

Sprawozdanie z listy 2

Eksploracja danych

Daria Grzelak, 277533

2025-04-28

Spis treści

1 Dyskretyzacja (przedziałowanie) cech ciągłych	1
1.1 Wykorzystane metody	2
1.2 Opis danych	2
1.3 Wybór cech do dyskretyzacji	2
1.4 Dyskretyzacja z wykorzystaniem różnych metod	6
1.5 Wnioski	18
2 Analiza składowych głównych (PCA – Principal Component Analysis)	18
2.1 Wykorzystane metody	19
2.2 Opis danych	19
2.3 Przygotowanie danych	19
2.4 Wyznaczenie składowych głównych	21
2.5 Zmienna poszczególnych składowych	23
2.6 Wizualizacja danych	24
2.7 Korelacja zmiennych	26
2.8 Wnioski	28
3 Skalowanie wielowymiarowe (MDS – Multidimensional Scaling)	29
3.1 Wykorzystane metody	29
3.2 Przygotowanie i opis danych	29
3.3 Redukcja wymiaru	30
3.4 Wizualizacja danych	32
3.5 Wnioski	35

1 Dyskretyzacja (przedziałowanie) cech ciągłych

W pierwszej części niniejszego raportu przedstawię dyskretyzację cech ciągłych w celu otrzymania prostszego modelu, dla którego łatwiej będzie zinterpretować wyniki.

1.1 Wykorzystane metody

W mojej analizie wykorzystam następujące narzędzia:

- narzędzia analizy opisowej do wybrania najlepszych i najgorszych cech dyskryminacyjnych,
- cztery metody dyskretyzacji nienadzorowanej: według równej częstotliwości, według równej szerokości, oparta na algorytmie k-średnich, o przedziałach zadanych przez użytkownika,
- wykresy wykonane w celu wizualizacji wyników.

1.2 Opis danych

Dane, które będę analizować, to `iris` z R-pakietu datasets, zawierający obserwacje na temat trzech gatunków irysów.

```
# Wczytanie danych
data(iris)
attach(iris)
```

Nazwa	Typ	Opis
Sepal.Length	Liczbowa	Długość działki kielicha
Sepal.Width	Liczbowa	Szerokość działki kielicha
Petal.Length	Liczbowa	Długość płatka
Petal.Width	Liczbowa	Szerokość płatka
Species	Jakościowa	Gatunek

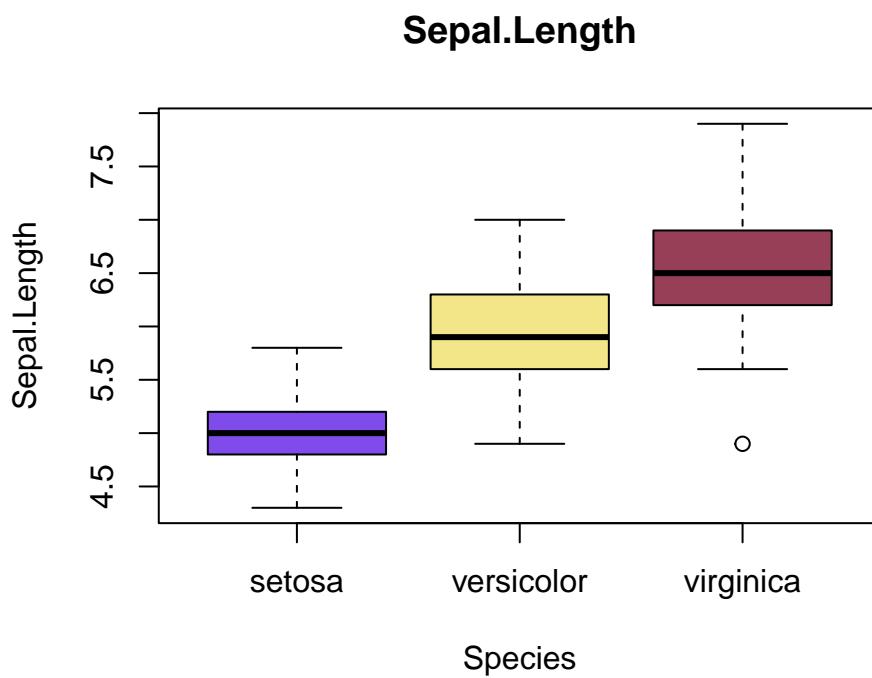
Tabela 1: Opis cech zawartych w danych iris

W tabeli 1 umieszczam opis cech zawartych w tych danych.

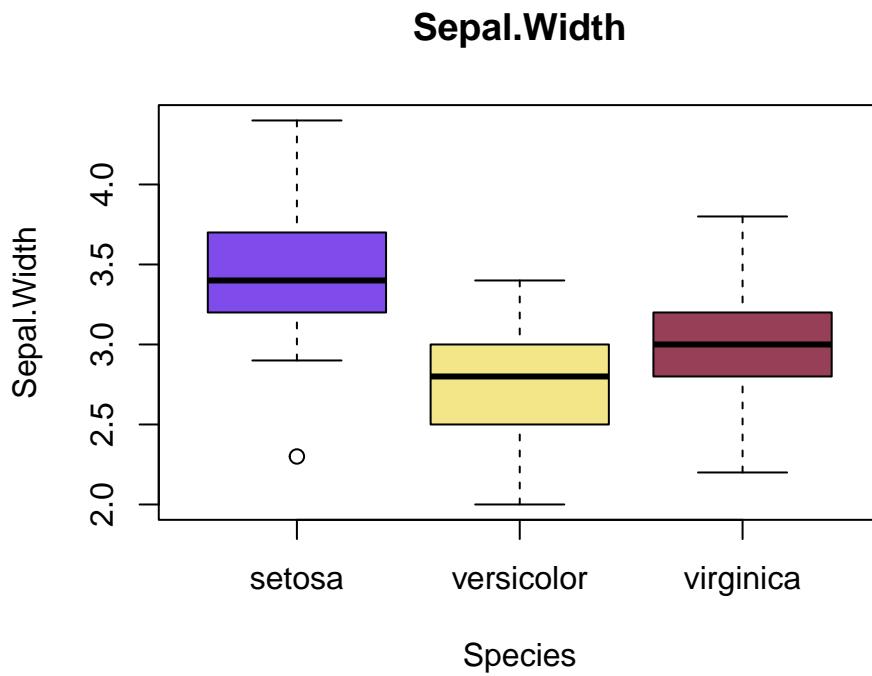
1.3 Wybór cech do dyskretyzacji

W tej części przeanalizuję poszczególne cechy, aby stwierdzić, na postawie której z nich da się najlepiej oraz najgorzej rozróżnić gatunki irysów. Następnie dla tych dwóch cech (najlepszej i najgorszej) dokonam dalszych analiz.

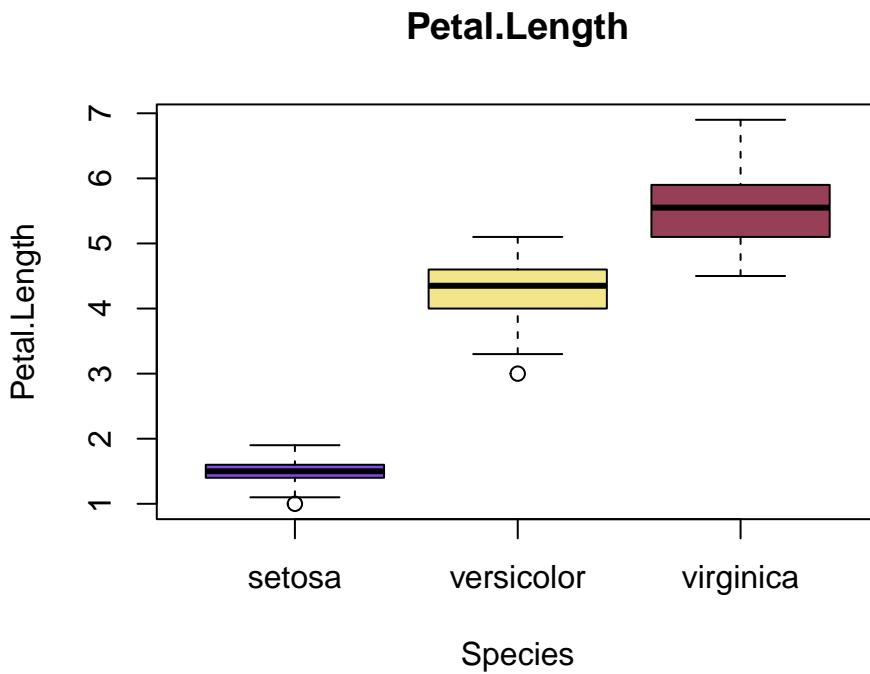
Analizy dokonam, tworząc wykresy pudełkowe dla poszczególnych zmiennych, z rozróżnieniem na gatunki irysów.



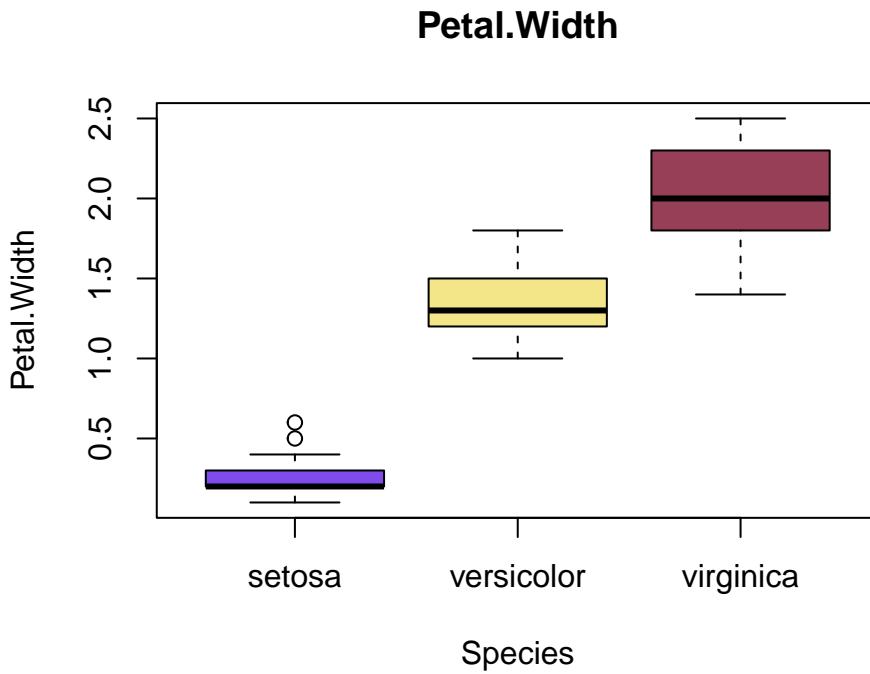
Rysunek 1: Wykres pułapkowy dla zmiennej Sepal.Length



