

Sprawozdanie z listy nr 2

Eksploracja Danych

Dawid Skowroński 282241, Mateusz Cieślak 272633

2025-04-30

Spis treści

1 Zadanie 1	1
1.1 Wczytanie i wstępne przygotowanie danych	1
1.2 Wybór cech	2
1.3 Porównanie metod dyskretyzacji nienadzorowanej	3
1.3.1 Zmienna Petal.Length	3
1.3.2 Zmienna Sepal.Width	7
1.4 Wnioski	11
2 Zadanie 2	11
2.1 Wczytanie i wstępne przygotowanie danych	11
2.2 Eksploracja danych ilościowych	13
2.3 Analiza głównych składowych (PCA)	14
2.4 Skumulowana wariancja	15
2.5 Wizualizacja zmiennych i miast w przestrzeni PCA	16
2.6 Korelacje zmiennych i biplot	19
2.7 Wnioski końcowe	20
3 Zadanie 3	21
3.1 Przygotowanie danych	21
3.2 Wstępna analiza danych	23
3.3 Redukcja wymiaru na bazie MDS	23
3.4 Wizualizacja danych	25
3.5 Podsumowanie	28

Spis wykresów

1 Wykresy pudełkowe przedstawiające rozkład poszczególnych cech zbioru iris z podziałem na gatunki	2
2 Wykres rozrzutu zmiennej Petal.Length	3
3 Dyskretyzacja zmiennej Petal.Length oparta na jednakowej częstości. Porównanie z rzeczywistymi klasami	4
4 Dyskretyzacja zmiennej Petal.Length oparta na przedziałach jednakowej długości. Porównanie z rzeczywistymi klasami	4
5 Dyskretyzacja zmiennej Petal.Length oparta na metodzie k-średnich. Porównanie z rzeczywistymi klasami	5
6 Dyskretyzacja zmiennej Petal.Length oparta na zadanych ręcznie przedziałach. Porównanie z rzeczywistymi klasami	5
7 Porównanie metod dyskretyzacji zmiennej Petal.Length - wykresy mozaikowe	6

8	Wykres rozrzutu zmiennej Sepal.Width	7
9	Dyskretyzacja zmiennej Sepal.Width oparta na jednakowej częstości. Porównanie z rzeczywistymi klasami	7
10	Dyskretyzacja zmiennej Sepal.Width oparta na przedziałach jednakowej długości. Porównanie z rzeczywistymi klasami	8
11	Dyskretyzacja zmiennej Sepal.Width oparta na metodzie k-średnich. Porównanie z rzeczywistymi klasami	8
12	Dyskretyzacja zmiennej Sepal.Width oparta na zadanych ręcznie przedziałach. Porównanie z rzeczywistymi klasami	9
13	Porównanie metod dyskretyzacji zmiennej Sepal.Width - wykresy mozaikowe	10
14	Wykresy pudełkowe przedstawiające rozkład poszczególnych cech zbioru City Quality of Life Dataset	13
15	Wykresy pudełkowe przedstawiające rozrzuł pierwszych trzech składowych PCA	14
16	Wykres słupkowy przedstawiający skumulowaną wariancję dla kolejnych składowych	15
17	Wykres zmiennych PCA w przestrzeni 2d i wykres obserwacji PCA pogrupowany według kontynentu	16
18	Podpisane obserwacje w celu identyfikacji miast najmocniej odstających.	18
19	Macierz korelacji i biplot dla wszystkich zmiennych ilościowych	19
20	Macierz korelacji i biplot dla wszystkich zmiennych ilościowych	20
21	Macierz korelacji wybranych cech zbioru titanic train	23
22	Wykresy Sheparda zbioru titanic train dla wybranych wymiarów	24
23	Wykres wartości zmiennej STRESS w zależności od liczby wymiarów	25
24	Dwuwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Survived)	25
25	Dwuwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Sex i Pclass)	26
26	Trójwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Survived)	27
27	Dwuwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Age i Survived)	28

Spis tabel

1	Struktura zbioru danych iris	1
2	Wartości brakujące w zbiorze iris	1
3	Liczba obserwacji w zbiorze iris ze względu na gatunek	2
4	Zgodność przypisania obiektów do poszczególnych grup (zmienna Sepal.Width)	6
5	Zgodność przypisania obiektów do poszczególnych grup (zmienna Sepal.Width)	10
6	Ładunki pierwszych trzech głównych składowych (PCA)	14
7	Struktura zbioru danych titanic train	21
8	Wartości brakujące w zbiorze Titanic	22

1 Zadanie 1

1.1 Wczytanie i wstępne przygotowanie danych

Tabela 1: Struktura zbioru danych iris

Zmienna	Opis	Typ	Przykładowe wartości
Sepal.Length	Długość działki kielicha	numeric	5.1, 4.9, 4.7, 4.6
Sepal.Width	Szerokość działki kielicha	numeric	3.5, 3, 3.2, 3.1
Petal.Length	Długość płatka	numeric	1.4, 1.4, 1.3, 1.5
Petal.Width	Szerokość płatka	numeric	0.2, 0.2, 0.2, 0.2
Species	Gatunek irysa	factor	setosa, setosa, setosa, setosa

Zbiór danych `iris` zawiera pomiary o dla trzech gatunków irysów (`setosa`, `virginica`, `versicolor`). Opis poszczególnych zmiennych wraz z przykładowymi wartościami przedstawiony jest w tabeli 1.

Celem analizy jest:

- Wybranie cechy o najlepszej i najgorszej zdolności dyskryminacyjnej.
- Porównanie skuteczności nienadzorowanych metod dyskretyzacji (`equal width`, `equal frequency`, `k-means`, `przedziały zadane przez użytkownika`).
- Ocena, czy wyniki różnią się w zależności od jakości danej cechy.

Sprawdzamy czy w zbiorze występują wartości brakujące

Tabela 2: Wartości brakujące w zbiorze iris

zmienna	liczba NA
Sepal.Length	0
Sepal.Width	0
Petal.Length	0
Petal.Width	0
Species	0

