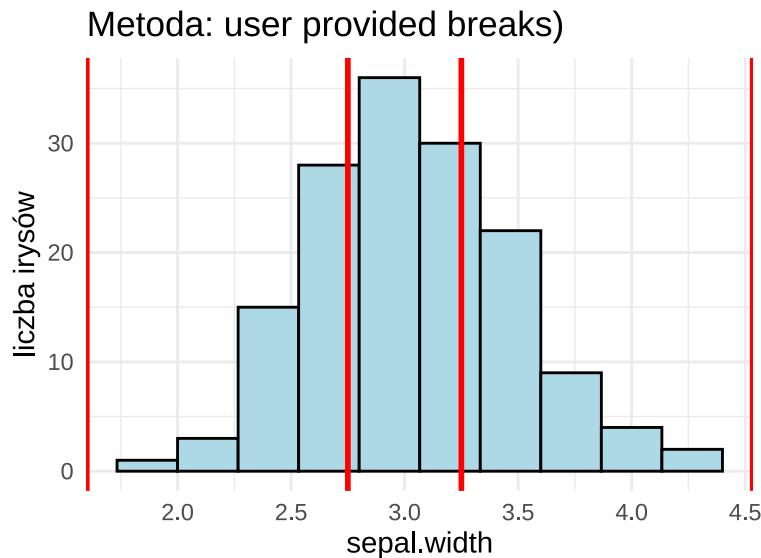
	setosa	versicolor	virginica
small	50	0	0
medium	0	48	6
large	0	2	44

```
## Cases in matched pairs: 94.67 %
##      small     medium     large
## "setosa" "versicolor" "virginica"
```

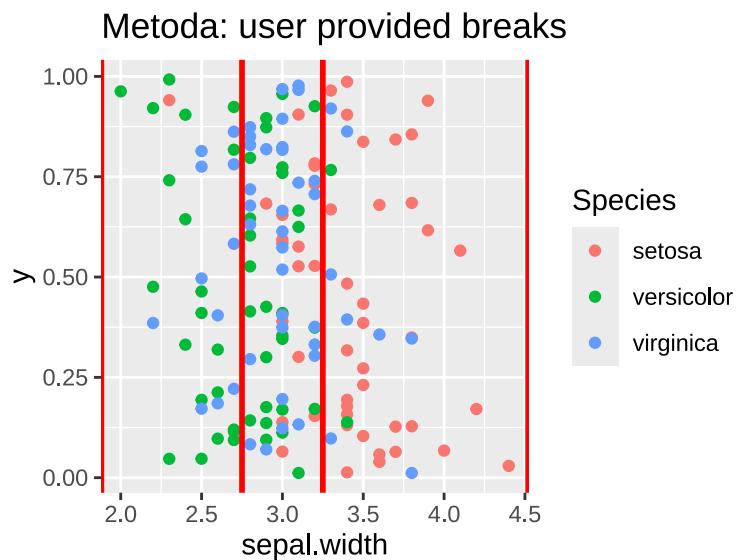
Wyniki dyskretyzacji metodą wyboru przedziałów przez użytkownika dla cechy **petal.length** przedstawiono na Rysunku 22, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 23. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 24) oraz tabela kontyngencji (Tabela 7). Ponadto wynik funkcji **matchClasses()** wskazuje na 94.67 % zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

#### 1.4.2 Cecha sepal.width

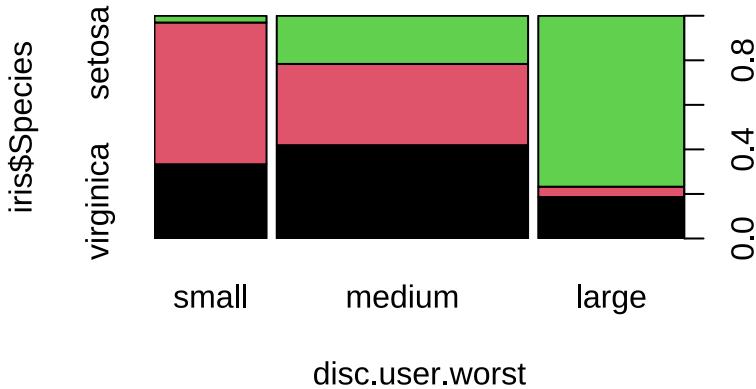
```
## disc.user.worst  
##   small medium  large  
##     33      74      43
```



Rysunek 25: Dyskretyzacja z przedziałami zadanymi przez użytkownika - porównanie z rzeczywistymi klasami



Rysunek 26: Dyskretyzacja z przedziałami zadanymi przez użytkownika - wykres rozrzutu



Rysunek 27: Dyskretyzacja z przedziałami zadanymi przez użytkownika - wykres mozaikowy

Tabela 8: Porównanie przedziałów dyskretyzacji z rzeczywistymi klasami

	setosa	versicolor	virginica
small	1	21	11
medium	16	27	31
large	33	2	8

```
## Cases in matched pairs: 56.67 %
##      small     medium      large
## "versicolor" "virginica" "setosa"
```

Wyniki dyskretyzacji metodą wyboru przedziałów przez użytkownika dla cechy **sepal.width** przedstawiono na Rysunku 25, gdzie zaznaczono wyznaczone przedziały. Rozkład obserwacji w przestrzeni cechy z uwzględnieniem gatunków uwidoczniono na Rysunku 26. Relację między przedziałami dyskretyzacji a rzeczywistymi klasami gatunków przedstawia wykres mozaikowy (Rysunek 27) oraz tabela kontyngencji (Tabela 8). Ponadto wynik funkcji **matchClasses()** wskazuje na 56.67 % zgodność pomiędzy przedziałami dyskretyzacji a rzeczywistymi klasami gatunków.

## 1.5 Wnioski

- dla **petal.length** (najlepsza cecha) średnia skuteczność metod wyniosła około 95 %
- dla **sepal.width** (najgorsza cecha) średnia skuteczność to około 55 %
- skuteczność metod zależy głównie od dyskryminacyjności cechy
- dla dobrych cech nawet proste metody są efektywne
- dla złych cech żadna metoda nie gwarantuje dobrej separacji

## 2 Analiza składowych głównych (Principal Component Analysis (PCA))

W tej sekcji zajmiemy się analizą składowych głównych (PCA) w oparciu o zbiór danych `uaScoresDataFrame.csv`. Analiza składowych głównych jest techniką redukcji wymiarowości, która pozwala na przekształcenie zbioru danych o wielu zmiennych w zbiór o mniejszej liczbie zmiennych, zwanych składowymi głównymi.