Rysunek 2: Wykres pułapkowy dla zmiennej Sepal.Width



Rysunek 3: Wykres pudełkowy dla zmiennej Petal.Length

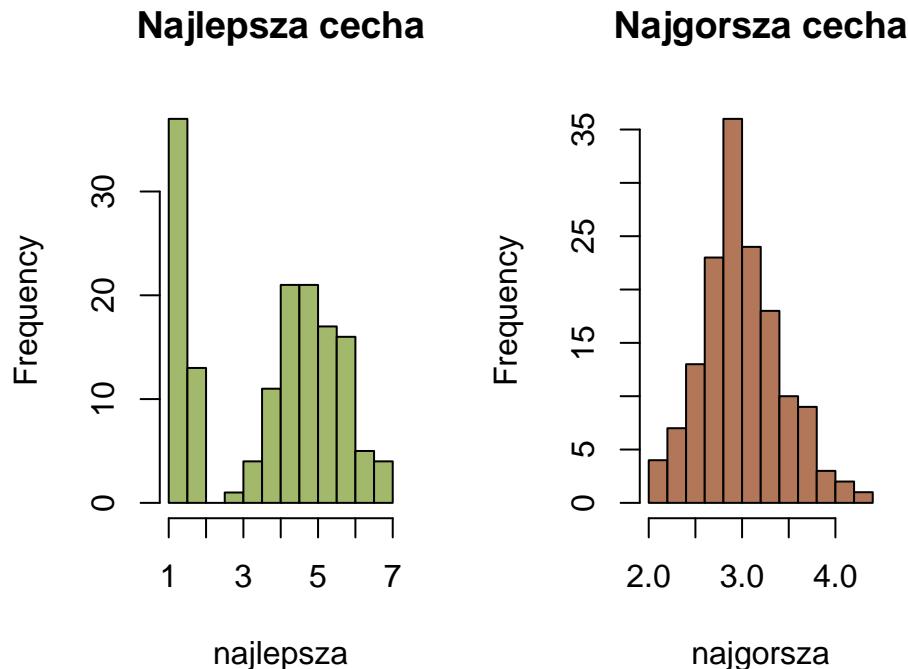


Rysunek 4: Wykres pudełkowy dla zmiennej Petal.Width

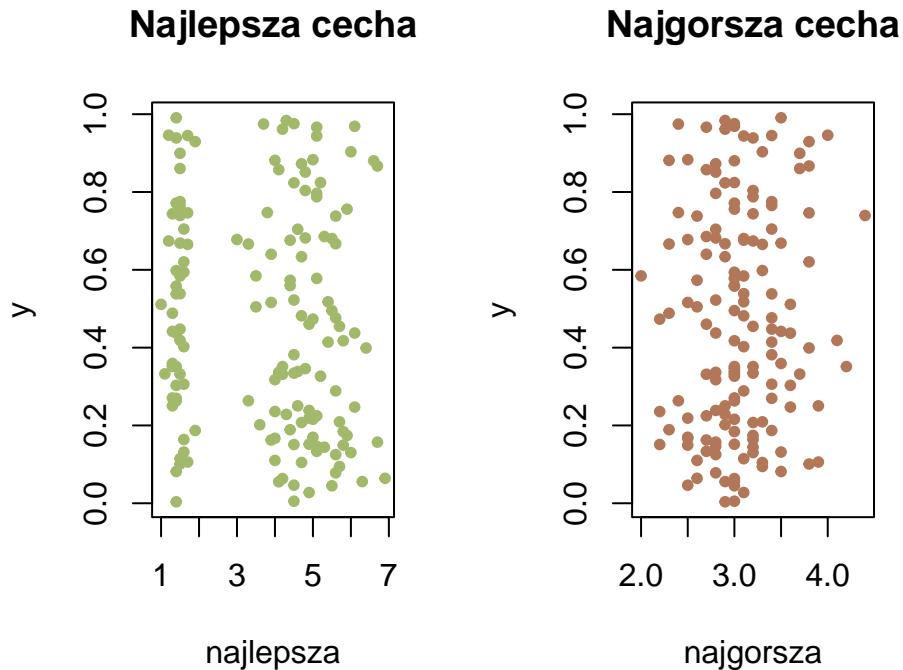
Na podstawie rysunków 1, 2, 3 oraz 4 stwierdziłem, że cechą o najlepszej zdolności dyskryminacyjnej jest Petal.Length, a cechą o najgorszej – Sepal.Width. Zatem do dyskretyzacji utworzę podzbiory tylko z tymi cechami.

```
# Utworzenie odpowiednich podzbiorów  
najlepsza <- iris[, "Petal.Length"]  
najgorsza <- iris[, "Sepal.Width"]
```

Przed zastosowaniem dyskretyzacji przedstawię jeszcze dane wyjściowe na histogramach oraz wykresach rozrzutu.



Rysunek 5: Histogramy dla wybranych cech



Rysunek 6: Wykresy rozrzutu dla wybranych cech

1.4 Dyskretyzacja z wykorzystaniem różnych metod

1.4.1 Metoda oparta na równych częstościach

Metoda ta jest oparta na staraniu się, aby do każdego przedziału trafiła taka sama liczba obiektów.

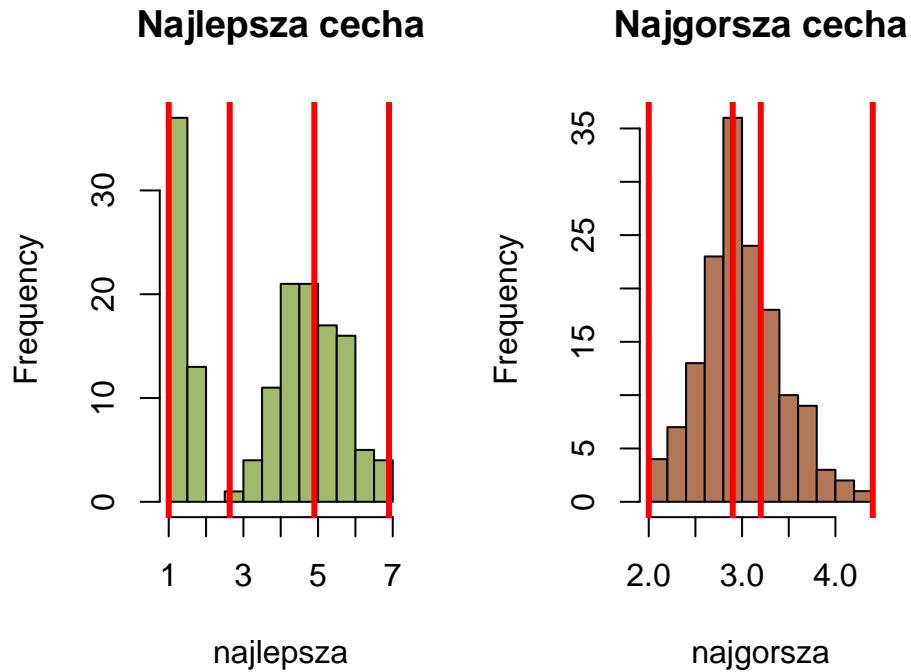
```
# Wykonanie dyskretyzacji
najlepsza.frequency <- discretize(najlepsza, breaks=3)
najgorsza.frequency <- discretize(najgorsza, breaks=3)
```

```
# Wyświetlenie tabel dla obu cech
table(najlepsza.frequency)
```

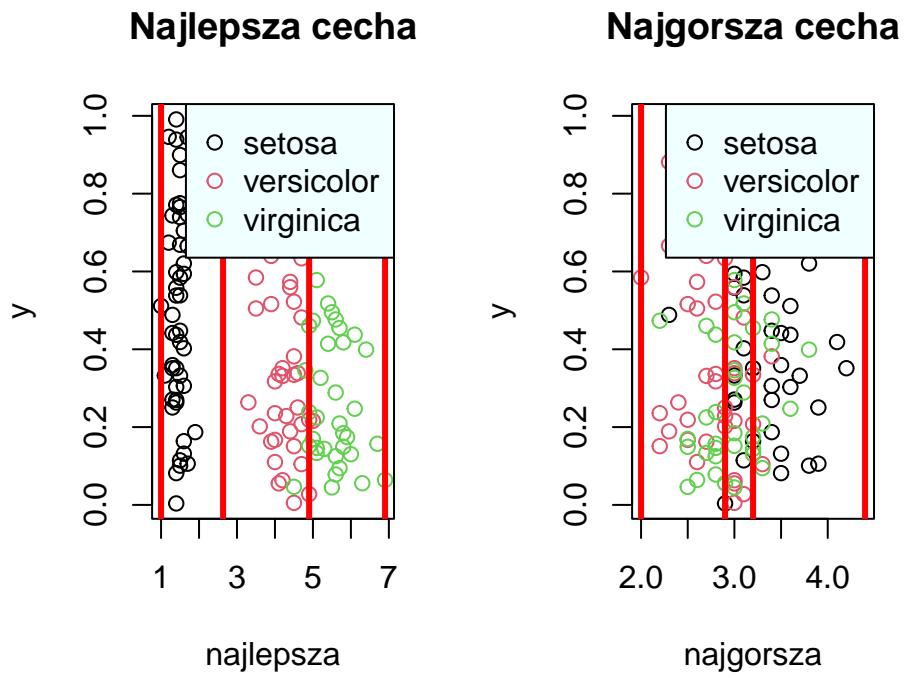
```
## najlepsza.frequency
## [1,2.63) [2.63,4.9) [4.9,6.9]
##      50        49        51
table(najgorsza.frequency)
```

```
## najgorsza.frequency
## [2,2.9) [2.9,3.2) [3.2,4.4]
##      47        47        56
```

Wyniki tej dyskretyzacji przedstawię także na wykresach – histogramach oraz wykresach rozrzutu.



Rysunek 7: Histogramy dla dyskretyzacji opartej na równych częstościach dla najlepszej i najgorszej cechy



Rysunek 8: Wykresy rozrzutu dla dyskretyzacji opartej na równych częstościach dla najlepszej i najgorszej cechy

Na rysunku 8 łatwo zauważać, że dyskretyzacja dla najlepszej cechy dała istotnie lepsze rezultaty niż dla najgorszej.

Wyznaczę jeszcze tabele dwudzielcze i współczynniki zgodności dla obu cech, aby ostatecznie ocenić skuteczność dyskretyzacji.

```
# przypisanie tabel (będą użyte później)
tab.najlepsza.frequency <- table(najlepsza.frequency, iris$Species)
tab.najgorsza.frequency <- table(najgorsza.frequency, iris$Species)
```

Wyświetlenie tabel

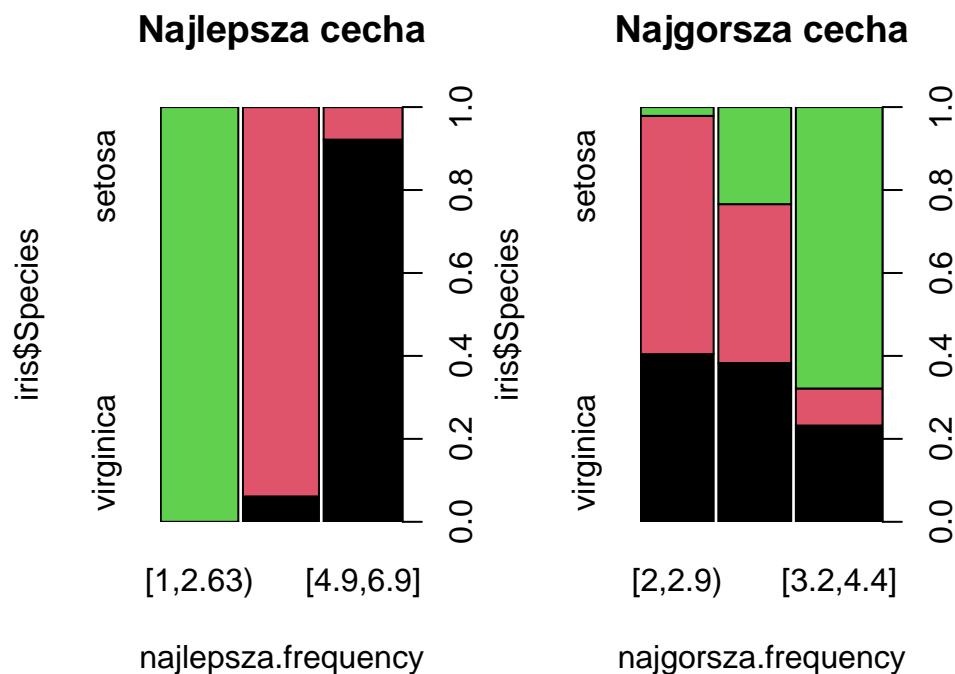
```
tab.najlepsza.frequency
```

```
##
## najlepsza.frequency setosa versicolor virginica
## [1,2.63)      50      0      0
## [2.63,4.9)     0      46      3
## [4.9,6.9]      0      4      47
```

```
tab.najgorsza.frequency
```

```
##
## najgorsza.frequency setosa versicolor virginica
## [2,2.9)      1      27      19
## [2.9,3.2)    11      18      18
## [3.2,4.4]    38      5      13
```

Wyniki z tabel przedstawię również w formie graficznej na wykresach.



Rysunek 9: Wykresy zgodności dla dyskretyzacji opartej na równych częstościach dla najlepszej i najgorszej cechy

Rysunek 9 potwierdza to, co pokazują tabele, że najlepsza cecha daje całkiem zadowalające efekty, natomiast najgorsza – niezbyt.

```
matchClasses(tab.najlepsza.frequency)

## Cases in matched pairs: 95.33 %

##      [1,2.63)  [2.63,4.9)  [4.9,6.9]
##      "setosa" "versicolor" "virginica"

matchClasses(tab.najgorsza.frequency)

## Cases in matched pairs: 55.33 %

##      [2,2.9)  [2.9,3.2)  [3.2,4.4]
##      "versicolor" "versicolor" "setosa"
```

Współczynniki zgodności potwierdzają to wszystko, ponieważ dla najlepszej cechy współczynnik zgodności wynosi ponad 95%.