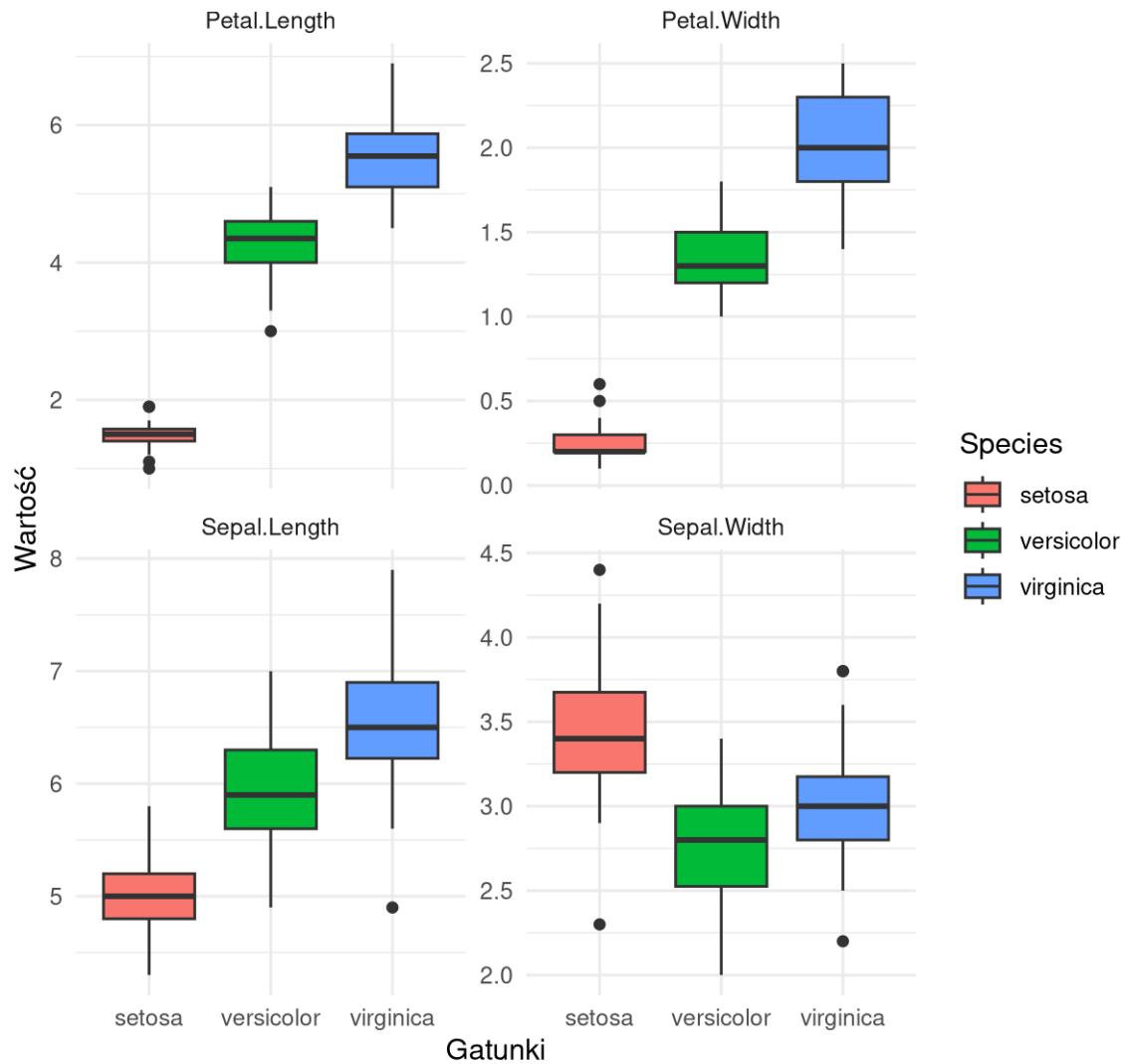
Tabela 3: Liczba obserwacji w zbiorze iris ze względu na gatunek

zmienna	liczba obserwacji
setosa	50
versicolor	50
virginica	50

Analizując tabelę 2 nie zauważamy wartości `NA`. Dodatkowo liczba obserwacji w zbiorze iris jest równoliczna ze względu na gatunek, co obrazuje tabela 3.

1.2 Wybór cech

Rozkłady cech względem poszczególnych gatunków



Wykres 1: Wykresy pudełkowe przedstawiające rozkład poszczególnych cech zbioru iris z podziałem na gatunki

Z wykresów 1 możemy zaobserwować, która zmienna ma najlepsze zdolności dyskryminacyjne. Najlepszą cechą wydaje się być Petal.Length. Wykresy pudełkowe najmniej na siebie nachodzą, zauważalne są duże różnice między gatunkami (w szczególności między gatunkiem setosa a pozostałymi). Jedynie w grupach versicolor i virginica występuje nieznaczne nakładanie się wartości przyjmowane przez tę zmienną.

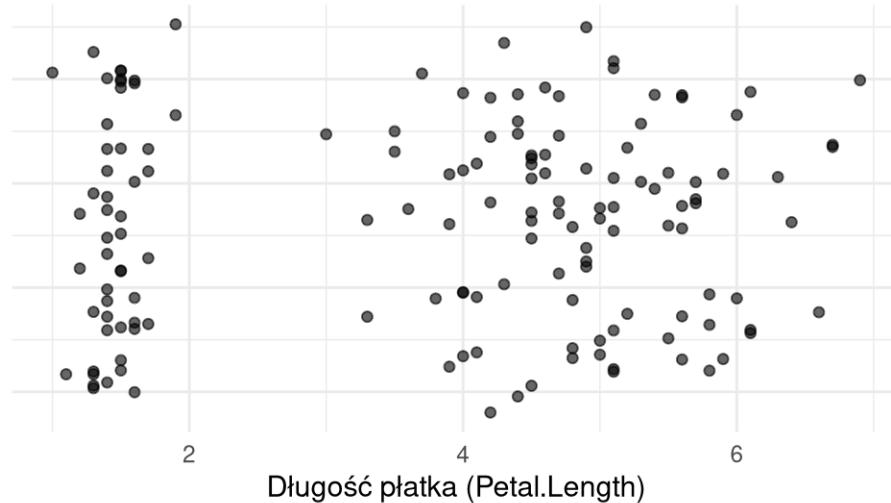
Cechą o najgorszej zdolności dyskryminacyjnej wydaje się być zmienna Sepal.Width. Rozkłady zmiennych pokrywają się w znacznej części, może to zaburzyć klarowność dyskretyzacji. W szczególności, w grupie setosa i virginica można zauważać obserwacje odstające, których wartości pokrywają się z wartościami w innych grupach.

1.3 Porównanie metod dyskretyzacji nienadzorowanej

1.3.1 Zmienna Petal.Length

Ustalamy docelową liczbę przedziałów na 3 i dokonujemy dyskretyzacji zmiennej Petal.Length metodą opartą na równych częstościach