### 2.1 Przygotowanie danych

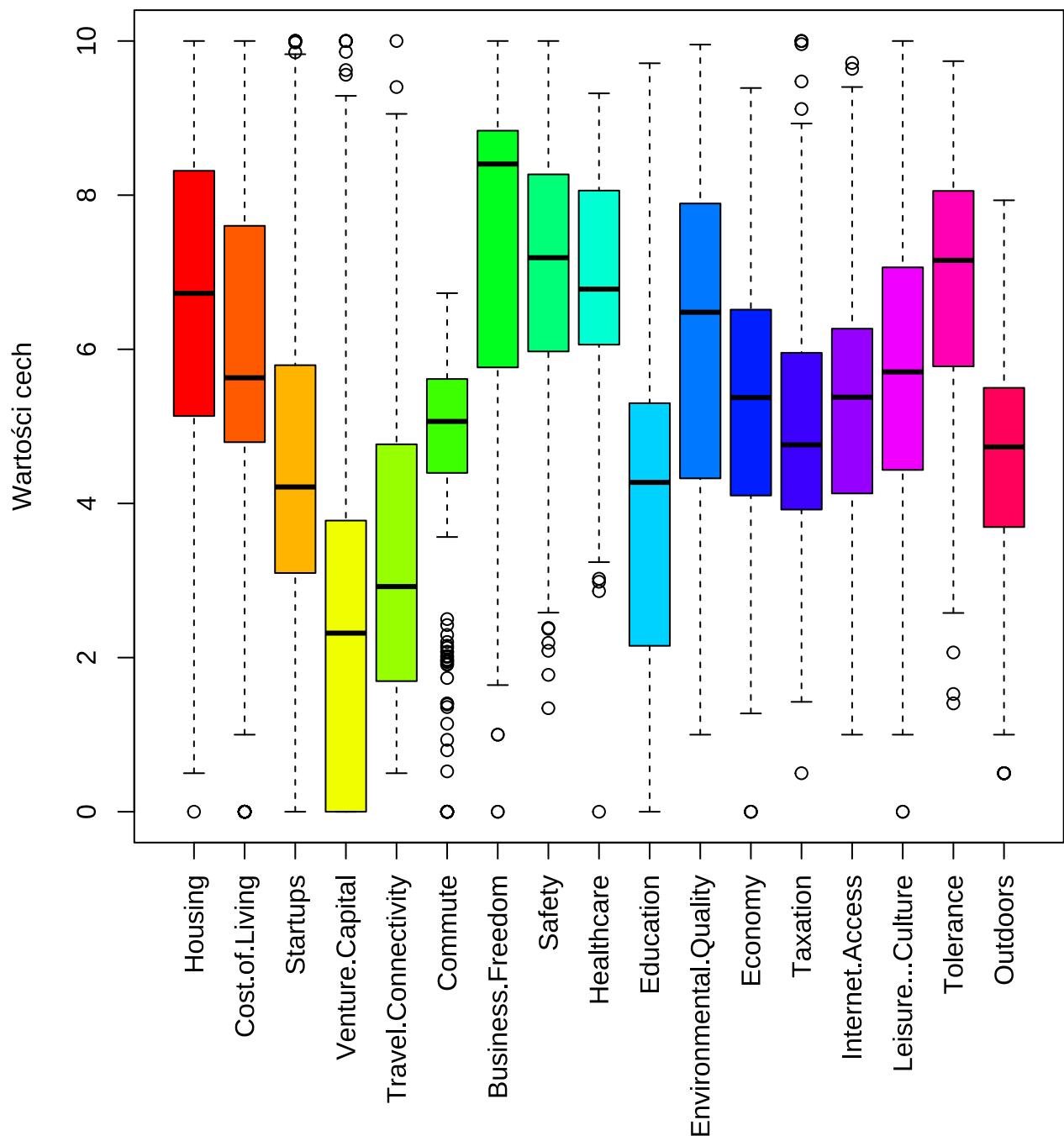
```
## Wymiary zbioru danych: 266 21
```

Zbior danych składa się z 4 cech jakościowych i 17 cech ilościowych. Do dalszej analizy należy wybrać wyłącznie cechy ilościowe, czyli wskaźniki jakości życia w zakresie 0-10

Zmienna	Wariancja
Venture.Capital	6.5201
Cost.of.Living	5.9883
Housing	5.2646
Education	4.8974
Environmental.Quality	4.8396
Startups	4.6347
Business.Freedom	4.4497
Travel.Connectivity	4.3749
Leisure...Culture	4.0271
Internet.Access	3.5052
Safety	3.0507
Tolerance	2.9745
Taxation	2.8554
Outdoors	2.5337
Commute	2.3197
Economy	2.3019
Healthcare	2.1960

Tabela 9: Wariancje poszczególnych cech

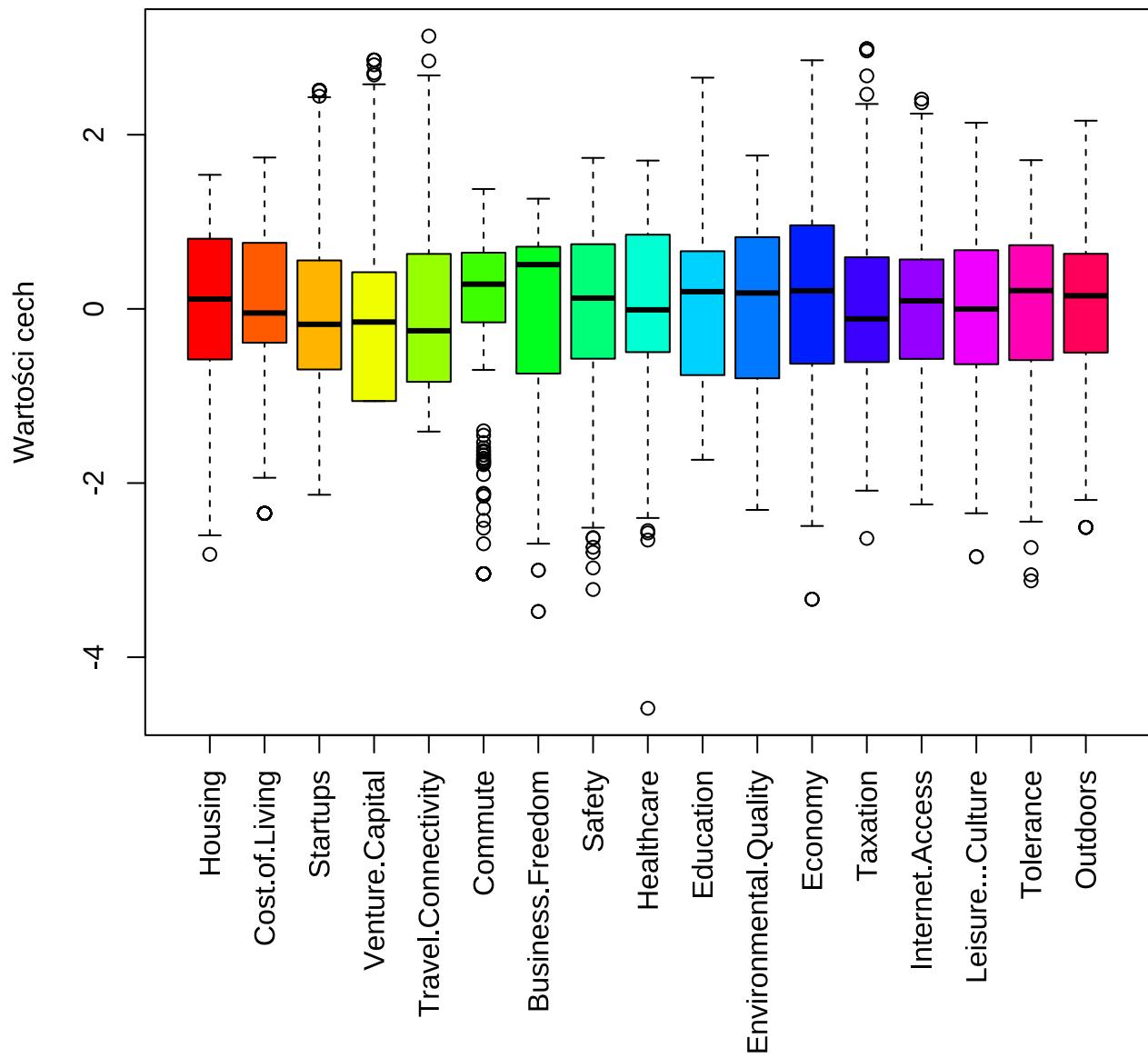
## Wykresy pudełkowe dla poszczególnych cech



Rysunek 28: Wykresy pudełkowe dla poszczególnych cech

Porównanie wariancji (Tabela 9) i wykresów pudełkowych dla poszczególnych zmiennych (Rysunek 28) wskazuje na konieczność uwzględnienia standaryzacji danych