1.4.2 Metoda oparta na równych szerokościach

Druga z metod dyskretyzacji polega na podzieleniu wartości na przedziały równej szerokości.

```
# Wykonanie dyskretyzacji
najlepsza.interval <- discretize(najlepsza, method="interval", breaks=3)
najgorsza.interval <- discretize(najgorsza, method="interval", breaks=3)

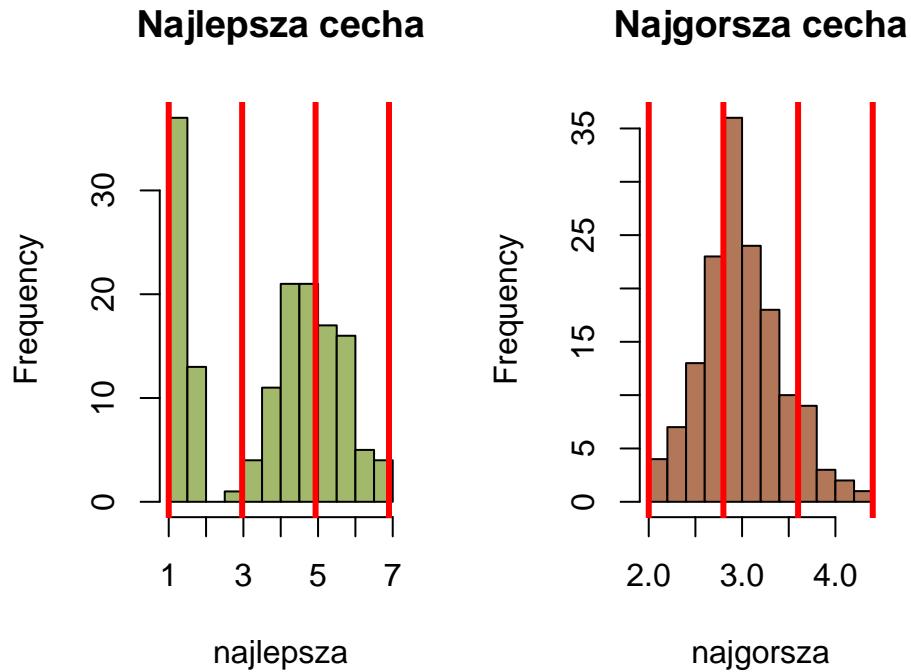
# Wyświetlenie tabel dla obu cech
table(najlepsza.interval)

## najlepsza.interval
##      [1,2.97)  [2.97,4.93)  [4.93,6.9]
##      50          54          46

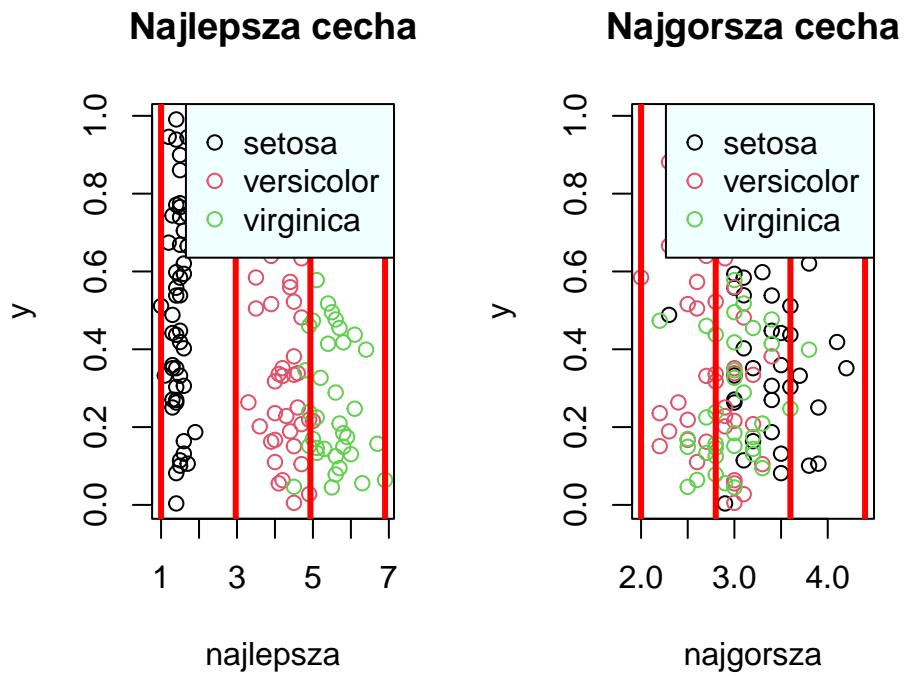
table(najgorsza.interval)

## najgorsza.interval
##      [2,2.8)  [2.8,3.6)  [3.6,4.4]
##      47          88          15
```

Wyniki tej dyskretyzacji przedstawię także na wykresach – histogramach oraz wykresach rozrzutu.



Rysunek 10: Histogramy dla dyskretyzacji opartej na równych szerokościach dla najlepszej i najgorszej cechy



Rysunek 11: Wykresy rozrzutu dla dyskretyzacji opartej na równych szerokościach dla najlepszej i najgorszej cechy

Jak widać na rysunku 11, dla najlepszej cechy dyskretyzacja te prezentuje się dość podobnie do tej względem równych częstości, natomiast znaczna różnica występuje dla najgorszej cechy.

Wyznaczę jeszcze tabele dwudzielcze i współczynniki zgodności dla obu cech, aby ostatecznie ocenić skuteczność dyskretyzacji.

```
# przypisanie tabel (będą użyte później)
tab.najlepsza.interval <- table(najlepsza.interval, iris$Species)
tab.najgorsza.interval <- table(najgorsza.interval, iris$Species)
```

Wyświetlenie tabel

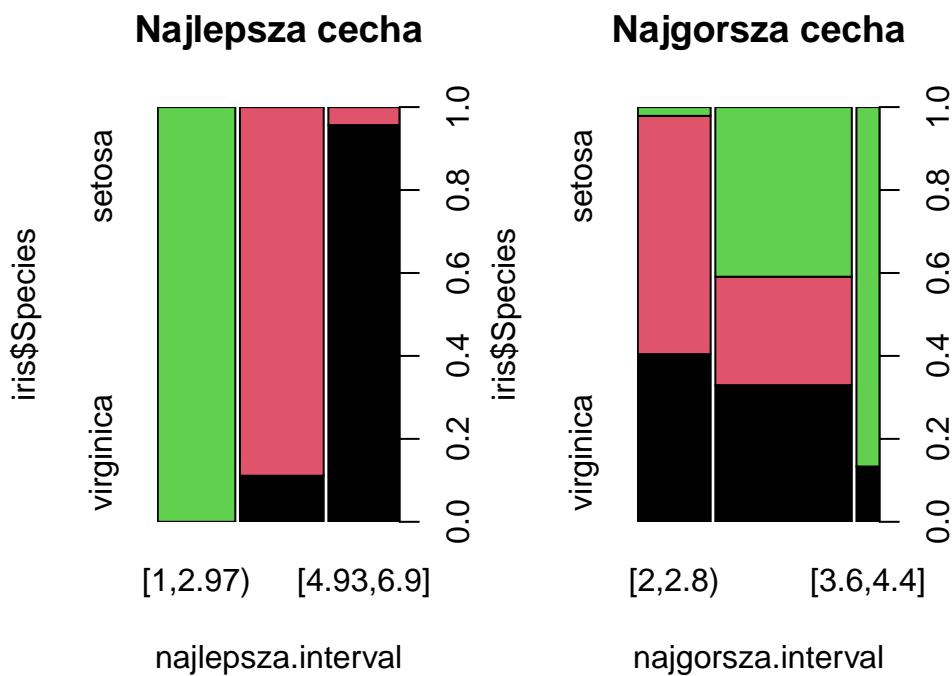
```
tab.najlepsza.interval
```

```
##
## najlepsza.interval setosa versicolor virginica
## [1,2.97)      50      0      0
## [2.97,4.93)    0      48      6
## [4.93,6.9]     0      2     44
```

```
tab.najgorsza.interval
```

```
##
## najgorsza.interval setosa versicolor virginica
## [2,2.8)       1      27      19
## [2.8,3.6)     36      23      29
## [3.6,4.4]     13      0      2
```

Wyniki z tabel przedstawię również w formie graficznej na wykresach.



Rysunek 12: Wykresy zgodności dla dyskretyzacji opartej na równych szerokościach dla najlepszej i najgorszej cechy

Jak wcześniej, rysunek 12 potwierdza to, co pokazują tabele, że najlepsza cecha daje całkiem zadowalające efekty, natomiast najgorsza – niezbyt.

```
matchClasses(tab.najlepsza.interval)

## Cases in matched pairs: 94.67 %

##      [1,2.97)  [2.97,4.93)  [4.93,6.9]
##      "setosa" "versicolor" "virginica"

matchClasses(tab.najgorsza.interval)

## Cases in matched pairs: 50.67 %

##      [2,2.8)  [2.8,3.6)  [3.6,4.4]
## "versicolor" "setosa"   "setosa"
```

1.4.3 Metoda oparta na algorytmie grupowania

Trzecią metodą dyskretyzacji jest dyskretyzacja przy zastosowaniu algorytmu grupowania (klasteryzacji) k-means.

```
# Wykonanie dyskretyzacji
najlepsza.cluster <- discretize(najlepsza, method="cluster", breaks=3)
najgorsza.cluster <- discretize(najgorsza, method="cluster", breaks=3)

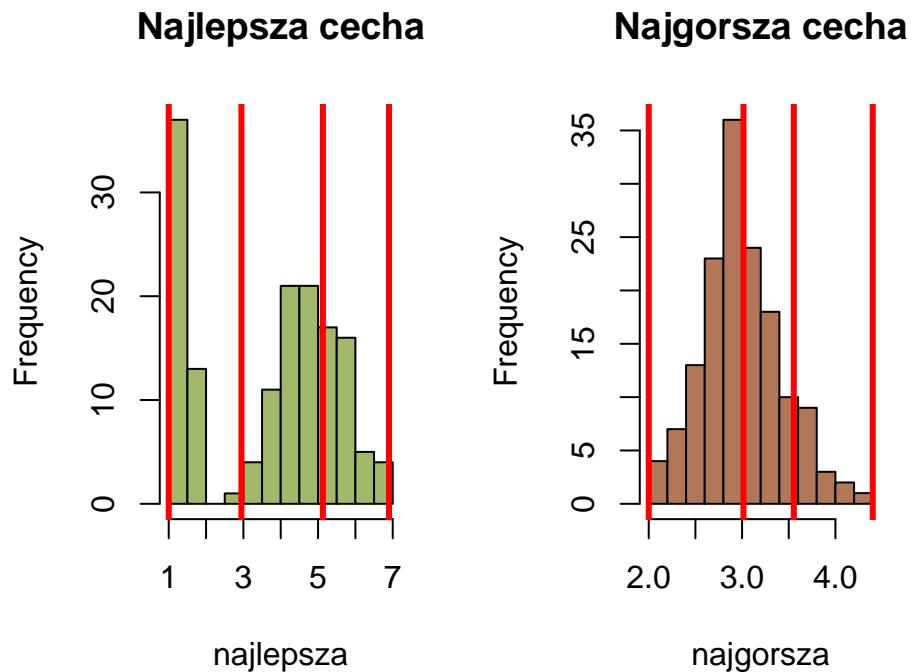
# Wyświetlenie tabel dla obu cech
table(najlepsza.cluster)

## najlepsza.cluster
##      [1,2.95)  [2.95,5.13)  [5.13,6.9]
##      50          66          34

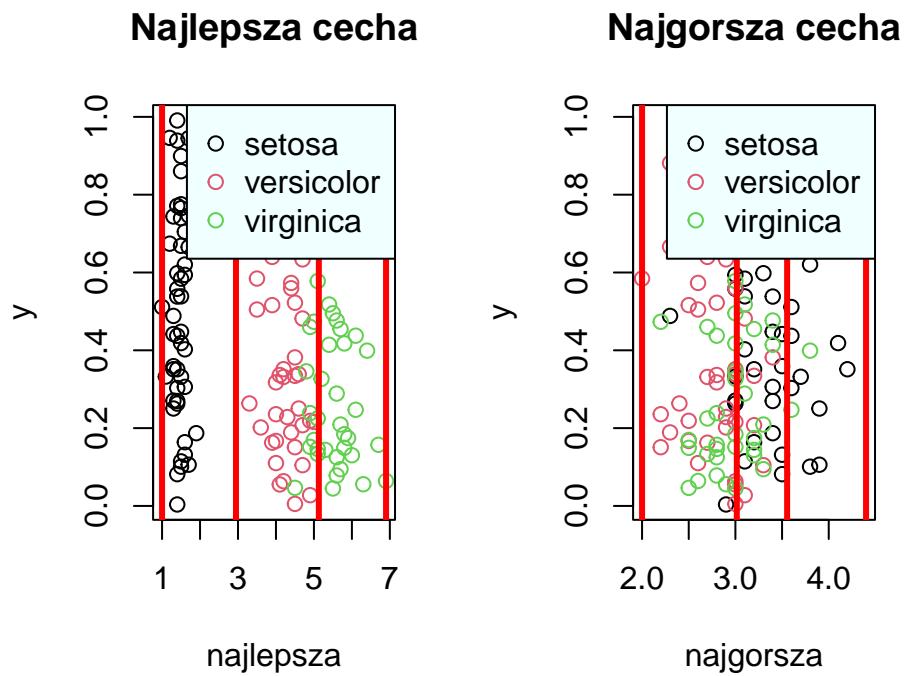
table(najgorsza.cluster)

## najgorsza.cluster
##      [2,3.02)  [3.02,3.55)  [3.55,4.4]
##      83          48          19
```

Wyniki tej dyskretyzacji przedstawię także na wykresach – histogramach oraz wykresach rozrzutu.