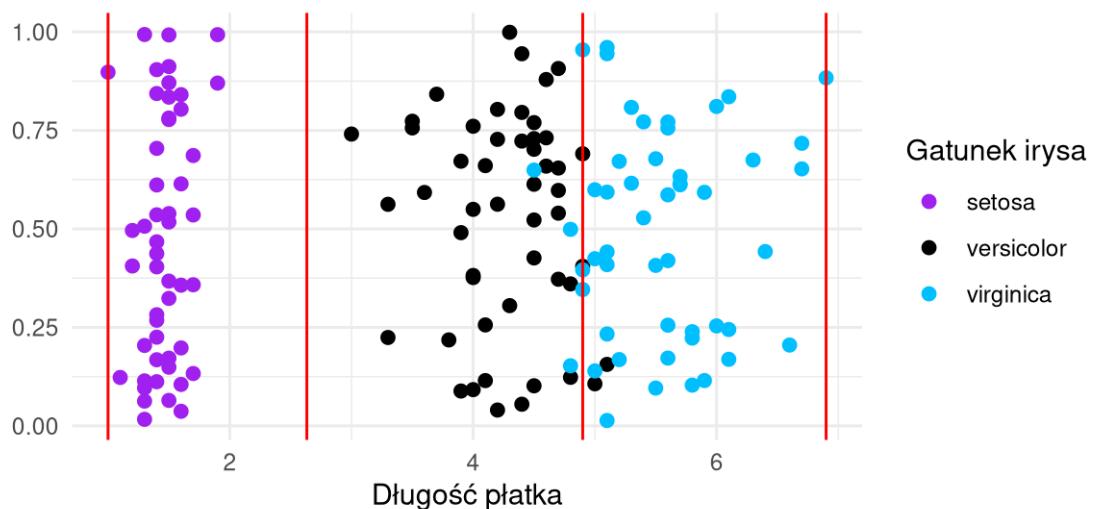
Rozrzut zmiennej Petal.Length



Wykres 2: Wykres rozrzutu zmiennej Petal.Length

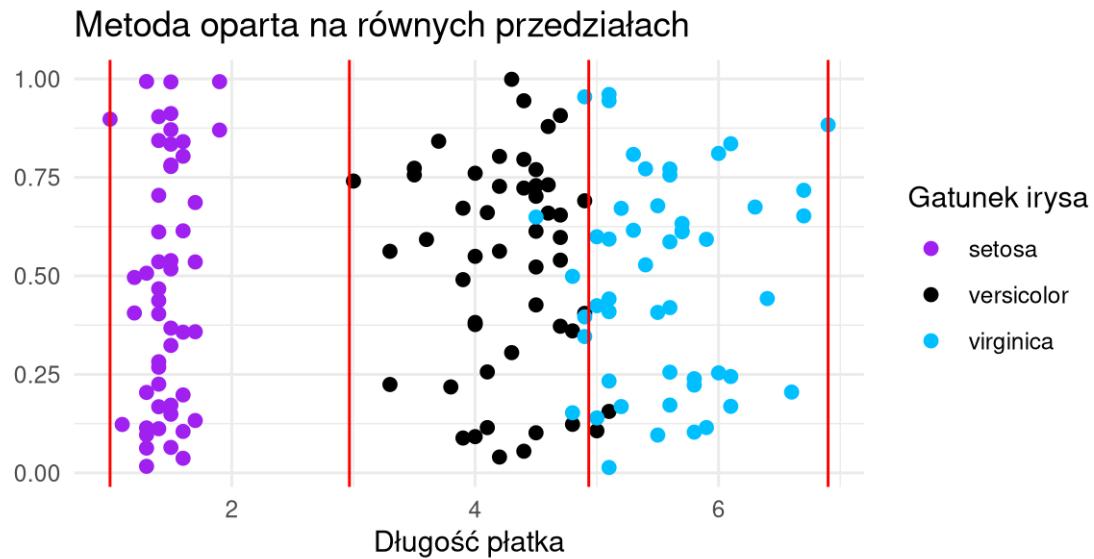
Wykres 2 obrazuje rozproszenie zmiennej Petal.Length (dla przejrzystości zastosowano losowe rozmieszczenie punktów względem współrzędnej y). Już we wstępnej analizie możemy zauważać, że dane rozmieszczone są w dwóch "skupiskach".

Metoda oparta na równych częstościach



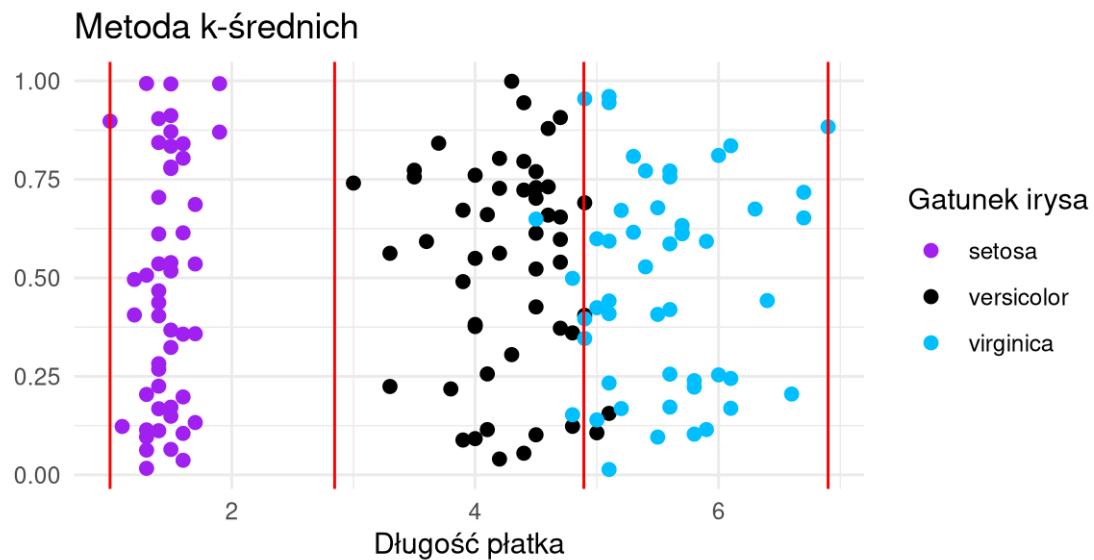
Wykres 3: Dyskretyzacja zmiennej Petal.Length oparta na jednakowej częstości. Porównanie z rzeczywistymi klasami

Stosując metodę opartą na przedziałach o jednakowej szerokości możemy zaobserwować bardzo dobrą dyskretyzację zmiennej Sepal.Length. Widoczna jest całkowita separacja klasy setosa. Pozostałe gatunki również są bardzo dobrze odseparowane.



Wykres 4: Dyskretyzacja zmiennej Petal.Length oparta na przedziałach jednakości długości. Porównanie z rzeczywistymi klasami

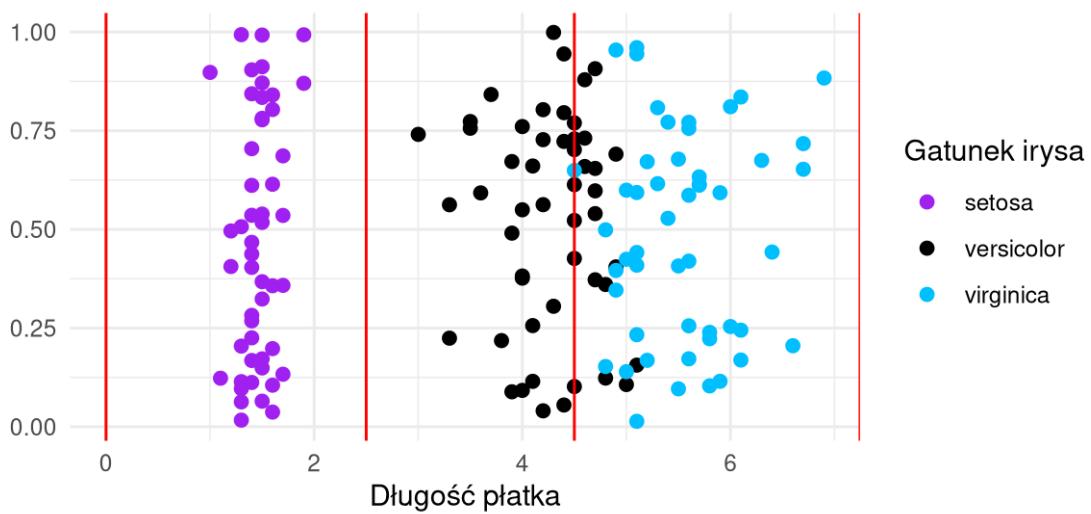
Stosując metodę opartą na przedziałach jednakości uzyskujemy wyniki zbliżone co do metody opartej na równolicznych przedziałach.