## Wykresy pudełkowe dla poszczególnych cech po standaryzacji

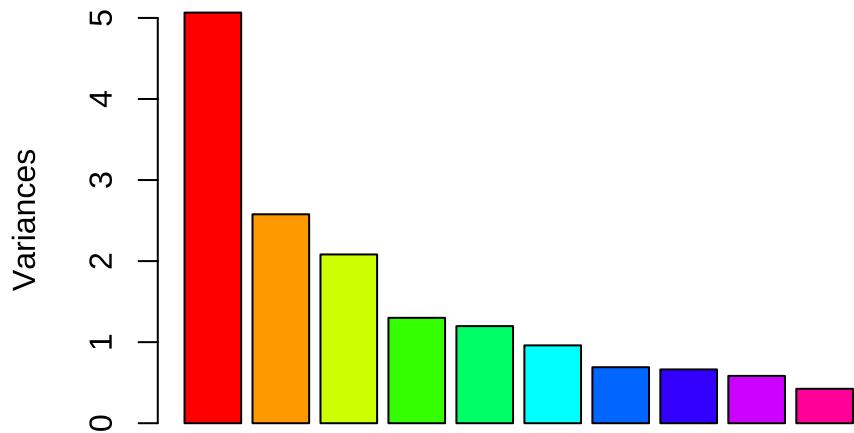


Rysunek 29: Wykresy pudełkowe dla poszczególnych cech po standaryzacji

Wykresy pudełkowe po standaryzacji 29 pozwoliły ujednolicić wariancję zmiennych.

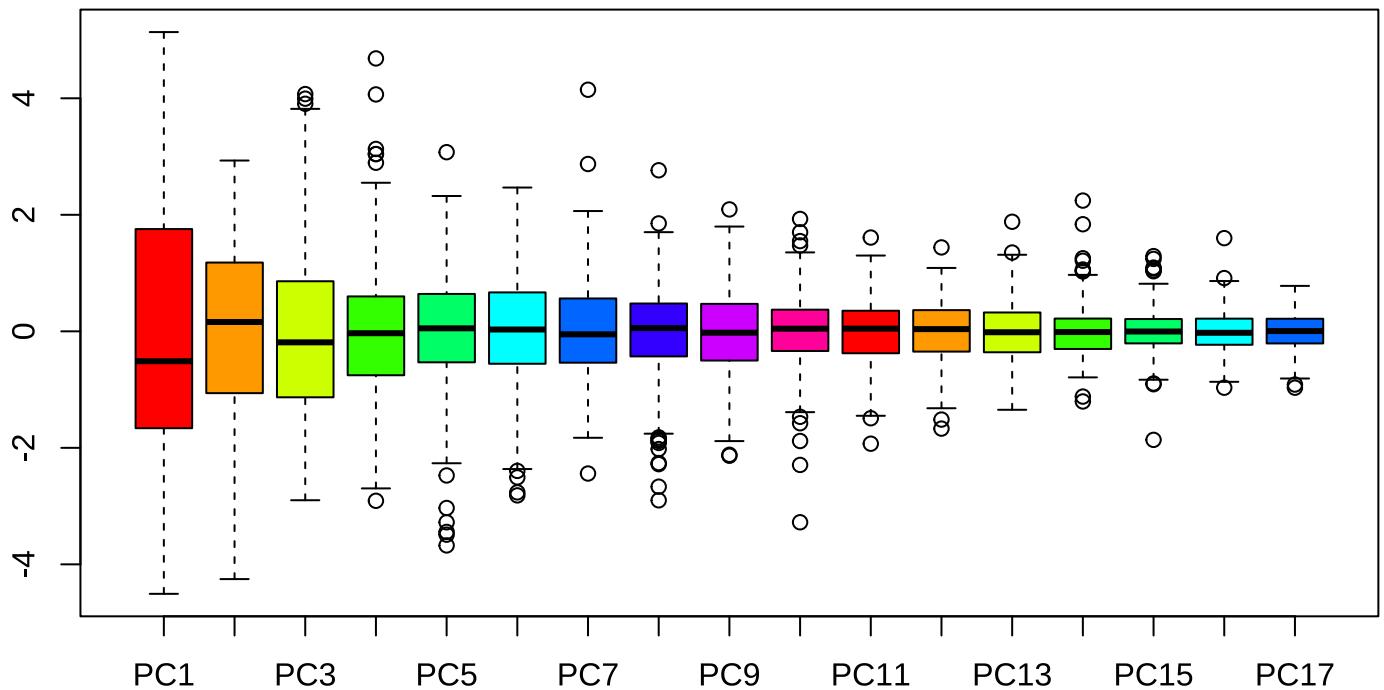
## 2.2 Analiza składowych głównych

### Wariancje poszczególnych składowych



Rysunek 30: Wariancje poszczególnych składowych

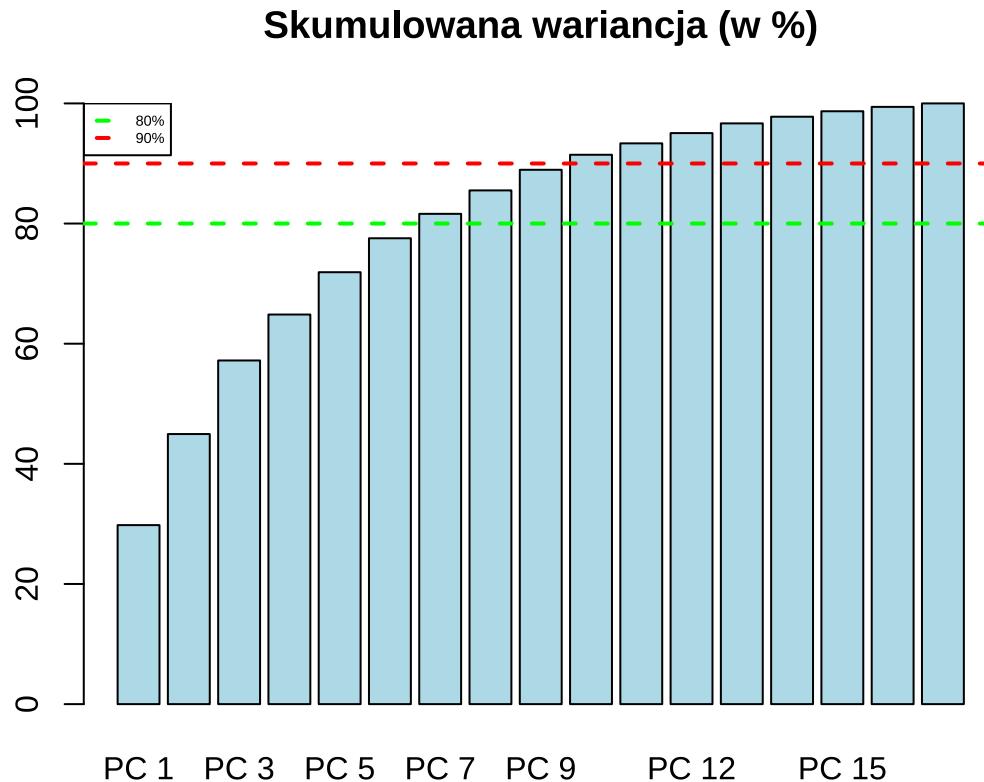
### Wykresy pudełkowe dla poszczególnych składowych głównych



Rysunek 31: Wykresy pudełkowe dla poszczególnych składowych głównych

Na wykresie (Rysunek 30) widzimy względny udział każdej składowej głównej w wyjaśnianiu wariancji danych. Możemy zauważać, że pierwsze kilka składowych dominuje. Wykresy

pułapkowe dla składowych głównych przedstawione (Rysunek 31) pokazują ich rozkład i zmienność. Warto zauważyć, że rozrzut wartości maleje z kolejnymi składowymi, co jest zgodne z zasadą działania PCA.