Rysunek 13: Histogramy dla dyskretyzacji k-means dla najlepszej i najgorszej cechy



Rysunek 14: Wykresy rozrzutu dla dyskretyzacji k-means dla najlepszej i najgorszej cechy

Na rysunku 14 można dostrzec, że dla najlepszej cechy dyskretyzacja dość różni się od poprzednich metod, natomiast dla najgorszej nieco przypomina tę opartą na równych szerokościach.

Wyznaczę jeszcze tabele dwudzielcze i współczynniki zgodności dla obu cech, aby ostatecznie ocenić skuteczność dyskretyzacji.

```
# przypisanie tabel (będą użyte później)
tab.najlepsza.cluster <- table(najlepsza.cluster, iris$Species)
tab.najgorsza.cluster <- table(najgorsza.cluster, iris$Species)
```

Wyświetlenie tabel

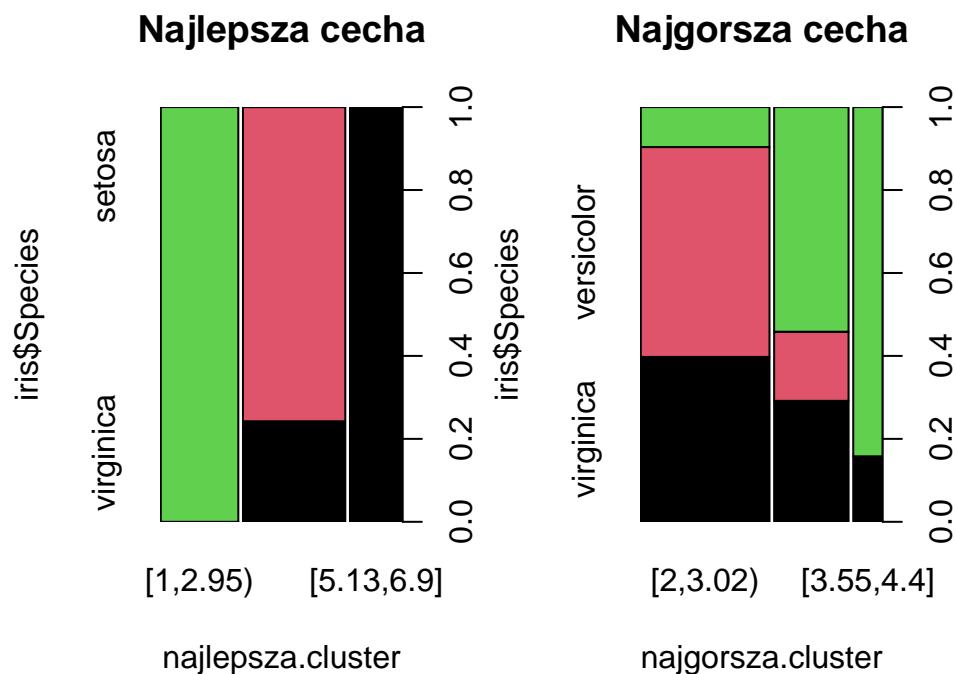
```
tab.najlepsza.cluster
```

```
##
## najlepsza.cluster setosa versicolor virginica
## [1,2.95)      50      0      0
## [2.95,5.13)    0      50     16
## [5.13,6.9]     0      0     34
```

```
tab.najgorsza.cluster
```

```
##
## najgorsza.cluster setosa versicolor virginica
## [2,3.02)      8      42     33
## [3.02,3.55)   26      8     14
## [3.55,4.4]    16      0      3
```

Wyniki z tabel przedstawię również w formie graficznej na wykresach.



Rysunek 15: Wykresy zgodności dla dyskretyzacji k-means dla najlepszej i najgorszej cechy

Rysunek 15 oraz wcześniejsze tabele pokazują, że ta metoda daje bardzo dobre wyniki dla dwóch gatunków irysów dla najlepszej cechy, źle natomiast rozdziela trzeci z nich, za to dyskretyzacja dla najgorszej cechy nie daje zbyt zadowalających efektów.

Ostatnią rzeczą do wyznaczenia są współczynniki zgodności.

```
matchClasses(tab.najlepsza.cluster)
```

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

```
matchClasses(tab.najgorsza.cluster)
```

```
## Cases in matched pairs: 56 %  
##      [2,3.02)  [3.02,3.55)  [3.55,4.4]  
##      "versicolor" "setosa" "setosa"
```

1.4.4 Metoda z przedziałami zadanymi przez użytkownika

Ostatnia z metod polega na ręcznym zadaniu przedziałów do dyskretyzacji. Dla najlepszej metody zadam przedziały od minus nieskończoności do 2, od 2 do 5 i od 5 do nieskończoności, natomiast dla najgorszej – od minus nieskończoności do 3, od 3 do 3,3 i od 3,3 do nieskończoności.

```
# Wykonanie dyskretyzacji
```

```
najlepsza.fixed <- discretize(najlepsza, method="fixed", breaks = c(-Inf, 2, 5, Inf), 1  
najgorsza.fixed <- discretize(najgorsza, method="fixed", breaks = c(-Inf, 3, 3.3, Inf), 1
```

```
# Wyświetlenie tabel dla obu cech
```

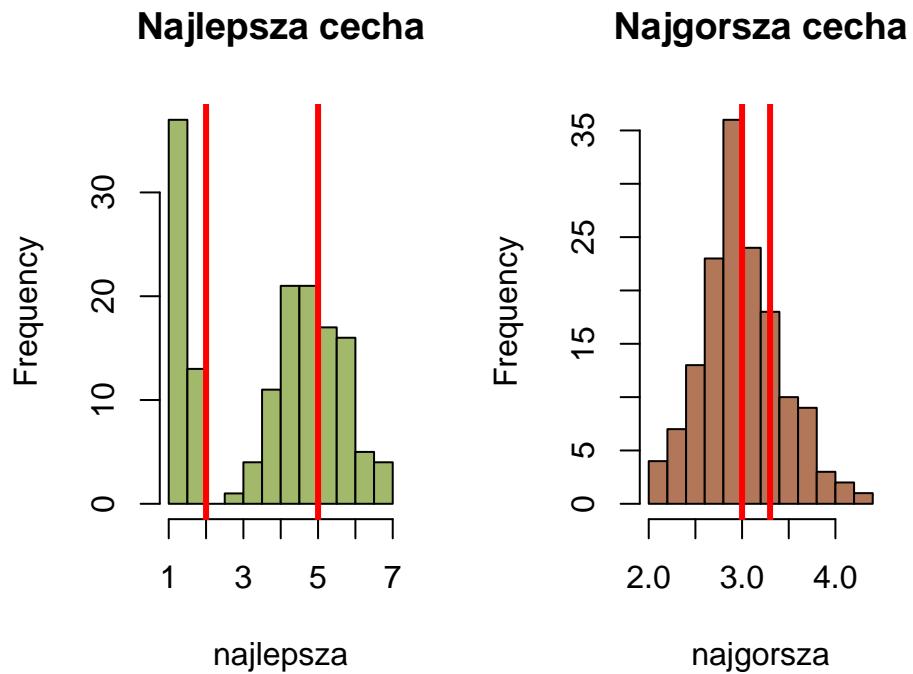
```
table(najlepsza.fixed)
```

```
## najlepsza.fixed  
##   small medium  large  
##     50      54      46
```

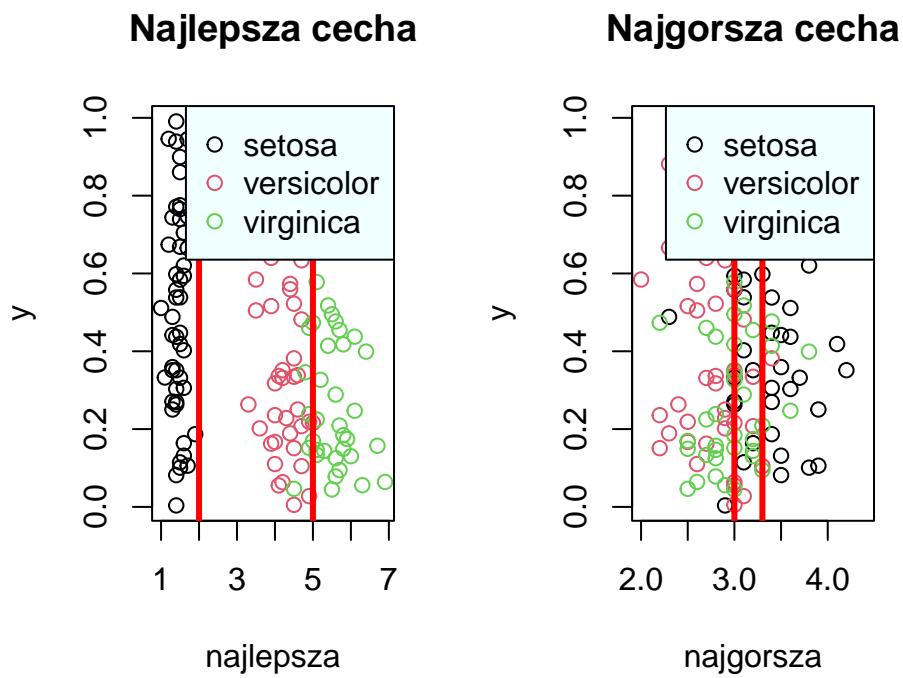
```
table(najgorsza.fixed)
```

```
## najgorsza.fixed  
##   small medium  large  
##     57      50      43
```

Wyniki tej dyskretyzacji przedstawię także na wykresach – histogramach oraz wykresach rozrzutu.



Rysunek 16: Histogramy dla dyskretyzacji według własnych przedziałów dla najlepszej i najgorszej cechy



Rysunek 17: Wykresy rozrzutu dla dyskretyzacji według własnych przedziałów dla najlepszej i najgorszej cechy

Jak widać na rysunku 17, dla dobranych przedziałów dyskretyzacja ta przypomina nieco tę opartą na równych częstościach.

Wyznaczę jeszcze tabele dwudzielcze i współczynniki zgodności dla obu cech, aby ostatecznie ocenić skuteczność dyskretyzacji.

```
# przypisanie tabel (będą użyte później)
tab.najlepsza.fixed <- table(najlepsza.fixed, iris$Species)
tab.najgorsza.fixed <- table(najgorsza.fixed, iris$Species)
```

Wyświetlenie tabel

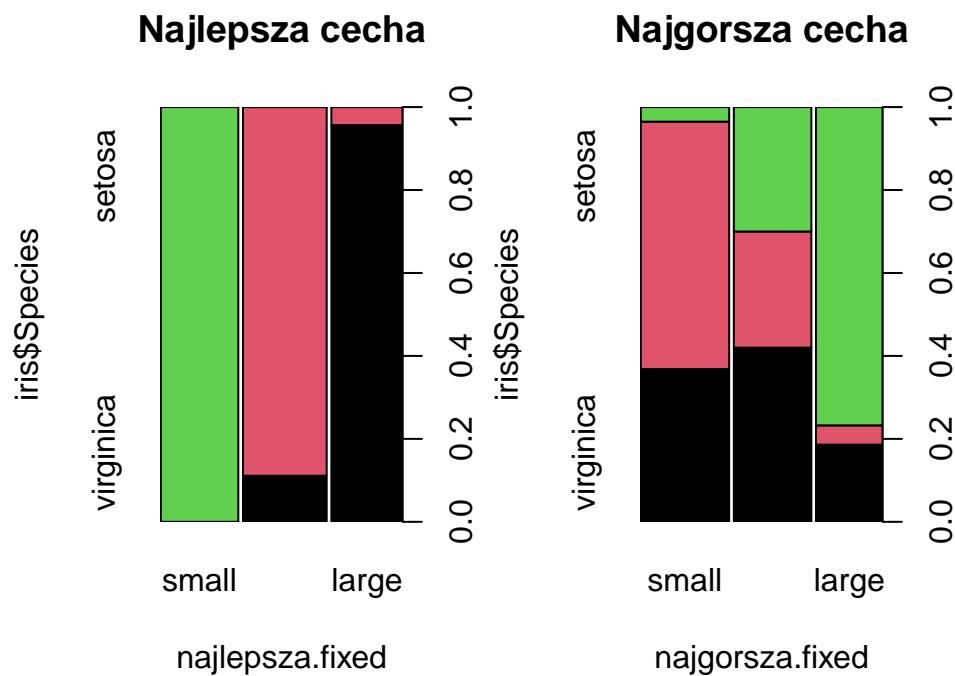
```
tab.najlepsza.fixed
```

```
##
## najlepsza.fixed setosa versicolor virginica
##      small      50         0         0
##     medium       0        48         6
##      large       0         2        44
```

```
tab.najgorsza.fixed
```

```
##
## najgorsza.fixed setosa versicolor virginica
##      small      2        34        21
##     medium     15        14        21
##      large     33         2         8
```

Wyniki z tabel przedstawię również w formie graficznej na wykresach.