Wykres 5: Dyskretyzacja zmiennej Petal.Length oparta na metodzie k-średnich. Porównanie z rzeczywistymi klasami

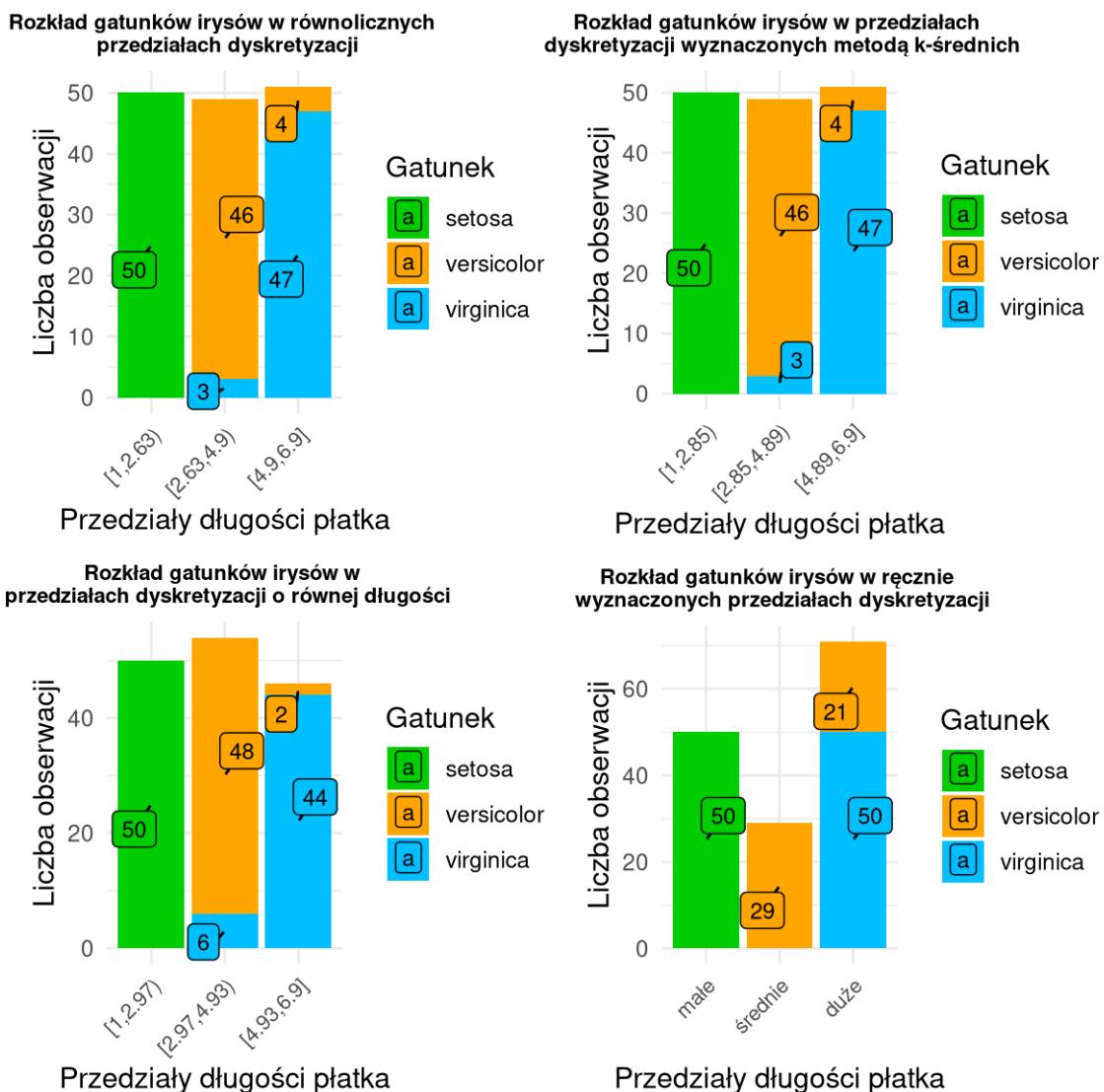
Metoda oparta na k-średnich (wykres 5) zwraca zbliżone rezultaty jak metoda równolicznych przedziałów i jednakowych odległości.

Podział zadany ręcznie



Wykres 6: Dyskretyzacja zmiennej Petal.Length oparta na zadanych ręcznie przedziałach. Porównanie z rzeczywistymi klasami

Dokładność metody opartej na przedziałach zadanych przez użytkownika (wykres 6) zwraca nieco gorsze wyniki (część obserwacji z gatunku `versicolor` znajduje się w klastrze gatunku `virginica`). W przypadku zmiennej `Sepal.Length` dyskretyzacja gatunku `setosa` jest intuicyjna, gdyż jest odseparowana od pozostałych gatunków. Lecz zadanie przedziałów dla pozostałych zmiennych jest kwestią subiektywną, przez co dokładność tej metody może się znacznie ważyć.



Wykres 7: Porównanie metod dyskretyzacji zmiennej Petal.Length - wykresy mozaikowe

Wykresy mozaikowe 7 podsumowują wszystkie metody dyskretyzacji zastosowane dla zmiennej `Petal.Length` oraz rozłożenie gatunków irysów w poszczególnych. Możemy zauważać, że podział gatunków jest najmniej klarowny dla metody ręcznie wyznaczonych przedziałów. Drugą najgorszą metodą jest ta oparta na równych przedziałach.

Tabela 4: Zgodność przypisania obiektów do poszczególnych grup (zmienna Sepal.Width)

Metoda	Współczynnik zgodności
Równa liczność	0.9533
Równe przedziały	0.9467
Metoda k-średnich	0.9533
Ręcznie ustalone przedziały	0.8600

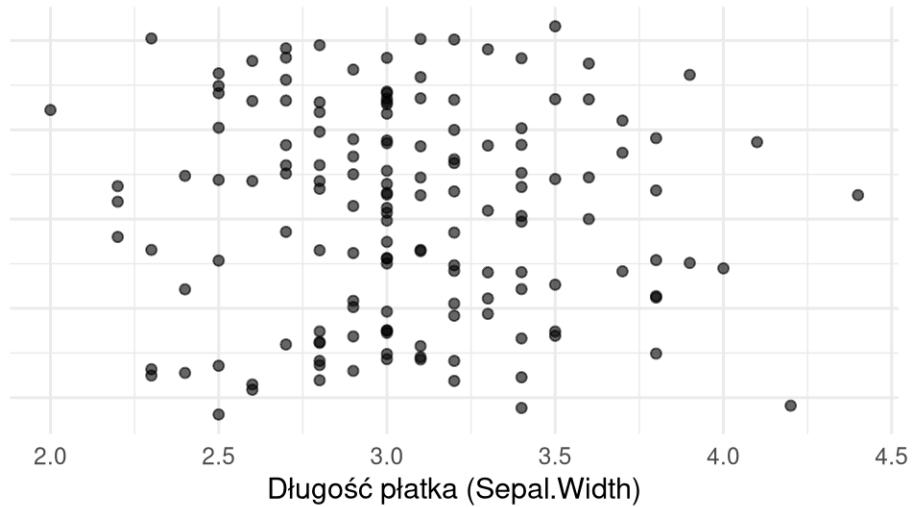
Tabela 4 przedstawia współczynniki zgodności dla wszystkich metod dyskretyzacji. Możemy zauważać, że

zarówna metoda k-średnich jak i metoda równolicznych przedziałów zwraca najlepszą zgodność (**0.953**) z rzeczywistymi klasami (gatunkami). Potwierdza to wstępna obserwację z wykresów pudełkowych - `Petal.Length` naturalnie dobrze separuje gatunki. Wszystkie metody skutecznie oddzielają `setosa`, problemem jest głównie nakładanie `versicolor` i `virginica`

1.3.2 Zmienna Sepal.Width

Dla zmiennej `Sepal.Width` przeprowadzamy podobną analizę.