Rysunek 32: Skumulowana wariancja (w %)

Wnioski: Jak widzimy (Rysunek 32) ograniczając się do siedmiu pierwszych składowych (PC1, PC2, PC3, PC4, PC5, PC6, PC7) zachowujemy ponad 80% całkowitej zmienności danych. Z kolei, dziewięć pierwszych składowych (tj. PC1, PC2, PC3, PC4, PC5, PC6, PC7, PC8, PC9) wyjaśniają prawie 90% całkowitej zmienności.

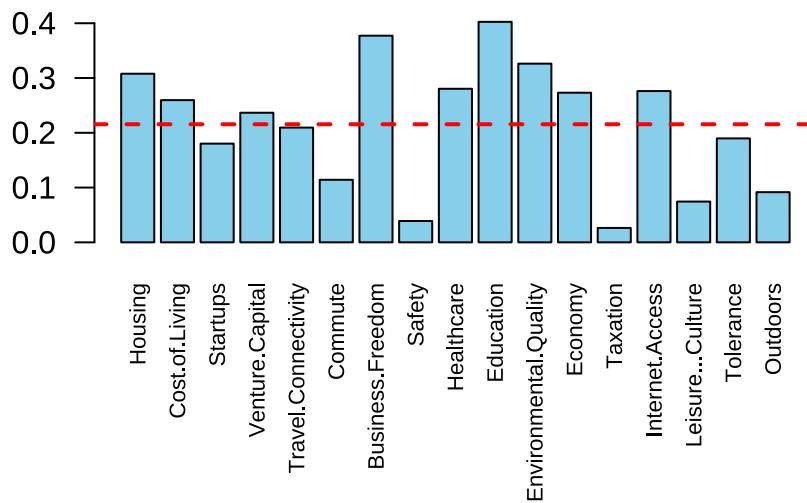
Przyjrzyjmy się wektorom ładunków dla pierwszych trzech składowych głównych:

PC1.Zmienna	PC1.Ladunek	PC2.Zmienna	PC2.Ladunek	PC3.Zmienna	PC3.Ladunek
Education	-0.4026	Startups	-0.4834	Commute	-0.5057
Business.Freedom	-0.3773	Venture.Capital	-0.4275	Travel.Connectivity	-0.3398
Environmental.Quality	-0.3262	Leisure...Culture	-0.3647	Safety	-0.3330
Housing	0.3078	Tolerance	0.3551	Cost.of.Living	-0.3305
Healthcare	-0.2804	Safety	0.2871	Housing	-0.3135
Internet.Access	-0.2762	Environmental.Quality	0.2525	Economy	0.3087
Economy	-0.2732	Healthcare	0.2419	Leisure...Culture	-0.3051
Cost.of.Living	0.2596	Outdoors	-0.1934	Healthcare	-0.2810
Venture.Capital	-0.2366	Cost.of.Living	-0.1758	Outdoors	-0.1486
Travel.Connectivity	-0.2095	Travel.Connectivity	-0.1353	Tolerance	-0.1027
Tolerance	-0.1897	Taxation	0.1074	Education	-0.0739
Startups	-0.1802	Business.Freedom	0.0982	Environmental.Quality	0.0536
Commute	-0.1142	Economy	-0.0740	Internet.Access	0.0284
Outdoors	-0.0916	Housing	0.0534	Business.Freedom	0.0241
Leisure...Culture	-0.0744	Education	-0.0491	Taxation	-0.0202
Safety	-0.0389	Commute	0.0259	Venture.Capital	0.0149
Taxation	0.0263	Internet.Access	0.0227	Startups	0.0061

Tabela 10: Wektory ładunków dla pierwszych trzech składowych głównych (posortowane według wartości bezwzględnej)

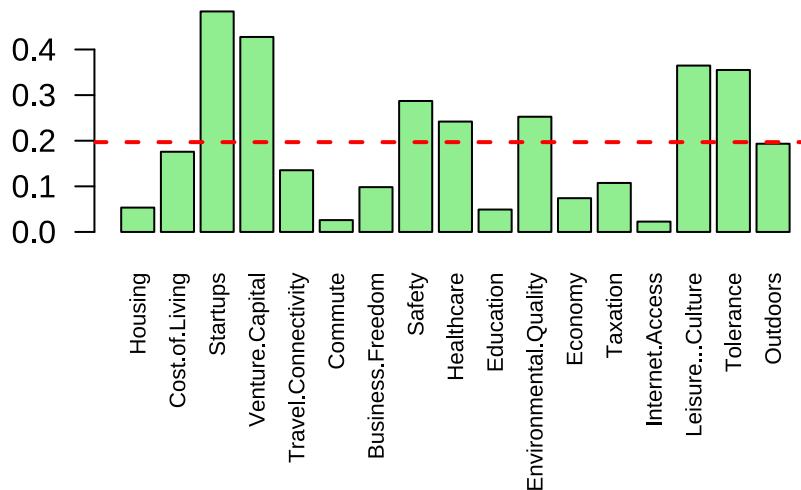
Przedstawione wartości wektorów ładunków (Tabela 10) wskazują na wkład poszczególnych zmiennych oryginalnych do składowych głównych.