Rysunek 18: Wykresy zgodności dla dyskretyzacji według własnych przedziałów dla najlepszej i najgorszej cechy

Rysunek 18 potwierdza to, co pokazują tabele, że najlepsza cecha daje całkiem zadowalające efekty, natomiast najgorsza – niezbyt, nawet pomimo ręcznego dobierania przedziałów.

```
matchClasses(tab.najlepsza.fixed)

## Cases in matched pairs: 94.67 %

##      small      medium      large
## "setosa" "versicolor" "virginica"

matchClasses(tab.najgorsza.fixed)

## Cases in matched pairs: 58.67 %

##      small      medium      large
## "versicolor" "virginica" "setosa"
```

1.5 Wnioski

Porównując wszystkie metody dyskretyzacji, można stwierdzić, że dla najlepszej cechy najlepsze okazały się algorytmy oparte na równych częstotliwościach i k-means, natomiast drugie w kolejności okazały się algorytmy oparte na równych szerokościach i o przedziałach zadanych przez użytkownika (z tymi konkretnymi przedziałami). Dla wszystkich tych algorytmów współczynniki zgodności wynoszą ponad 90% i dla wszystkich metod dyskretyzacji są bardzo podobne.

Natomiast dla najgorszej cechy współczynniki zgodności są bardziej zróżnicowane, lecz wszystkie mieszczą się w przedziale 50–60%. Najlepszy okazał się algorytm z przedziałami zadanymi przez użytkownika, następnie po kolei k-means i równe częstotliwości, natomiast najgorszy z algorytmów, zadający przedziały o równej szerokości, okazał się najgorszy i istotnie gorszy od algorytmu równych częstotliwości (o prawie 5 punktów procentowych).

Oczywiście wyniki dla najlepszej cechy różnią się istotnie od wyników dla najgorszej cechy – wszystkie przetestowane metody dyskretyzacji pokazują, że zdecydowanie bardziej opłaca się je stosować dla cech o dobrej zdolności dyskryminacyjnej, gdyż ich dyskretyzacja sensownie dzieli gatunki. Natomiast dyskretyzacja dla najgorszej cechy ma skuteczność w okolicy 50%, dlatego wykorzystanie tej cechy stanowi niezbyt dobry pomysł.

Na koniec zadania odłączam także dla porządku dane iris.

```
# Odłączenie danych iris (dla porządku)
detach(iris)
```

2 Analiza składowych głównych (PCA – Principal Component Analysis)

W tej części niniejszego raportu przedstawię analizę składową głównych.

2.1 Wykorzystane metody

Metody, które wykorzystam, to:

- wyznaczenie składowych głównych i ich analiza (PCA),
- badanie wariancji i wyjaśnionej wariancji, aby wyjaśnić jak najwięcej zmienności danych,
- narzędzia graficzne (wykresy) do prezentacji i analizy wyników.

2.2 Opis danych

Dane analizowane w tej części to City Quality Of Life Dataset, czyli dane opisujące jakość życia w poszczególnych miastach. Wszystkie cechy ilościowe przyjmują wartości z przedziału 0–10, gdzie większa wartość oznacza lepszy wynik.

Nazwa	Typ	Opis
X	Liczbowa	Identyfikator miasta
UA_Name	Jakościowa	Nazwa miasta
UA_Country	Jakościowa	Kraj
UA_Continent	Jakościowa	Kontynent
Housing	Liczbowa	Warunki mieszkaniowe
Cost.of.Living	Liczbowa	Koszt utrzymania
Startups	Liczbowa	Startupy
Venture.Capital	Liczbowa	Inwestycje w młode przedsiębiorstwa
Travel.Connectivity	Liczbowa	Połączenia komunikacyjne
Commute	Liczbowa	Dojazd
Business.Freedom	Liczbowa	Wolność biznesowa
Safety	Liczbowa	Bezpieczeństwo
Healthcare	Liczbowa	Opieka zdrowotna
Education	Liczbowa	Edukacja
Environmental.Quality	Liczbowa	Jakość środowiska
Economy	Liczbowa	Ekonomia
Taxation	Liczbowa	Opodatkowanie
Internet.Access	Liczbowa	Dostęp do Internetu
Leisure...Culture	Liczbowa	Kultura czasu wolnego
Tolerance	Liczbowa	Tolerancja
Outdoors	Liczbowa	Plener

Tabela 2: Opis cech zawartych w danych City Quality Of Life Dataset

W tabeli 2 znajduje się wykaz i opis poszczególnych cech.

2.3 Przygotowanie danych

Pierwszym krokiem jest przygotowanie danych. Na samym początku wczytam dane oraz wybór podzbiór zawierający wyłącznie dane liczbowe (poza identyfikatorem, oczywiście).

Utworzymy także zbiór, na którym wykorzystano standaryzację, aby w późniejszej części sprawdzić, czy standaryzacja ma istotny wpływ na wyniki analizy.

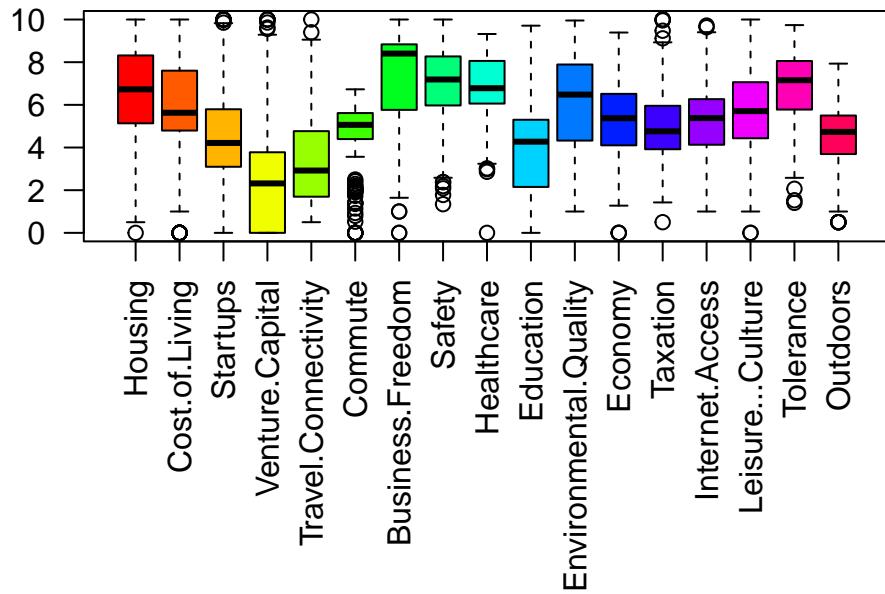
```
# Wczytanie danych
city <- read.csv(file="uaScoresDataFrame.csv", stringsAsFactors=TRUE)

# Wybranie podzbioru z wartościami numerycznymi
city.main <- city[,5:21]

# Standaryzacja
city.main.scaled <- scale(city.main, center=TRUE, scale=TRUE)
```

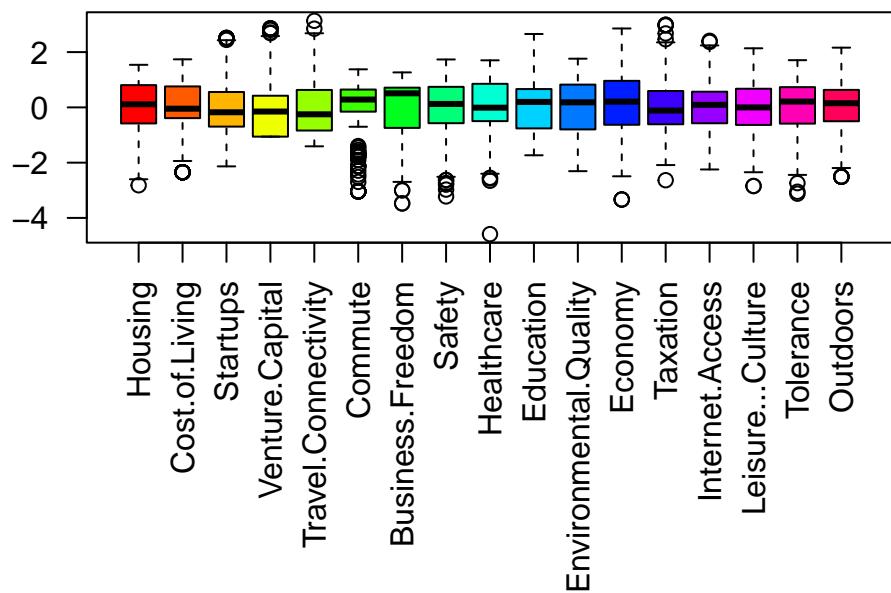
Wszystkie dane wczytały się poprawnie, więc nie ma konieczności zamiany typów danych. Następnie porównam zmienność poszczególnych cech przed standaryzacją i po niej.

Wykresy pudełkowe dla poszczególnych zmiennych liczbowych



Rysunek 19: Wykresy pudełkowe dla cech liczbowych city

Wykresy pudełkowe dla poszczególnych zmiennych liczbowych po standaryzacji



Rysunek 20: Wykresy pudełkowe dla cech liczbowych city po standaryzacji

2.4 Wyznaczenie składowych głównych

Następnym krokiem jest wyznaczenie składowych głównych.

```
# Wyznaczenie składowych głównych dla wersji zwykłej i ustandaryzowanej
city.main.pca <- prcomp(city.main, retx=T, center=T)
city.main.scaled.pca <- prcomp(city.main.scaled, retx=T, center=T)
```

Teraz dla obu wersji przeanalizuję wektory ładunków pierwszych trzech głównych składowych.

```
# Wyświetlenie wektorów ładunków dla danych niestandaryzowanych
city.main.pca$rotation[,1:3]
```

	PC1	PC2	PC3
## Housing	0.37242737	0.06107944	0.42548429
## Cost.of.Living	0.32567873	0.33295425	0.41958014
## Startups	-0.21991774	0.47359165	-0.10914131
## Venture.Capital	-0.33424051	0.50010881	-0.15041545
## Travel.Connectivity	-0.20628382	0.16849102	0.31523643
## Commute	-0.05513830	0.04776381	0.40294718
## Business.Freedom	-0.35377810	-0.14761183	0.17572131
## Safety	-0.01454403	-0.14955910	0.25561233
## Healthcare	-0.16662613	-0.14185046	0.29803822
## Education	-0.41727758	0.01770795	0.18591295
## Environmental.Quality	-0.31750403	-0.31097977	0.14525283
## Economy	-0.19562681	-0.01049967	-0.12758032

```

## Taxation          0.02611105 -0.05441808  0.01299327
## Internet.Access -0.23299282 -0.04876310  0.11912313
## Leisure...Culture -0.07803217  0.36230342  0.21996309
## Tolerance         -0.12951735 -0.25876198  0.16051457
## Outdoors          -0.07343196  0.12397834  0.03390511

```

Dla danych nieustandardyzowanych można zauważać następujące własności tych składowych:

- PC1 największą uwagę zwraca na warunki mieszkaniowe i koszty utrzymania, co się oczywiście naturalnie ze sobą łączy,
- PC2 zwraca uwagę na zmienne Startups i Venture.Capital, więc mówi przede wszystkim o warunkach dla startujących przedsiębiorstw,
- PC3 poza warunkami mieszkaniowymi i kosztami utrzymania sporą wagę kładzie na połączenia transportowe i dojazd do pracy, więc mówi nie tylko o samych warunkach mieszkania, ale też o lokalizacji.