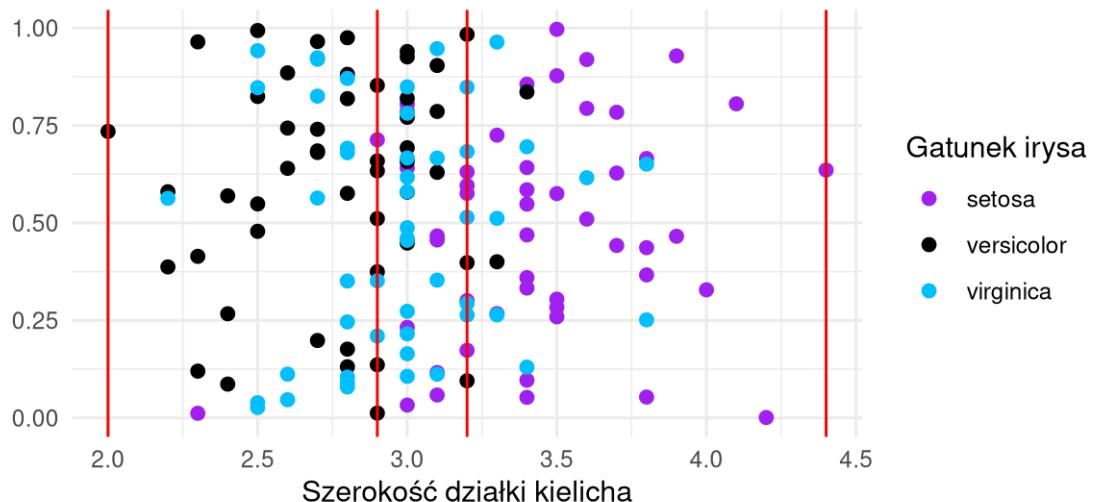
Rozrzut zmiennej `Sepal.Width`



Wykres 8: Wykres rozrzutu zmiennej `Sepal.Width`

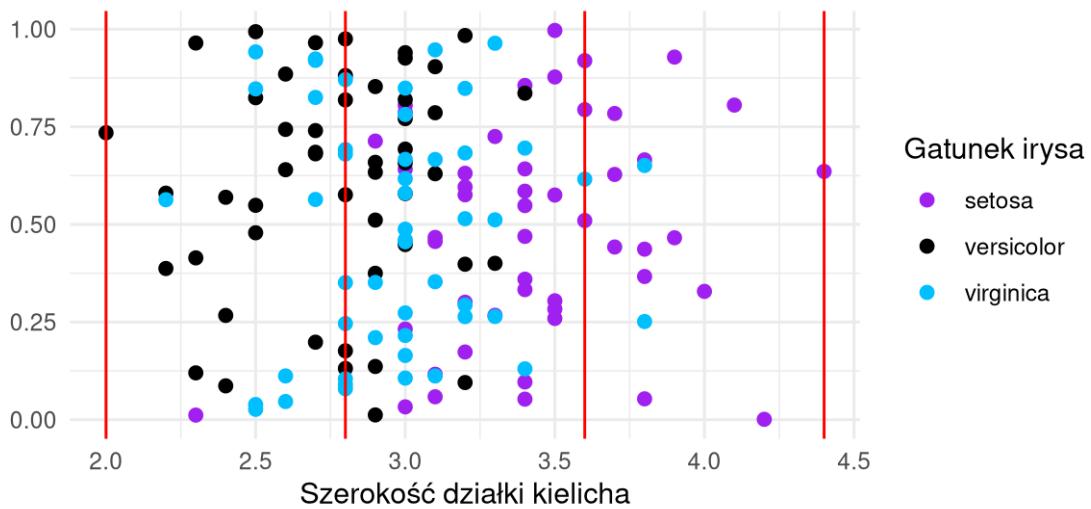
Wykres 8 przedstawia rozrzut zmiennej `Sepal.Width`. W przypadku tej zmiennej podział danych na skupiska nie występuje.

Metoda oparta na równych częstościach



Wykres 9: Dyskretyzacja zmiennej `Sepal.Width` oparta na jednakowej częstości. Porównanie z rzeczywistymi klasami

Metoda oparta na równych przedziałach



Wykres 10: Dyskretyzacja zmiennej Sepal.Width oparta na przedziałach jednakowej długości. Porównanie z rzeczywistymi klasami

Metoda k-średnich



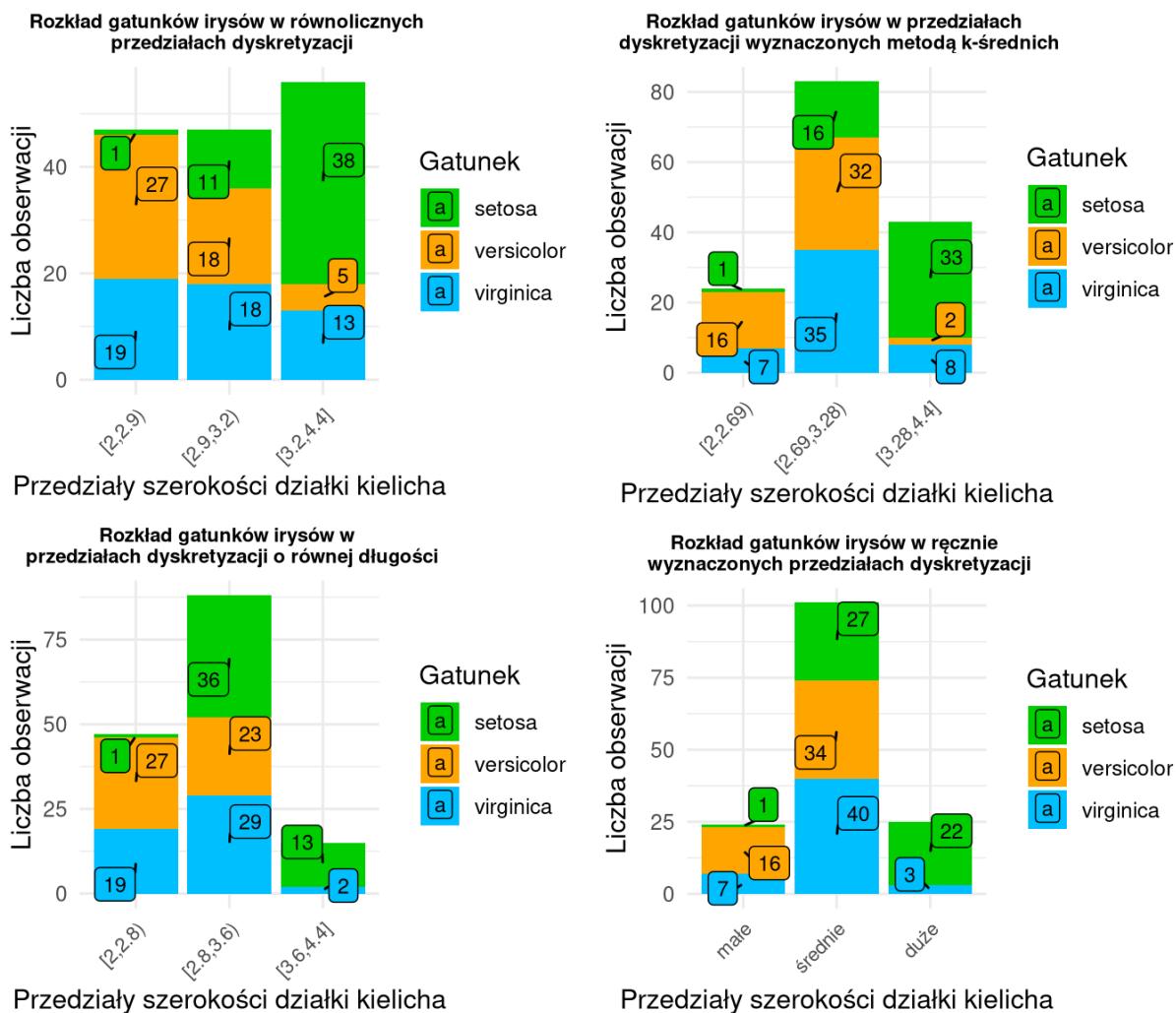
Wykres 11: Dyskretyzacja zmiennej Sepal.Width oparta na metodzie k-średnich. Porównanie z rzeczywistymi klasami