### Wkład zmiennych do PC1



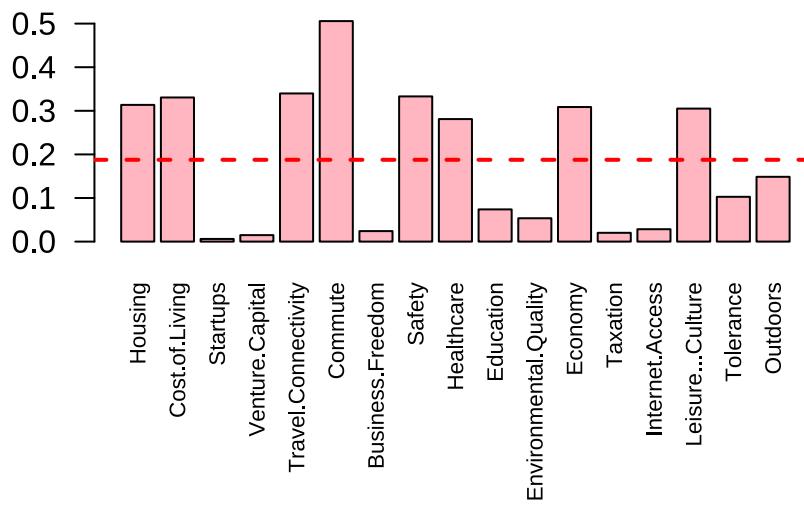
Rysunek 33: Wkład zmiennych do PC1

### Wkład zmiennych do PC2



Rysunek 34: Wkład zmiennych do PC2

### Wkład zmiennych do PC3

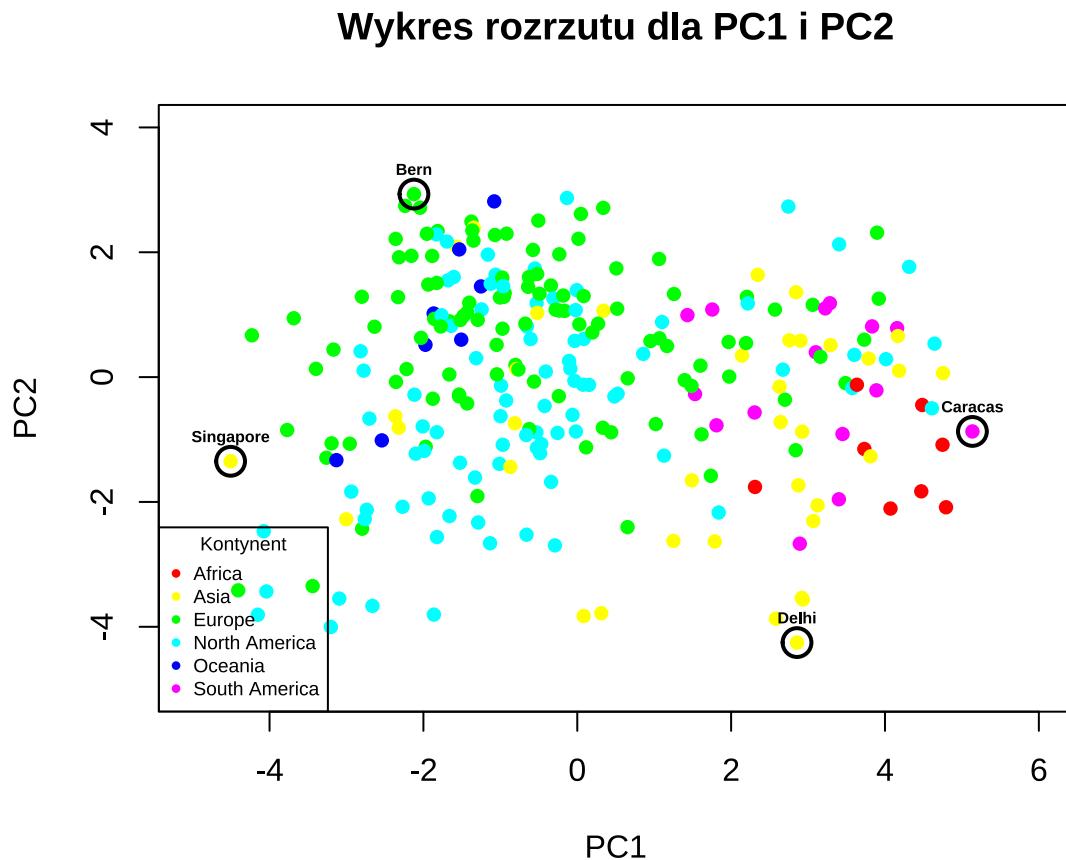


Rysunek 35: Wkład zmiennych do PC3

Na wykresach (Rysunki 33, 34, 35) można zobaczyć bezwzględny wkład zmiennych do poszczególnych składowych głównych. Czerwoną linią zaznaczono średnią wielkość wkładu.

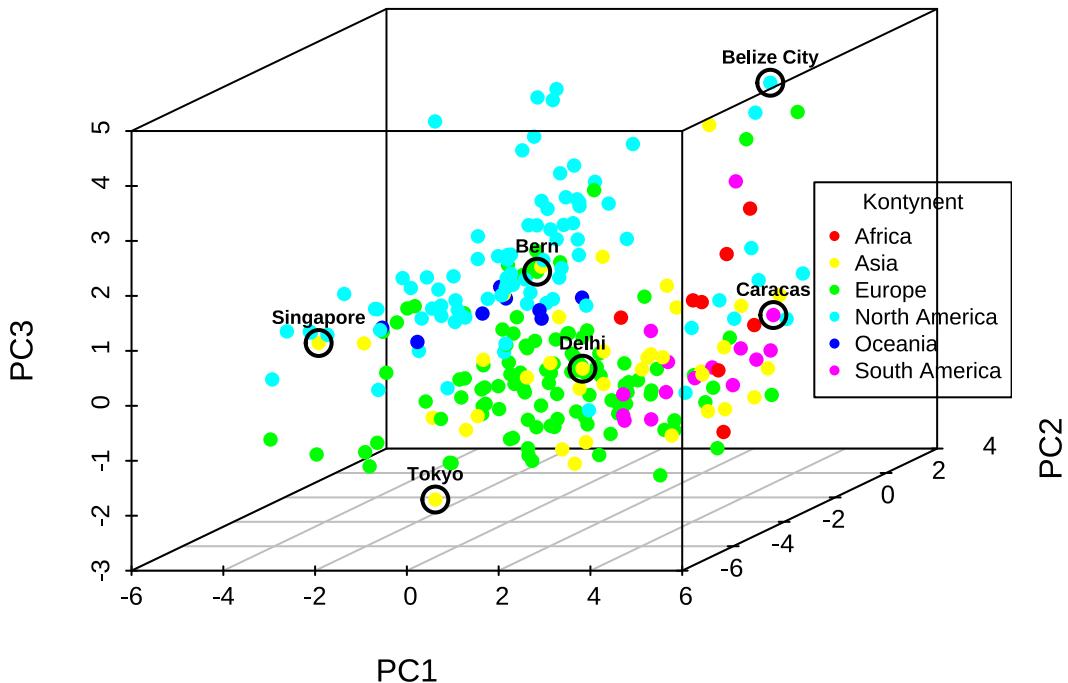
## 2.3 Wizualizacja danych wielowymiarowych

Wykorzystamy wyznaczone składowe główne do wizualizacji danych. Najpierw wygenerujemy wykres rozrzutu dla pierwszych dwóch składowych głównych:



Rysunek 36: Wykres rozrzutu dla pierwszych dwóch składowych głównych

## Wykres rozrzutu 3D dla PC1, PC2 i PC3



Rysunek 37: Wykres rozrzutu dla pierwszych trzech składowych głównych

Na wykresie dwuwymiarowym (Rysunek 36) widoczne jest grupowanie się miast zgodnie z ich przynależnością kontynentalną, co sugeruje istnienie regionalnych wzorców jakości życia. Zauważalne są wyraźne skupiska miast europejskich, azjatyckich i amerykańskich, które formują naturalne grupy w przestrzeni pierwszych dwóch składowych głównych. Ta obserwacja wskazuje, że czynniki geograficzne, kulturowe i ekonomiczne związane z położeniem kontynentalnym mają istotny wpływ na wskaźniki jakości życia.

Wykres trójwymiarowy (Rysunek 37) potwierdza te obserwacje, jednocześnie ujawniając dodatkową strukturę danych dzięki wprowadzeniu trzeciej składowej głównej. Warto zauważyć, że niektóre miasta, które wydawały się podobne w projekcji 2D, wykazują większe zróżnicowanie po uwzględnieniu trzeciej składowej.

Na wykresie 2D zidentyfikowano cztery odstające miasta:

- **Bern** – wyróżnia się wyjątkowo wysokimi wartościami dla PC2, co może wskazywać na wysoką jakość usług publicznych oraz wysokie wskaźniki bezpieczeństwa
- **Singapore** - wyróżnia się skrajnie niskimi wartościami PC1, co może odzwierciedlać jego unikalną kombinację wysokiego rozwoju gospodarczego, niskiego poziomu przestępcości, ale jednocześnie wysokich kosztów
- **Delhi** - wyróżnia się bardzo niskimi wartościami PC2, co może być związane z niższymi wskaźnikami ekonomicznymi i problemami z zanieczyszczeniem powietrza