Taką samą analizę wykonam dla danych ustandardyzowanych.

```
# Wyświetlenie wektorów ładunków dla danych ustandardyzowanych
city.main.scaled.pca$rotation[,1:3]
```

	PC1	PC2	PC3
## Housing	0.30782507	0.05335338	-0.313546491
## Cost.of.Living	0.25960906	-0.17578152	-0.330535177
## Startups	-0.18023854	-0.48344148	0.006099991
## Venture.Capital	-0.23659743	-0.42745094	0.014876825
## Travel.Connectivity	-0.20945432	-0.13530669	-0.339775966
## Commute	-0.11420445	0.02593103	-0.505735874
## Business.Freedom	-0.37728094	0.09821960	0.024104574
## Safety	-0.03893545	0.28710395	-0.333009986
## Healthcare	-0.28035896	0.24194822	-0.281024808
## Education	-0.40256205	-0.04907952	-0.073864482
## Environmental.Quality	-0.32622202	0.25253554	0.053571655
## Economy	-0.27317525	-0.07400327	0.308670514
## Taxation	0.02629917	0.10741513	-0.020184933
## Internet.Access	-0.27619221	0.02270564	0.028441569
## Leisure...Culture	-0.07444660	-0.36473241	-0.305054520
## Tolerance	-0.18974955	0.35509112	-0.102725071
## Outdoors	-0.09158656	-0.19338254	-0.148586789

Wnioski, które można wyciągnąć, są następujące:

- PC1, jak wcześniej, zwraca sporą uwagę na warunki mieszkaniowe i koszt utrzymania,
- PC2 natomiast największą wagę przypisuje tolerancji, bezpieczeństwu, warunkom środowiskowym i opiece zdrowotnej; można powiedzieć, że zwraca uwagę na dobry stan człowieka jako jednostki,
- PC3 największą wagę przypisuje zmiennej Economy – zwraca więc uwagę na aspekty ekonomiczne.

Tak więc standaryzacja ma wpływ na wektory ładunków poszczególnych składowych głównych – sprawia, że poszczególne składowe są bardziej różnorodne.

2.5 Zmienność poszczególnych składowych

Następnym krokiem jest sprawdzenie zmienności odpowiadającej poszczególnym składowym. Na tej podstawie odpowiedziem, ile składowych trzeba, by wyjaśnić odpowiednio 80% i 90% całkowitej zmienności danych.

```
# Wyświetlenie informacji o wyjaśnionej wariancji  
summary(city.main.pca)
```

```
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 4.7395 3.4008 2.7143 2.22244 1.92703 1.76884 1.62423  
## Proportion of Variance 0.3366 0.1733 0.1104 0.07401 0.05565 0.04688 0.03953  
## Cumulative Proportion 0.3366 0.5099 0.6203 0.69433 0.74997 0.79686 0.83639  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 1.50346 1.35002 1.12082 1.08781 1.06721 0.92301 0.89190  
## Proportion of Variance 0.03387 0.02731 0.01882 0.01773 0.01707 0.01277 0.01192  
## Cumulative Proportion 0.87026 0.89757 0.91640 0.93413 0.95119 0.96396 0.97588  
## PC15 PC16 PC17  
## Standard deviation 0.81516 0.71951 0.6538  
## Proportion of Variance 0.00996 0.00776 0.0064  
## Cumulative Proportion 0.98584 0.99360 1.0000
```

```
summary(city.main.scaled.pca)
```

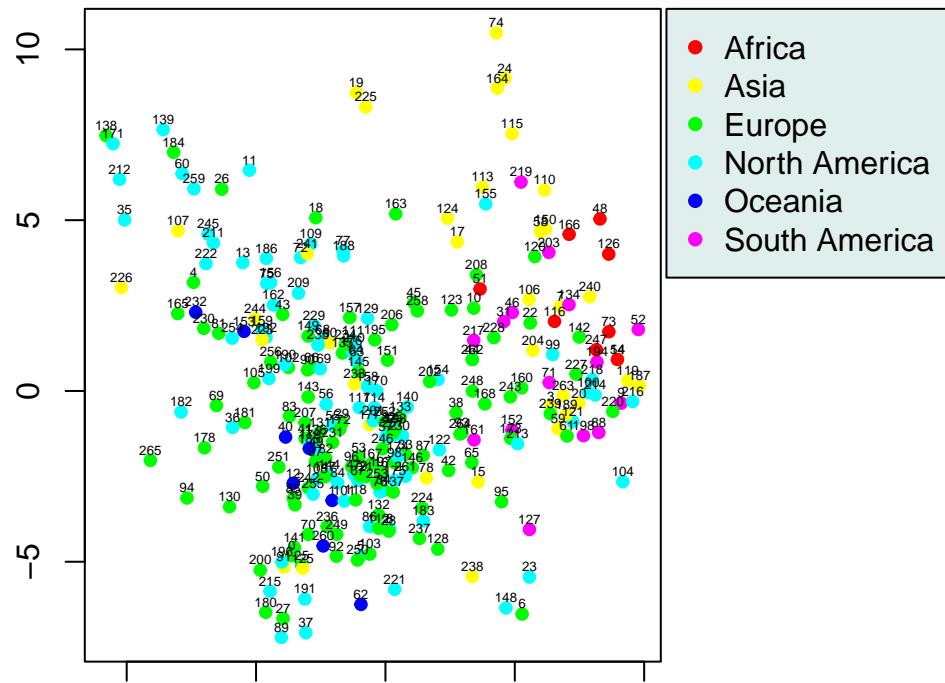
```
## Importance of components:  
## PC1 PC2 PC3 PC4 PC5 PC6 PC7  
## Standard deviation 2.251 1.6055 1.4430 1.14022 1.09451 0.9800 0.83117  
## Proportion of Variance 0.298 0.1516 0.1225 0.07648 0.07047 0.0565 0.04064  
## Cumulative Proportion 0.298 0.4496 0.5721 0.64856 0.71903 0.7755 0.81617  
## PC8 PC9 PC10 PC11 PC12 PC13 PC14  
## Standard deviation 0.81463 0.76387 0.65148 0.56860 0.53928 0.52414 0.43425  
## Proportion of Variance 0.03904 0.03432 0.02497 0.01902 0.01711 0.01616 0.01109  
## Cumulative Proportion 0.85520 0.88953 0.91449 0.93351 0.95062 0.96678 0.97787  
## PC15 PC16 PC17  
## Standard deviation 0.39266 0.35204 0.31319  
## Proportion of Variance 0.00907 0.00729 0.00577  
## Cumulative Proportion 0.98694 0.99423 1.00000
```

W obu przypadkach do wyjaśnienia 80% całkowitej zmienności danych potrzeba 7 składowych głównych, a do wyjaśnienia 90% – 10.

2.6 Wizualizacja danych

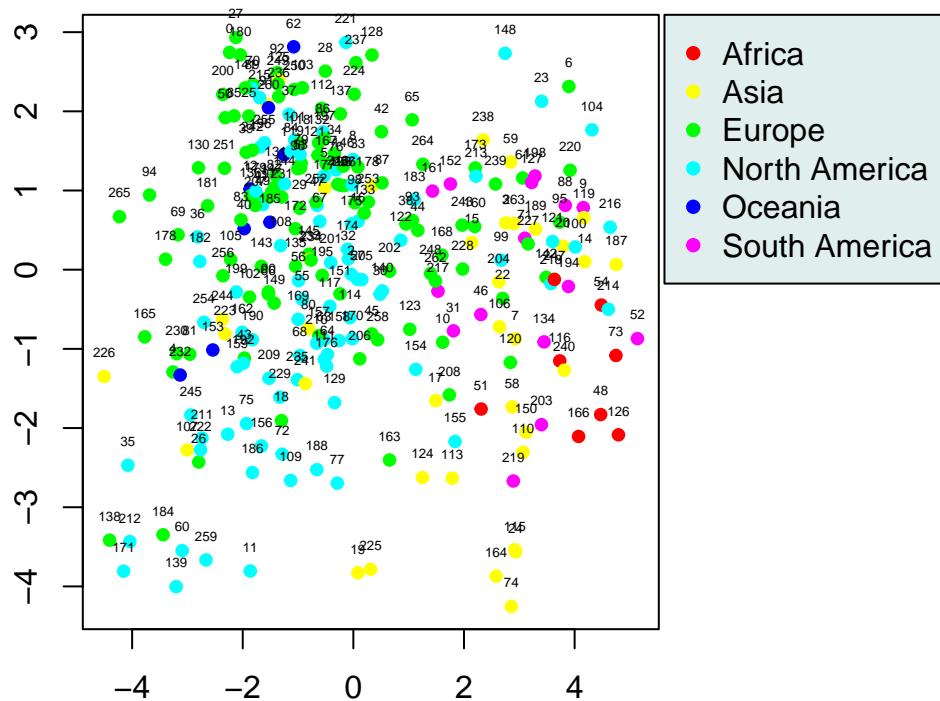
Następnym krokiem jest przedstawienie danych dla obu wersji na wykresach rozrzutu.

Rozrzut – dane nieustandardyzowane



Rysunek 21: Wykres rozrzutu 2D dla analizy składowych głównych (wersja nieustandardyzowana)

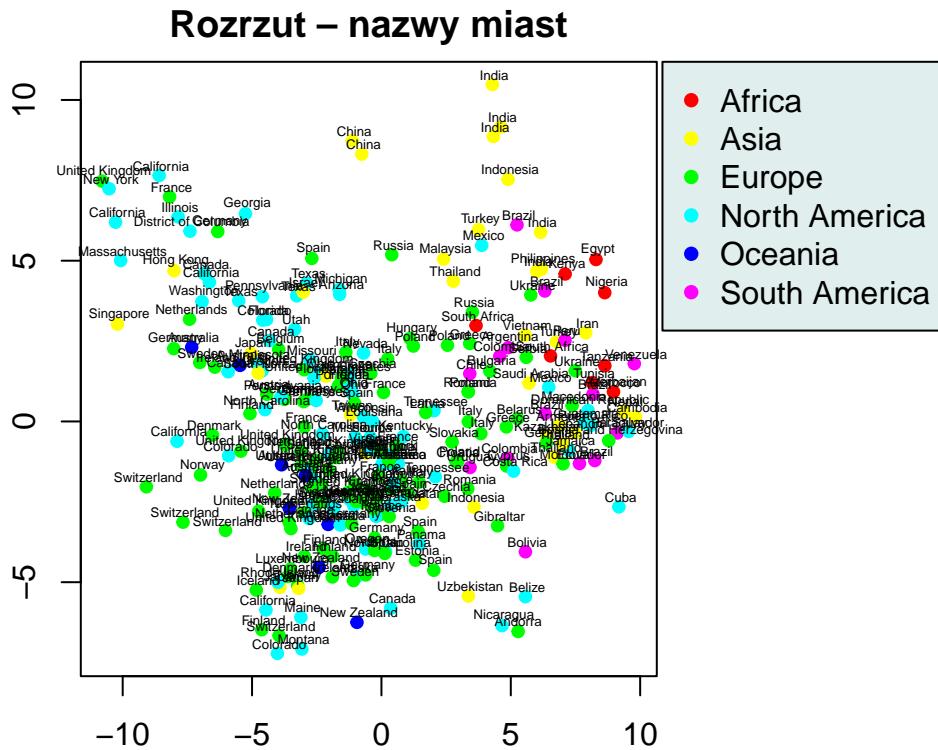
Rozrzut – dane ustandaryzowane



Rysunek 22: Wykres rozrzutu 2D dla analizy składowych głównych (wersja ustandaryzowana)

Łatwo można zauważyć, że dane na rysunkach 21 oraz 22 różnią się przede wszystkim tym, że te ustandaryzowane są odwrócone pionowo względem tych nieustandaryzowanych. Dane odrobinę tworzą skupiska, lecz nie całkowicie – sporo z nich jest rozproszonych. W jednym miejscu, po prawej stronie wykresu, znajdują się miasta Ameryki Południowej, Afryki oraz sporo miast azjatyckich, natomiast drugą główną grupę, po lewej, stanowią Europa, Ameryka Północna i Oceania. Część miast azjatyckich wpada do grupy lewej, natomiast do prawej wpada część miast Europy i Ameryki Północnej.

Aby wyciągnąć ostateczne wnioski, utworzę jeszcze jeden wykres, tym razem przypisujący punktom nazwy państw.



Rysunek 23: Wykres rozrzutu 2D dla analizy składowych głównych (wersja nieustandardyzowana), ale z nazwami państw

Ten wykres ostatecznie potwierdza obserwacje, że skupisko po lewej to głównie rozwinięte kraje Zachodu oraz bardziej rozwinięte kraje azjatyckie, takie jak Japonia. Natomiast skupisko po prawej to głównie kraje mniej rozwinięte, w tym kraje Europy nie należące do Unii Europejskiej.