Przedziały zadane ręcznie



Wykres 12: Dyskretyzacja zmiennej Sepal.Width oparta na zadanych ręcznie przedziałach. Porównanie z rzeczywistymi klasami

Analizując wykresy 9,10,11 oraz 12 zastosowania metod dyskretyzacji dla zmiennej Sepal.Width żadna z nich nie wykazuje dobrego wyodrębnienia na poszczególne gatunki.



Wykres 13: Porównanie metod dyskretyzacji zmiennej Sepal.Width - wykresy mozaikowe

Wykresy mozaikowe 13 wyraźnie pokazują, że żadna metoda dyskretyzacji Sepal.Width nie osiąga tak czystej separacji jak dla Petal.Length. Wszystkie wykresy pokazują silne nakładanie się gatunków w przedziałach.

Tabela 5: Zgodność przypisania obiektów do poszczególnych grup (zmienna Sepal.Width)

Metoda	Współczynnik zgodności
Równa liczność	0.5600
Równe przedziały	0.4126
Metoda k-średnich	0.4785
Ręcznie ustalone przedziały	0.3333

Zestawienie metod dyskretyzacji w tabeli 5 pokazuje, że żadna z metod nie osiąga zadowalającej separacji gatunków. Najlepszą metodą spośród wszystkich (współczynnik zgodności **0.56**). Najgorszy współczynnik ma natomiast mają natomiast przedziały ustalone ręcznie (**0.333**).

1.4 Wnioski

Analiza wyraźnie pokazuje, że skuteczność metod dyskretyzacji silnie zależy od właściwości dyskryminacyjnych samej cechy.

- Długość płatka `Petal.Length` skutecznie separuje poszczególne klasy.
- Metoda równolicznych przedziałów jest najlepsza zarówno dla `Sepal.Width` jak i `Petal.Length`.
- Najgorszą metodą w obu przypadkach jest ta, gdzie użytkownik wyznacza przedziały (jednak dokładność tej metody może się zmieniać).
- Dla `Sepal.Width` nawet najlepsza metoda daje niezadowalające wyniki, co sugeruje, że sama dyskretyzacja dla tej zmiennej może być niewystarczająca, aby wyróżnić poszczególne gatunki irysów.

2 Zadanie 2

2.1 Wczytanie i wstępne przygotowanie danych

```
qol_base <- read.csv(file = paste0(getwd(), "/uaScoresDataFrame.csv"),
  header = TRUE, stringsAsFactors = TRUE)

# Usunięcie zbędnej kolumny
qol <- qol_base[, -1]

# Dane były zbierane jako ocena od różnych osób,
# więc to mało prawdopodobne, aby wszystkie osoby
# ocenili dany aspekt na 0. W takim wypadku
# zakładamy, że 0 to wartość brakująca / brak
# odpowiedzi w ankcie.
qol[qol == 0] <- NA

# Sprawdzenie typów zmiennych
str(qol)

## 'data.frame': 266 obs. of 20 variables:
## $ UA_Name : Factor w/ 264 levels "Aarhus","Adelaide",...
## $ UA_Country : Factor w/ 135 levels "Alabama","Alaska",...
## $ UA_Continent : Factor w/ 6 levels "Africa","Asia",...
## $ Housing : num 6.13 6.31 7.26 9.28 3.05 ...
## $ Cost.of.Living : num 4.02 4.69 6.06 9.33 3.82 ...
## $ Startups : num 2.83 3.14 3.77 2.46 7.97 ...
## $ Venture.Capital : num 2.51 2.64 1.49 NA 6.11 ...
## $ Travel.Connectivity : num 3.54 1.78 1.46 4.59 8.32 ...
## $ Commute : num 6.31 5.34 5.06 5.87 6.12 ...
## $ Business.Freedom : num 9.94 9.4 8.67 5.57 8.84 ...
## $ Safety : num 9.62 7.93 1.34 7.31 8.5 ...
## $ Healthcare : num 8.7 7.94 6.43 4.55 7.91 ...
## $ Education : num 5.37 5.14 4.15 2.28 6.18 ...
## $ Environmental.Quality: num 7.63 8.33 7.32 3.86 7.6 ...
## $ Economy : num 4.89 6.07 6.51 5.27 5.05 ...
## $ Taxation : num 5.07 4.59 4.35 8.52 4.95 ...
## $ Internet.Access : num 8.37 4.34 5.4 2.89 4.52 ...
## $ Leisure...Culture : num 3.19 4.33 4.89 2.94 8.87 ...
## $ Tolerance : num 9.74 7.82 7.03 6.54 8.37 ...
## $ Outdoors : num 4.13 5.53 3.52 5.5 5.31 ...
```

```

# Konwersja kolumn z typem factor na character
# (UA_Continent zostawiamy jako factor, bo jest
# tylko 6 poziomów)
qol$UA_Name <- as.character(qol$UA_Name)
qol$UA_Country <- trimws(as.character(qol$UA_Country))

# Kolumny zawierające NA
colSums(is.na(qol))

##          UA_Name           UA_Country        UA_Continent
##             0                  0                   0
##      Housing       Cost.of.Living        Startups
##             1                  18                  2
## Venture.Capital   Travel.Connectivity Commute
##                74                  0                  11
## Business.Freedom            Safety Healthcare
##                 2                  0                  1
## Education Environmental.Quality Economy
##                 26                  0                  3
## Taxation        Internet.Access Leisure...Culture
##                 0                  0                  2
## Tolerance          Outdoors
##                 0                  0

# Liczba wierszy zawierających NA
sum(apply(qol, 1, function(x) any(is.na(x))))
```