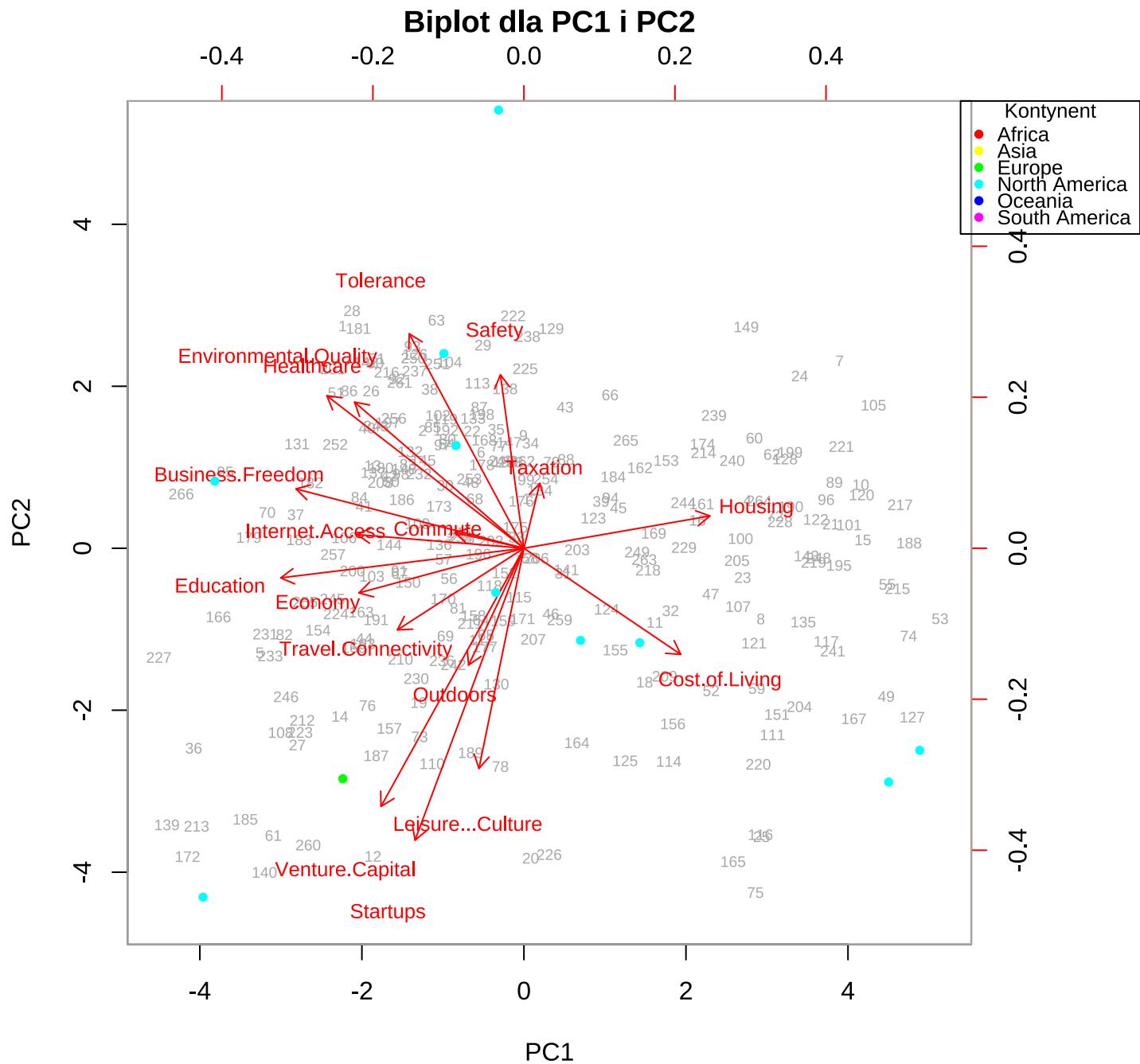
- **Caracas** - wyróżnia się ekstremalnie wysokie wartości PC1, co prawdopodobnie odzwierciedla poważne problemy ekonomiczne i wysoki poziom przestępcości

Na wykresie 3D dostrzegalne są dwa dodatkowe miasta odstające:

- **Tokyo** - wyróżnia się niskimi wartościami PC3, co może odzwierciedlać specyficzne cechy takie jak bardzo wysoka gęstość zaludnienia i zaawansowany system transportu publicznego
- **Belize City** – wyróżnia się wysokimi wartościami PC3, co może być związane z jego unikalnym położeniem geograficznym i specyfiką gospodarki opartej na turystyce

Zidentyfikowane miasta odstające reprezentują skrajne wartości składowych głównych, co świadczy o ich nietypowych warunkach jakości życia w porównaniu do większości analizowanych miast.

## 2.4 Korelacja zmiennych



Rysunek 38: Biplot dla pierwszych dwóch składowych głównych

Na podstawie biplotu (Rysunek 38) zidentyfikowano następujące istotne korelacje:

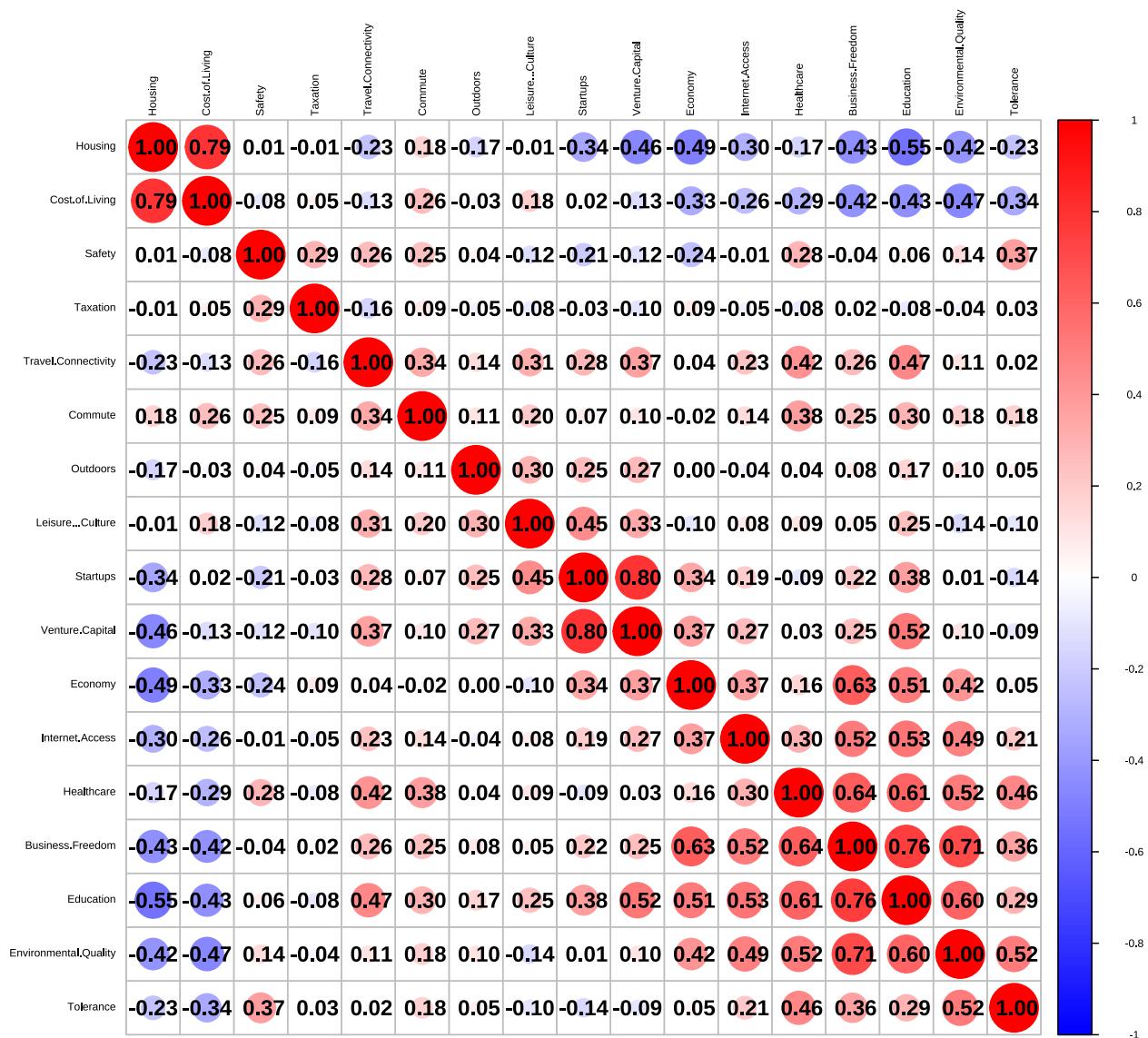
Silne korelacje dodatnie występują między:

- zmiennymi opisującymi jakość usług publicznych (Healthcare, Environmental Quality, Safety)
- wskaźnikami ekonomicznymi (Economy, Business Freedom, Internet Access)
- aspektami kulturalnymi (Leisure & Culture, Outdoors)

Wyraźne korelacje ujemne widoczne są między:

- Cost of Living a jakością usług publicznych
- Housing a zmiennymi Education i Economy
- Taxation a aspektami kulturalnymi

**Macierz korelacji zmiennych**



Rysunek 39: Macierz korelacji zmiennych

Macierz korelacji (Rysunek 39) potwierdza obserwacje z biplotu, dostarczając dokładnych wartości współczynników korelacji:

- współczynniki korelacji między zmiennymi z grupy usług publicznych, wskaźnikami ekonomicznymi i aspektami kulturalnymi są dodatnie
- ujemne korelacje między Cost of Living a wskaźnikami usług publicznych mają wartości w zakresie od -0.3 do -0.6

Obie metody analizy korelacji prowadzą do spójnych wniosków. Biplot dostarcza intuicyjnej wizualizacji struktury zależności, podczas gdy macierz korelacji precyzuje liczbowe wartości tych zależności.

## 2.5 Wnioski

- Potrzebne składowe: Pierwsze 4-5 składowych głównych wyjaśnia około 70-75% wariancji danych, co stanowi zadowalającą reprezentację zbioru. Pierwsze dwie składowe wyjaśniają około 40-45% wariancji.
- Interpretacja składowych:
  - PC1: Reprezentuje ogólną jakość życia i rozwój miasta (silnie związana z usługami publicznymi, bezpieczeństwem i rozwojem gospodarczym)
  - PC2: Kontrastuje aspekty kosztowe (mieszkanie, koszty życia) z jakością usług miejskich
- Grupy miast:
  - Widoczne jest grupowanie miast według kontynentów
  - Miasta europejskie wykazują tendencję do wysokich wartości w zakresie usług publicznych
  - Miasta północnoamerykańskie często charakteryzują się wysokimi kosztami życia
- Wpływ standaryzacji:
  - Standaryzacja była kluczowa dla uzyskania poprawnych wyników
  - Bez standaryzacji zmienne o większych wartościach bezwzględnych (np. koszty mieszkania) nieproporcjonalnie wpływałyby na wyniki PCA

## 3 Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))

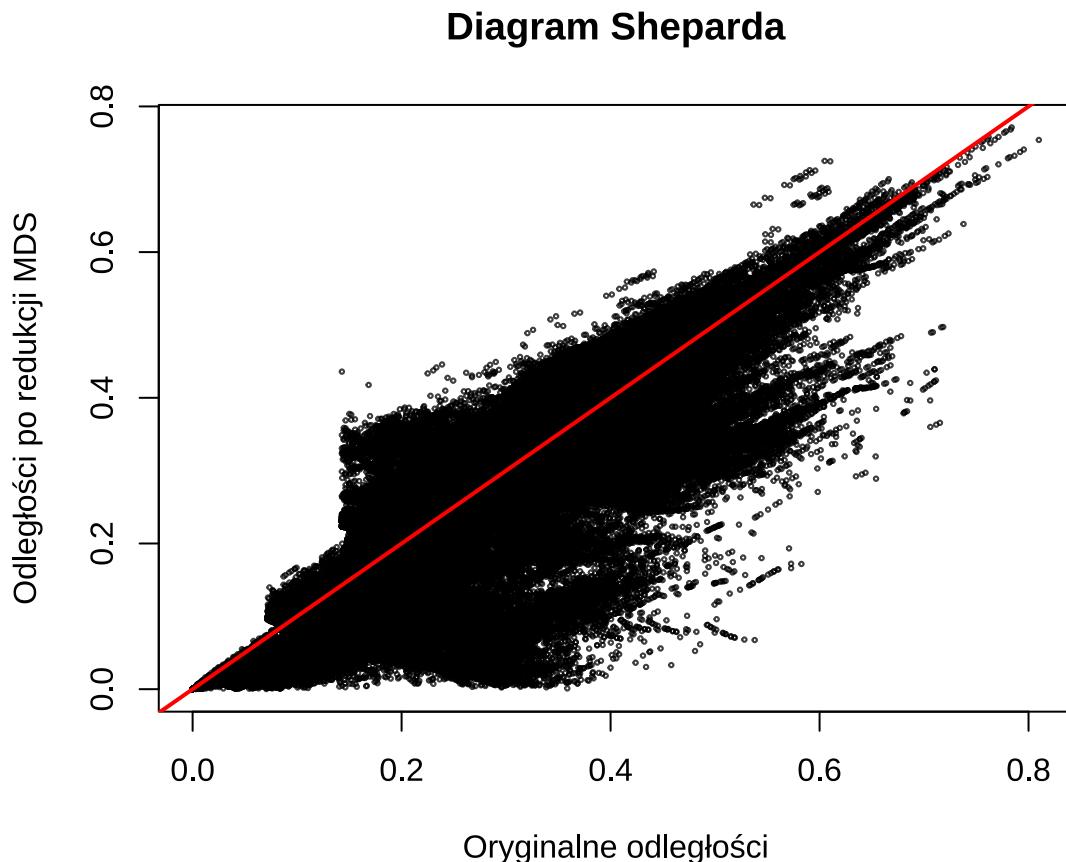
W tej sekcji zajmiemy się analizą danych z zestawu danych **Titanic**, który zawiera informacje o pasażerach statku Titanic. Naszym celem będzie zastosowanie metody MDS do redukcji wymiaru i wizualizacji danych.

### 3.1 Przygotowanie danych

## Wymiary zbioru danych: 714 7

- Usunięto zmienne identyfikacyjne pasażerów: PassengerId, Name, Ticket i Cabin.
- Zidentyfikowano braki danych wyłącznie w zmiennej Age (177 obserwacji)
- Po usunięciu wierszy z brakującymi wartościami uzyskano 714 kompletnych obserwacji.

### 3.2 Redukcja wymiaru na bazie MDS



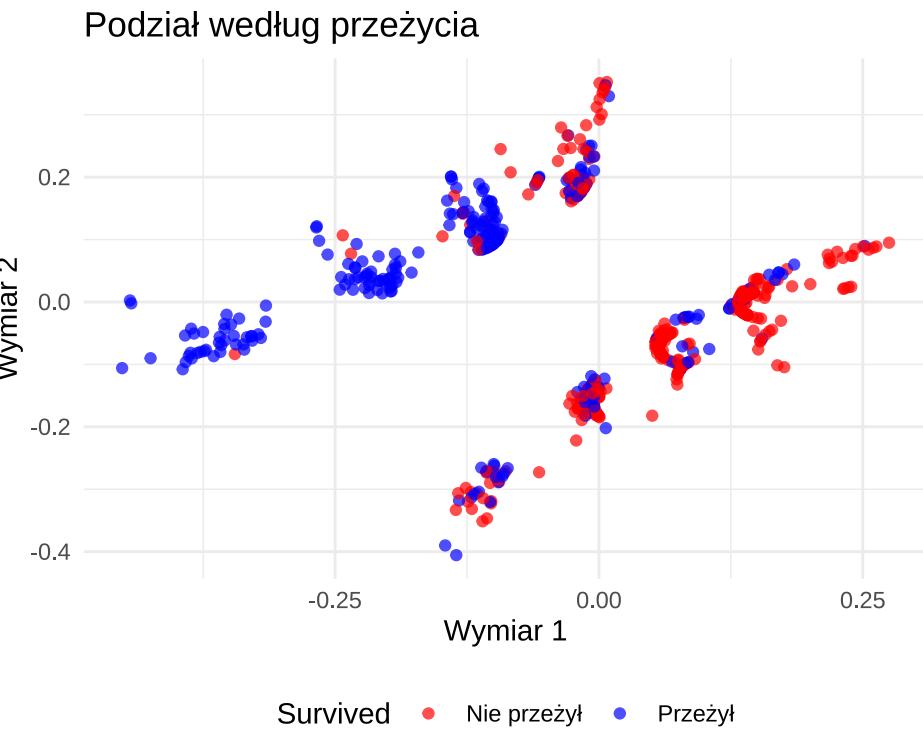
Rysunek 40: Diagram Sheparda dla skalowania MDS

```
## Znormalizowany STRESS: 0.246
```