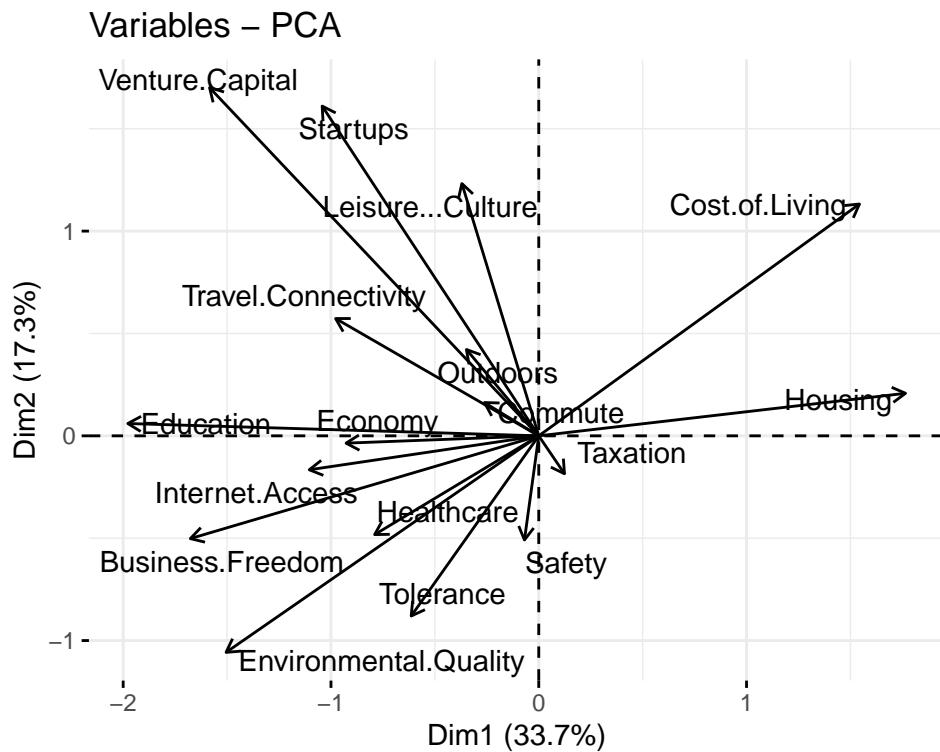
Można zauważyć też trzy obszary obserwacji odbiegających. Pierwszy z nich to lewy górnego róg z miastami z państw takich jak Wielka Brytania (Londyn) czy Francja (Paryż) i ze stanów Nowy Jork i Kalifornia – są to wszystko obszary dobrze rozwinięte. Po lewej stronie, ale nieco niżej, odbiega też część obserwacji ze Szwajcarii.

Uwagę zwraca też drugi obszar przy górnym brzegu wykresu z obserwacjami z Chin i Indii, czyli krajów cechujących się szczególnie wysoką gęstością zaludnienia. Są to przede wszystkim największe miasta tych krajów, takie jak Szanghaj, Pekin czy Delhi.

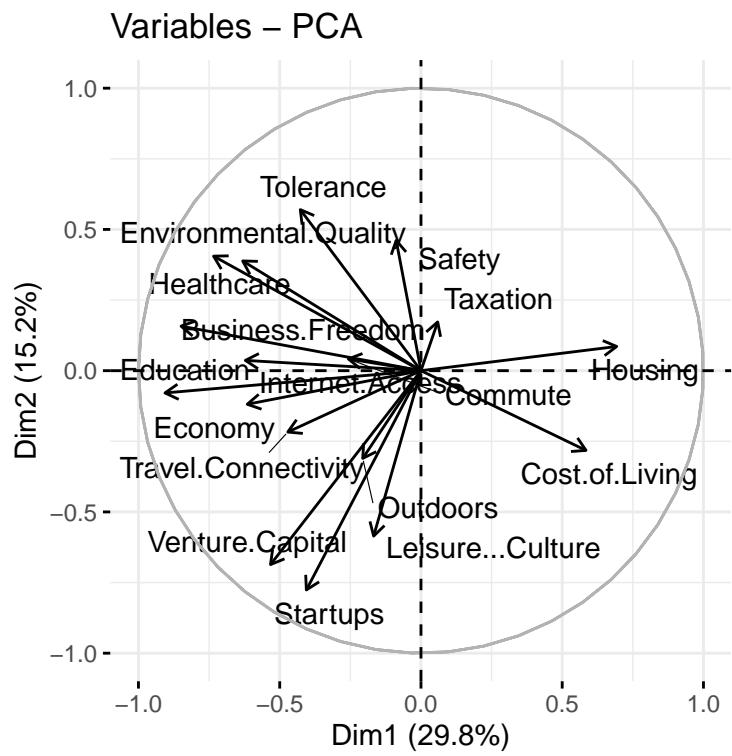
Ostatnim takim obszarem godnym wyróżnienia jest prawy dolny róg wykresu, z miastami z krajów słabo rozwiniętych i/lub komunistycznych, takich jak Kuba czy Boliwia, a także z małych krajów, takich jak Andora.

2.7 Korelacja zmiennych

Ostatnim krokiem przed wyciągnięciem ostatecznych wniosków jest zbadanie korelacji zmiennych. Ze względu na dużą liczbę zmiennych dwuwykres daje zbyt nieczytelne wyniki, zatem wykorzystam koło korelacji.



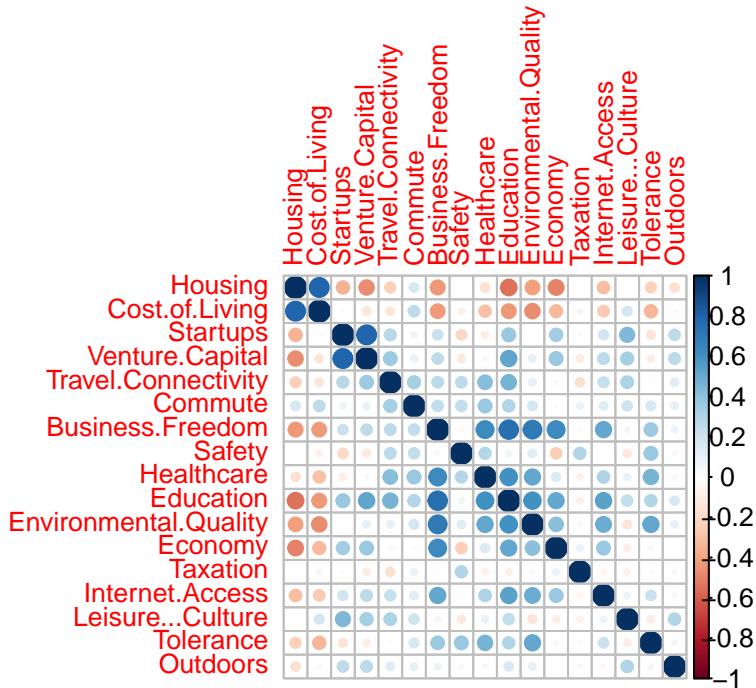
Rysunek 24: Koło korelacji dla danych nieustandardyzowanych



Rysunek 25: Koło korelacji dla danych ustandardyzowanych

Rysunki 24 oraz 25 pokazują dość podobne wyniki, jednak w większości „odwrócone” względem poziomej osi. Na obu widać, że dość blisko siebie są zmienne Housing i Cost.of.Living, natomiast większość pozostałych zmiennych znajduje się blisko siebie nawzajem i daleko od tamtych dwóch. Porównam te wyniki z macierzą korelacji, którą dla większej czytelności również zwizualizuję na wykresie.

Vizualizacja macierzy korelacji zmiennych z danych cit



Rysunek 26: Wizualizacja macierzy korelacji

Rysunek 26 potwierdza ogólny wniosek o tym, że zmienne Housing i Cost.of.Living są „dalej” od reszty zmiennych, a między większością pozostałych jest dodatnia korelacja.

2.8 Wnioski

Na podstawie obserwacji można wyciągnąć kilka wniosków.

1. Zwykle koszty utrzymania idą w parze z warunkami mieszkaniowymi, podobnie startupy idą w parze z venture. Istnieje jeszcze jedna grupa dość powiązanych zmiennych: wolność biznesowa, opieka zdrowotna, edukacja, jakość środowiska, ekonomia i, w nieco mniejszej mierze, dostęp do Internetu. Poza tym odrobinę uwagi zwraca korelacja zmiennych mówiących o tolerancji i jakości środowiska.
2. Obserwacje naturalnie układają się w dwie grupy krajów rozwiniętych oraz rozwijających się.
3. Do otrzymania zadowalającej reprezentacji danych potrzebujemy przynajmniej siedmiu składowych głównych, czyli liczby odpowiadającej prawie połowie wszystkich zmiennych.

4. Standaryzacja danych w tym przypadku na ogół nie miała istotnego wpływu na wyniki (co jest spowodowane tym, że już oryginalne dane były na tej samej skali), czasami nawet nieco utrudniła odczytanie wyników. Jedyną istotną różnicę stanowiła w przypadku interpretacji ładunków składowych głównych, gdzie dane standaryzowane dały większą różnorodność w interpretacji trzech pierwszych składowych głównych.

3 Skalowanie wielowymiarowe (MDS – Multidimensional Scaling)

W ostatniej części raportu omówię skalowanie wielowymiarowe.

3.1 Wykorzystane metody

Metody, które będę stosować w tej części, to:

- macierz odmienności,
- skalowanie wielowymiarowe (MDS),
- diagram Sheparda,
- graficzna prezentacja wyników na wykresach.

3.2 Przygotowanie i opis danych

Dane, na których będę pracować w tej części to `titanic` z pakietu o tej samej nazwie. W trakcie wstępnej obróbki zmieniłem na typ factor zmienne, które wczytały się jako chr, a także te, których wartości są liczbami, ale powinno się je odczytywać jako zmienne jakościowe (tj. `Survived`, gdzie 0 – fałsz, 1 – prawda oraz `Pclass`, gdzie cyfra oznacza numer klasy). Usunęłam także zmienne służące za identyfikatory (`PassengerID`, `Name`, `Ticket`, `Cabin`).

```
# Wczytanie danych
titanic <- titanic_train

# Zamiana typów zmiennych, które się błędnie wczytały
titanic$Sex <- as.factor(titanic$Sex)
titanic$Pclass <- as.factor(titanic$Pclass)
titanic$Embarked <- as.factor(titanic$Embarked)
titanic$Survived <- as.factor(titanic$Survived)

# Usunięcie zbędnych zmiennych (identyfikatorów)
titanic <- titanic[c(2:3, 5:8, 10, 12)]

# Utworzenie podzbioru pomijającego zmienną Survived (ten będę skalować)
titanic.to.scale=titanic[,2:8]
```

Po wstępnej obróbce mogę opisać dokładniej dane.

Nazwa	Typ	Opis
Survived	Jakościowa	Czy dana osoba przetrwała katastrofę
Pclass	Jakościowa	Klasa
Sex	Jakościowa	Płeć
Age	Liczbową	Wiek
SibSp	Liczbową	Liczba rodzeństwa/partnerów na statku
Parch	Liczbową	Liczba rodziców/dzieci na statku
Fare	Liczbową	Cena biletu
Embarked	Jakościowa	Miejsce wejścia na pokład

Tabela 3: Opis cech zawartych w danych titanic

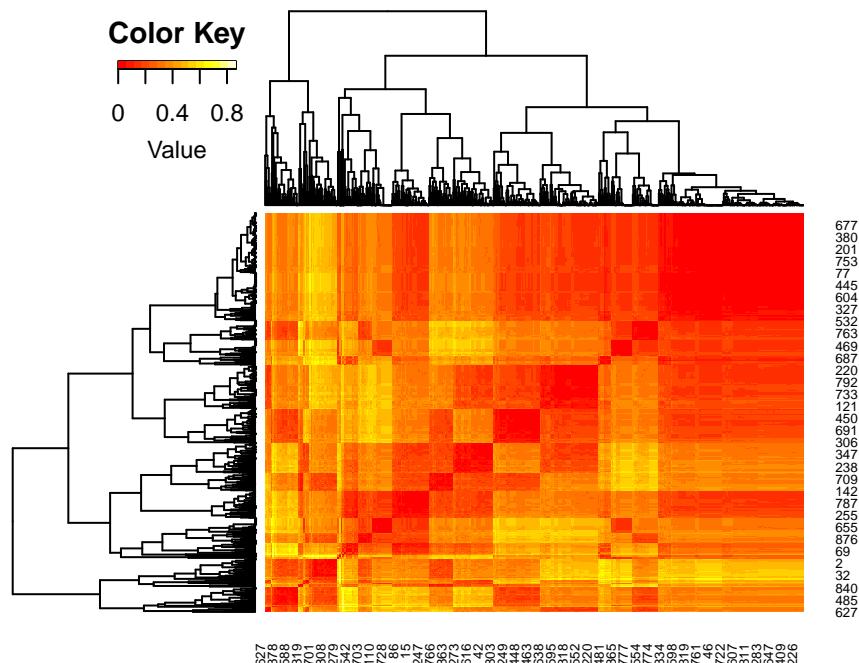
W tabeli 3 znajduje się opis danych titanic już po usunięciu zmiennych identyfikacyjnych.