[1] 88

```

# Mamy bardzo dużo wartości NA, większość
# znajduje się w kolumnie Venture.capital, ale w
# innych kolumnach również. Jeśli chcielibyśmy je
# usunąć pozbylibyśmy się 88 wierszy, więc
# zamiast tego użyjemy metody uzupełniania
# wartości brakujących na podstawie najbliższych
# sąsiadów.
```

```

qol_imputed <- kNN(qol, k = 5, imp.var = FALSE)

# Rozróżnienie miast o tej samej nazwie
qol$UA_Name[qol$UA_Name == "Birmingham" & qol$UA_Country ==
    "United Kingdom"] <- "Birmingham (UK)"
qol$UA_Name[qol$UA_Name == "Birmingham" & qol$UA_Country ==
    "Alabama"] <- "Birmingham (Alabama)"
qol$UA_Name[qol$UA_Name == "Portland" & qol$UA_Country ==
    "Maine"] <- "Portland (Maine)"
qol$UA_Name[qol$UA_Name == "Portland" & qol$UA_Country ==
    "Oregon"] <- "Portland (Oregon)"

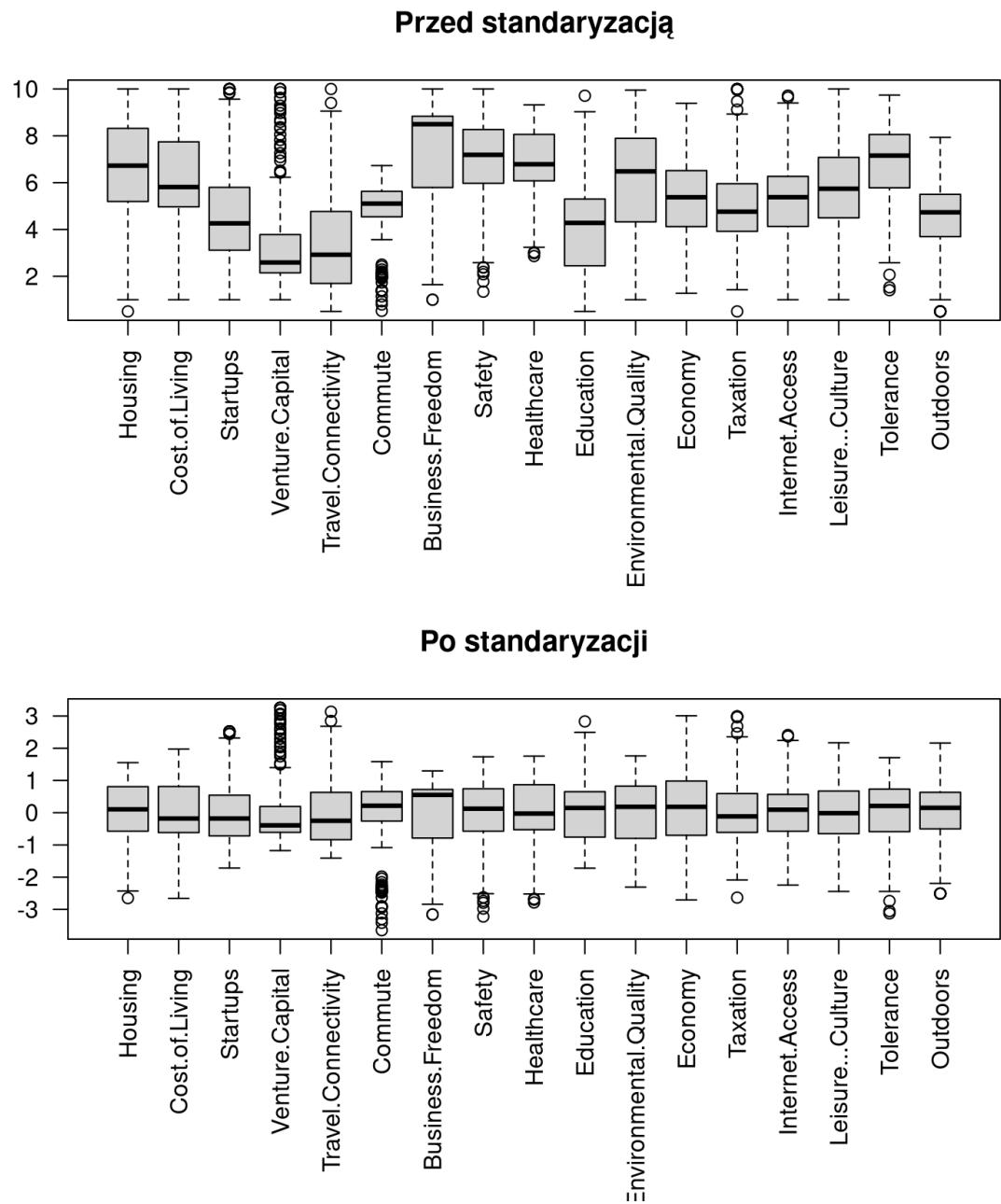
# Rozróżnienie krajów o tej samej nazwie
qol$UA_Country[qol$UA_Country == "Canada" & qol$UA_Continent ==
    "North America"] <- "Canada (NA)"
qol$UA_Country[qol$UA_Country == "Canada" & qol$UA_Continent ==
    "Europe"] <- "Canada (EU)"
qol$UA_Country[qol$UA_Country == "Georgia" & qol$UA_Continent ==
```

```

"North America"] <- "Georgia (NA)"
qol$UA_Country[qol$UA_Country == "Georgia" & qol$UA_Continent ==
"Europe"] <- "Georgia (EU)"

```

2.2 Eksploracja danych ilościowych

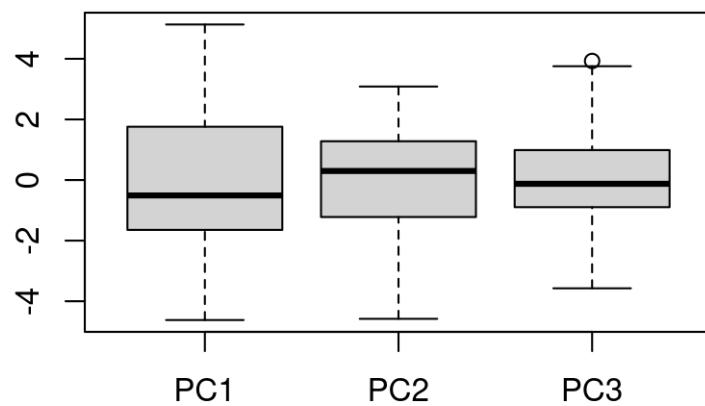


Wykres 14: Wykresy pudełkowe przedstawiające rozkład poszczególnych cech zbioru City Quality of Life Dataset

W naszym przypadku zastosowanie standaryzacji było konieczne, ponieważ zmienne mają różny rozrzut co widać na wykresie 14.