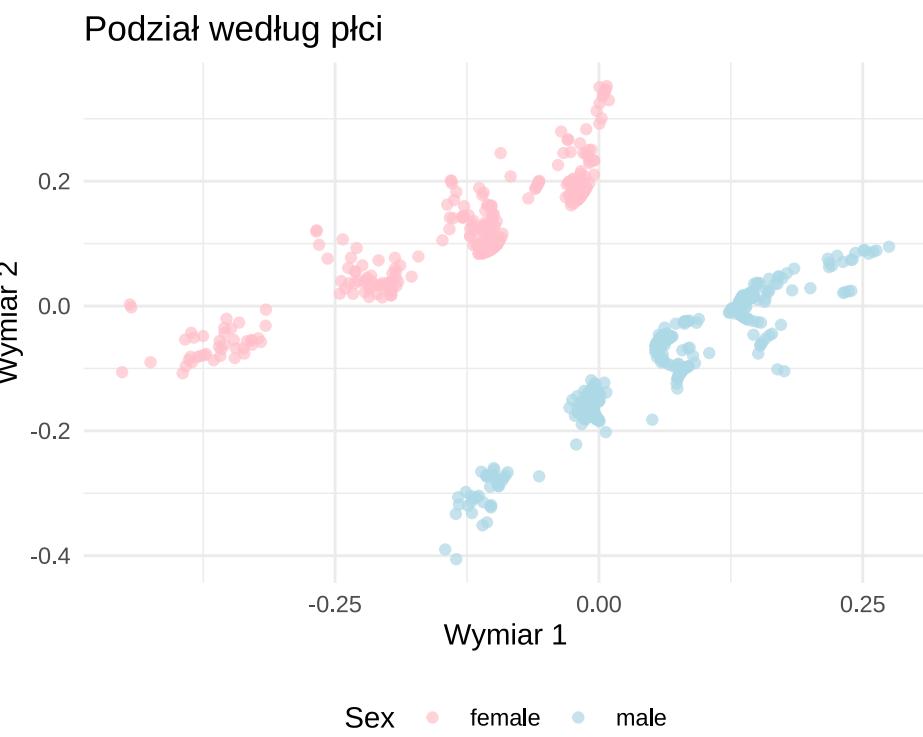
Diagram Sheparda (Rysunek 40) przedstawia zależność między oryginalnymi odległościami a odległościami w przestrzeni 2D po redukcji. Punkty leżące blisko czerwonej linii oznaczają dobrą zgodność między strukturami odległości.

Wartość STRESS wskazuje na jakość dopasowania. Im niższa wartość STRESS, tym lepsze dopasowanie. Znormalizowany Stress wyniósł 0.246, co wskazuje na dobrą zgodność struktur odległości.

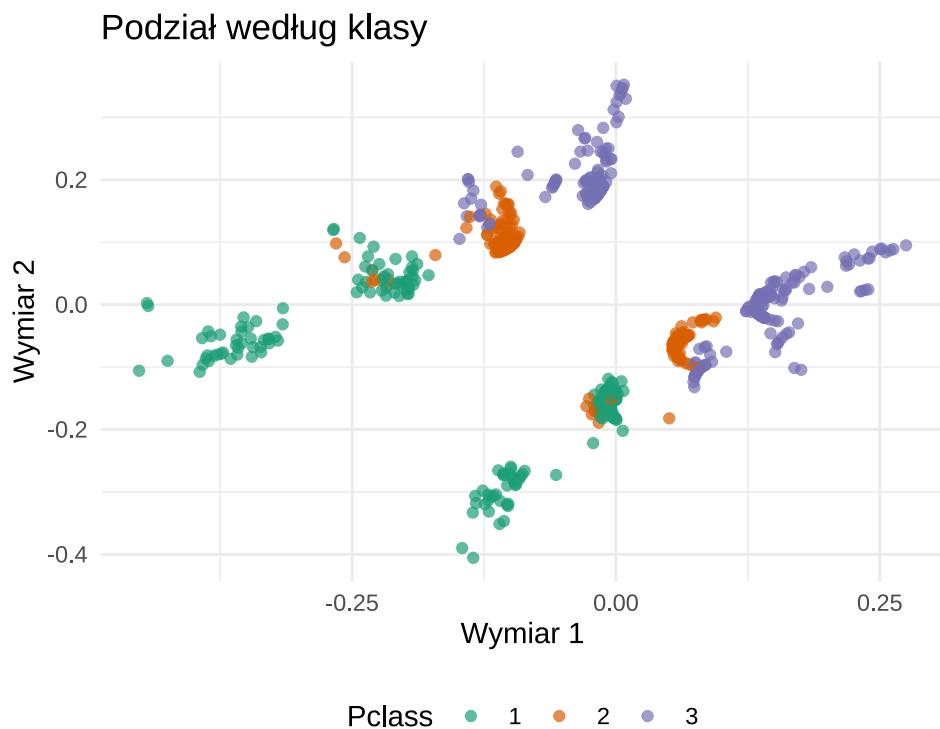
### 3.3 Wizualizacja wyników MDS



Rysunek 41: Rozkład wyników MDS według przeżycia



Rysunek 42: Rozkład wyników MDS według płci



Rysunek 43: Rozkład wyników MDS według klasy pasażerskiej

Podział według przeżycia (Rysunek 41):

- widoczna częściowa separacja grup
- osoby które przeżyły (niebieskie) skupią się w obszarze ujemnych wartości Wymiaru 1
- kilka wyraźnych odstających punktów w górnej części wykresu (potencjalni pasażerowie o nietypowych profilach)

Podział według płci (Rysunek 42):

- silna separacja między grupami, szczególnie wzdłuż Wymiaru 2
- kobiety (różowe) dominują w obszarze dodatnich wartości Wymiaru 2
- mężczyźni (niebieskie) skupieni w dolnej części, co sugeruje istotną rolę płci w strukturze danych

Podział według klasy (Rysunek 43):

- wyraźne przejście kolorów wzdłuż Wymiaru 1
- Klasa 1 (ciemnozielona) skupiona po prawej stronie
- Klasa 3 (jasnozielona) dominuje po lewej stronie wykresu

Potwierdza to silny związek klasy pasażerskiej z pozycją w przestrzeni MDS

### **3.4 Wnioski:**

- struktura danych w przestrzeni MDS jest silnie powiązana ze zmiennymi: Survived, Sex i Pclass
- najlepszą separację wizualną obserwujemy dla płci (Rysunek 42)
- przejście w klasach pasażerskich (Rysunek 43) sugeruje, że Wymiar 1 może być związany ze statusem ekonomicznym