3.3 Redukcja wymiaru

Następnym krokiem jest redukcja wymiaru danych, aby było je łatwiej zwizualizować. Na początku wyznaczę macierz odmienności, by następnie przeskalaować dane.

```
# Wyznaczenie macierzy odmienności
dissimilarities <- daisy(titanic.to.scale, type=list(), stand=T)
dis.matrix <- as.matrix(dissimilarities)
```

Po wyznaczeniu tej macierzy zwizualizuję ją w postaci mapy cieplnej.



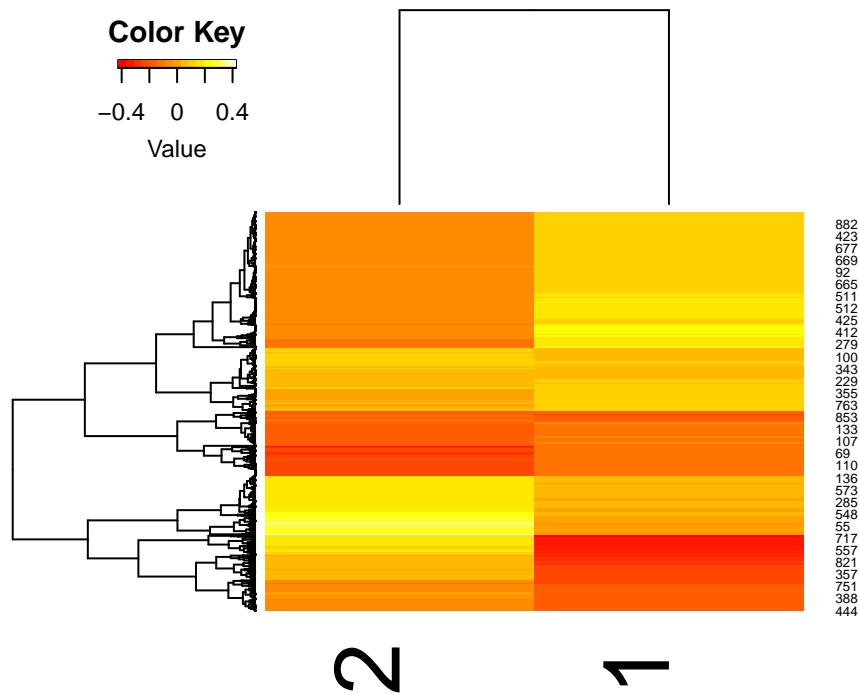
Rysunek 27: Mapa cieplna dla macierzy odmienności danych titanic

Kolejnym krokiem będzie przeskalowanie tych danych. Jako wymiar docelowej przestrzeni przyjmę $d = 2$.

```
# Przeskalowanie
mds.k2 <- cmdscale(dis.matrix, k=2)

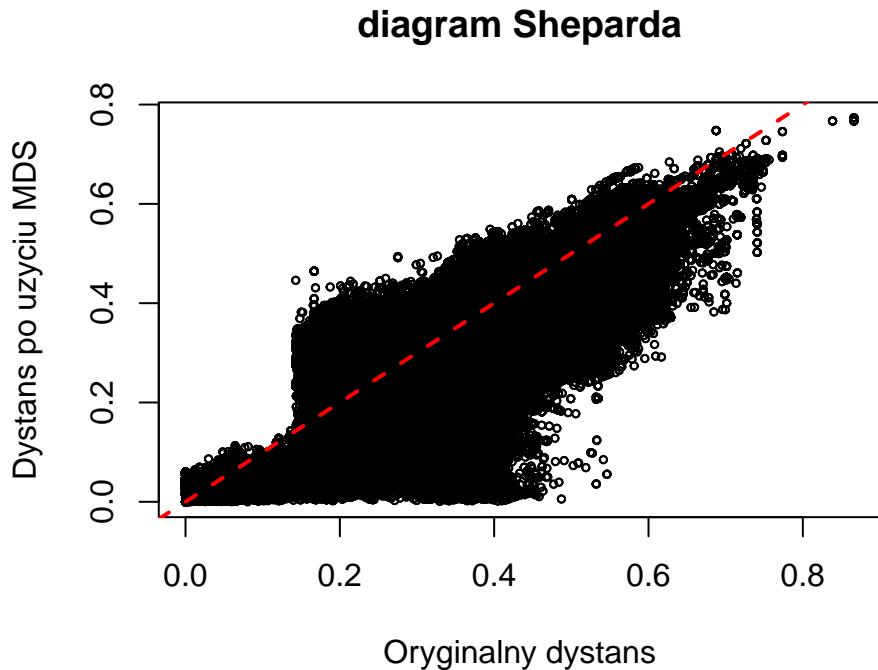
# Wyznaczenie odległości euklidesowych w nowej przestrzeni
dist.mds.k2 <- dist(mds.k2, method="euclidean")
dist.mds.k2 <- as.matrix(dist.mds.k2)
```

Nowe odległości również przedstawię na mapie cieplnej.



Rysunek 28: Mapa cieplna dla macierzy odmienności przeskalowanych danych titanic

Kolejnym krokiem będzie zbadanie jakości odwzorowania przy pomocy diagramu Sheparda.



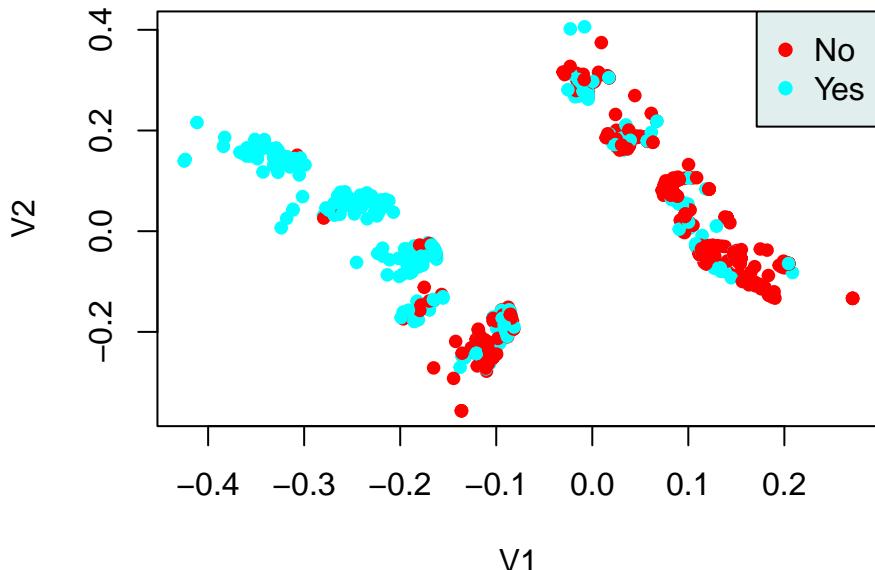
Rysunek 29: Diagram Sheparda dla przeskalowanych danych titanic

Rysunek 29 pokazuje, że odwzorowanie pozostawia wiele do życzenia, choć dane i tak są skupione blisko siebie, więc nie jest aż tak źle.

3.4 Wizualizacja danych

Ostatnim etapem tej analizy będzie wizualizacja danych z podziałem na poszczególne grupy. W pierwszej kolejności zwizualizuję dane względem tego, czy dana osoba przetrwała katastrofę, czy też nie.

Dane titanic – wykres rozrzutu 2D (Survived)

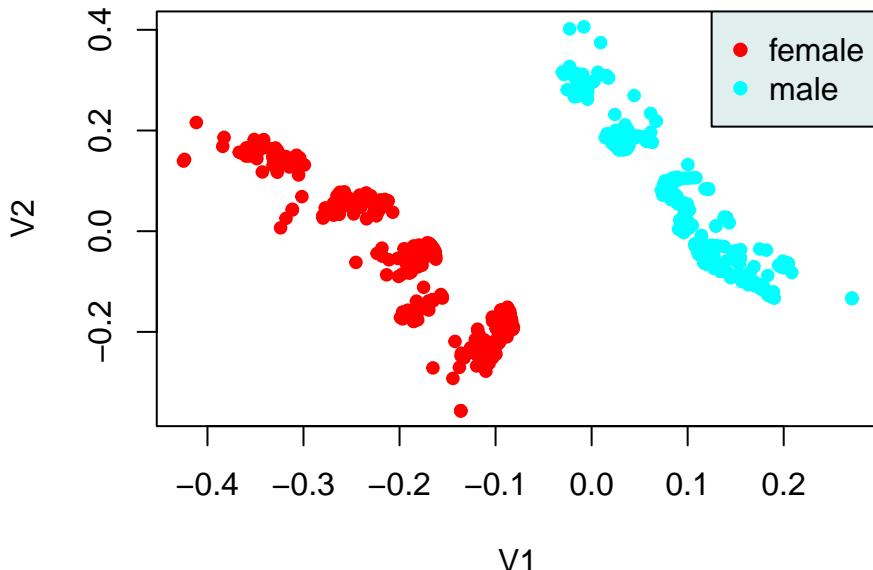


Rysunek 30: Wykres rozrzutu dla przeskalowanych danych titanic z podziałem względem zmiennej Survived

Na rysunku 30 łatwo można zauważać dwa skupiska, które w pewnym stopniu pokrywają się z informacją na temat przeżycia katastrofy, jednak nie w pełni – grupy częściowo się mieszają. Występują pojedyncze obserwacje odstające – najbardziej rzuca się w oczy czerwona kropka najdalej na prawo.

Dalej sprawdzę jeszcze wyniki względem płci pasażerów oraz ich klasy.

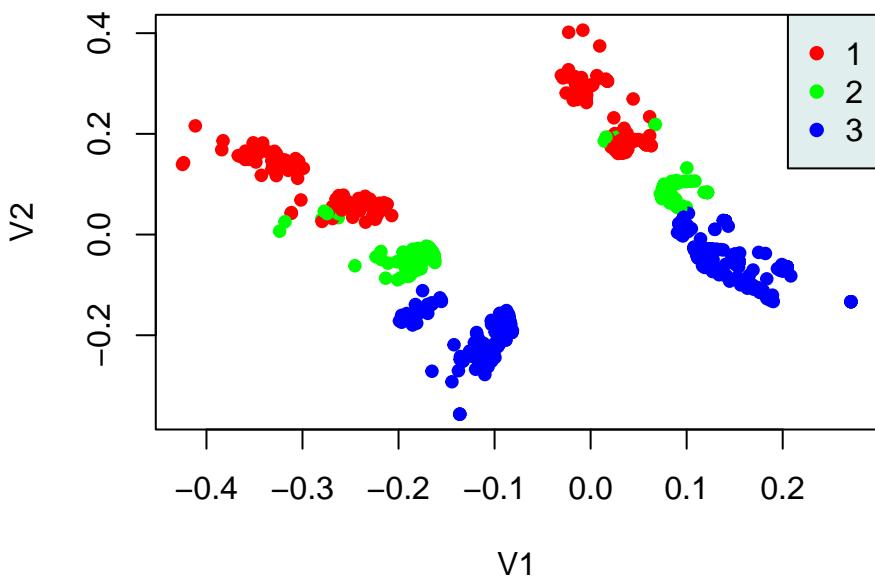
Dane titanic – wykres rozrzutu 2D (Sex)



Rysunek 31: Wykres rozrzutu dla przeskalowanych danych titanic z podziałem względem zmiennej Sex

Rysunek 31 pokazuje, że skupiska są związane przede wszystkim z płcią – skupisko po lewej jest skupiskiem kobiet, natomiast skupisko po prawej to mężczyźni.

Dane titanic – wykres rozrzutu 2D (Pclass)



Rysunek 32: Wykres rozrzutu dla przeskalowanych danych titanic z podziałem względem zmiennej Pclass

Jeśli chodzi o klasę, rysunek 32 pokazuje względnie równomierne rozmieszczenie klas w obu skupiskach.

3.5 Wnioski

W wizualizacji danych względem wybranych zmiennych widać, że wśród osób, które przetrwały katastrofę, były przede wszystkim kobiety – jest to zgodne z faktem, że najpierw ewakuowano kobiety i dzieci. Klasa, w której podróżowano, ma nieco mniejsze znaczenie, jednak można zauważać, że wśród kobiet dużo mniej przetrwało tych, które były w klasie trzeciej. W przypadku mężczyzn również można zauważać mniejszą liczbę osób, które przetrwały, wśród tych, które podróżowały w klasie trzeciej. Wynika to z faktu, że osoby z tej klasy miały utrudniony dostęp do ewakuacji.