2.3 Analiza głównych składowych (PCA)

Rozrzut pierwszych trzech składowych PCA



Wykres 15: Wykresy pudełkowe przedstawiające rozrzut pierwszych trzech składowych PCA

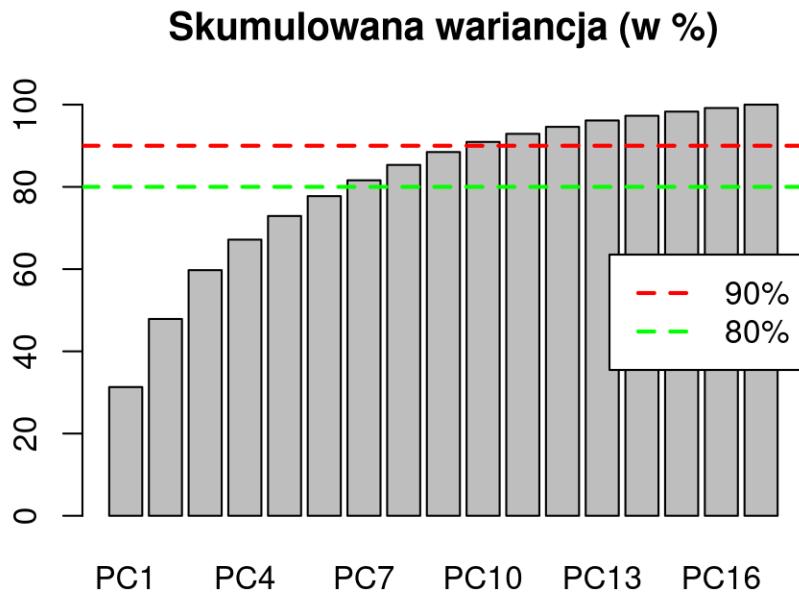
Tabela 6: Ładunki pierwszych trzech głównych składowych (PCA)

	PC1	PC2	PC3
Housing	0.314	0.157	-0.098
Cost.of.Living	0.374	-0.022	-0.109
Startups	-0.158	-0.450	-0.160
Venture.Capital	-0.167	-0.418	-0.192
Travel.Connectivity	-0.195	-0.045	-0.422
Commute	-0.091	0.283	-0.404
Business.Freedom	-0.370	0.081	0.165
Safety	-0.045	0.343	-0.333
Healthcare	-0.279	0.301	-0.161
Education	-0.387	-0.046	-0.058
Environmental.Quality	-0.327	0.204	0.166
Economy	-0.252	-0.176	0.390
Taxation	0.028	0.094	0.030
Internet.Access	-0.273	0.007	0.117
Leisure...Culture	-0.070	-0.295	-0.367
Tolerance	-0.194	0.333	-0.041
Outdoors	-0.086	-0.144	-0.276

Na podstawie tabeli 6 możemy wyciągnąć następujące wnioski wstępne:

- PC1: może oznaczać komfort ekonomiczny vs. jakość instytucjonalna i środowiskowa. Miasta z wysoką wartością PC1 mają niższe koszty życia i mieszkania, ale również niższą jakość edukacji, wolności biznesowej i środowiska.
- PC2: może oznaczać aktywność biznesową, startupową i kulturalną – mniejsze wartości PC2 to większa aktywność tych cech. Natomiast wyższe wartości PC2 oznaczają większą tolerancję, bezpieczeństwo i lepszą opiekę zdrowotną.
- PC3: może oznaczać równowagę między siłą gospodarki a jakością życia (dostępność, bezpieczeństwo, dojazdy).

2.4 Skumulowana wariancja

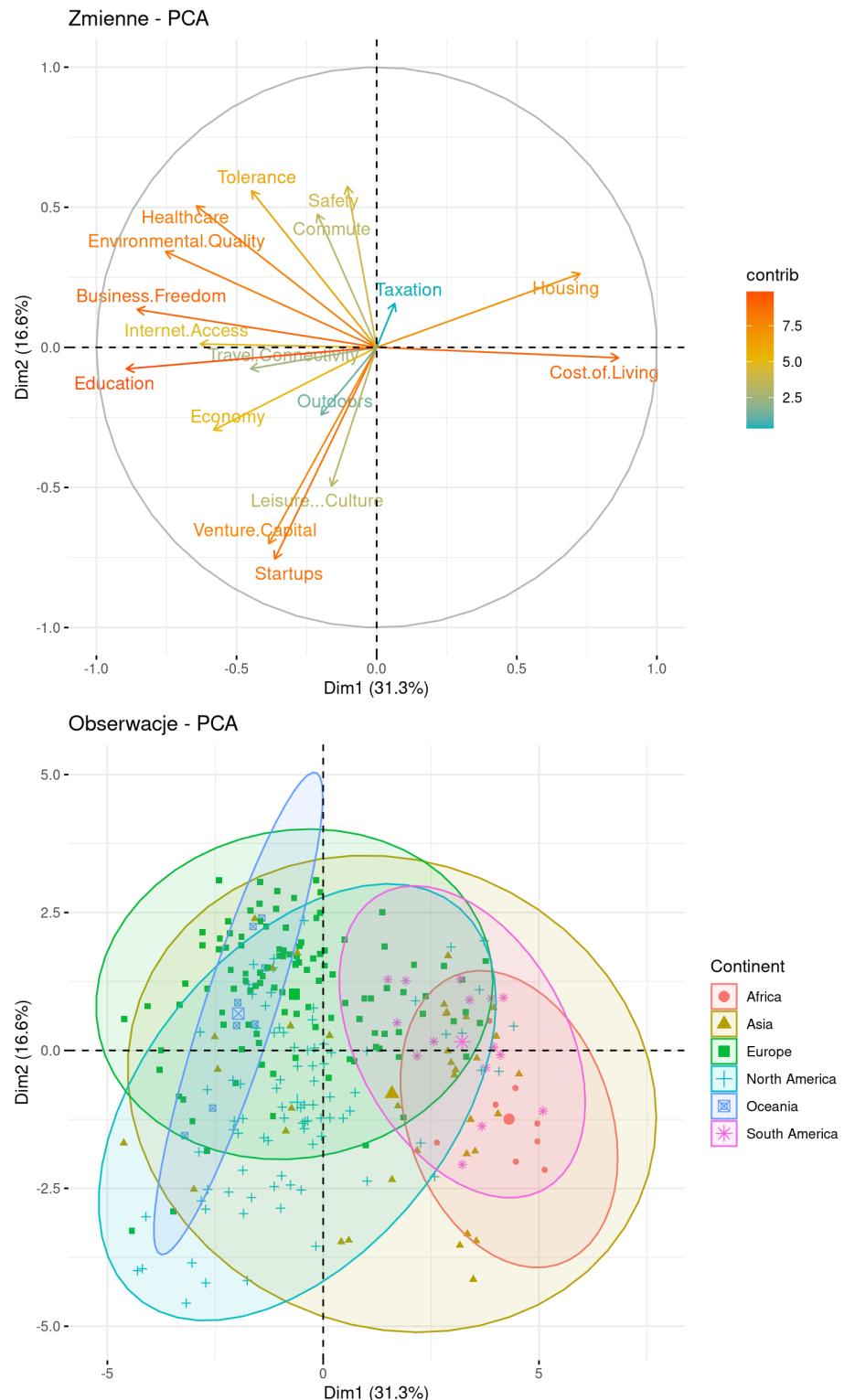


Wykres 16: Wykres słupkowy przedstawiający skumulowaną wariancję dla kolejnych składowych

Z wykresu 16 odczytujemy, że:

- Do wyjaśnienia 80% całkowitej zmienności danych jest potrzebnych 7 składowych głównych.
- Do wyjaśnienia 90% całkowitej zmienności danych jest potrzebnych 10 składowych głównych.

2.5 Wizualizacja zmiennych i miast w przestrzeni PCA



Wykres 17: Wykres zmiennych PCA w przestrzeni 2d i wykres obserwacji PCA pogrupowany według kontynentu

PC1: Największe ładunki:

- Cost.of.Living: +0.374
- Housing: +0.314
- Business.Freedom: -0.370
- Education: -0.387
- Environmental.Quality: -0.327

Interpretacja PC1: Miasta z niskim kosztem życia i mieszkaniem są po prawej stronie osi PC1. Miasta z lepszą edukacją, wolnością gospodarczą i jakością środowiska są po lewej stronie.

PC2: Największe ładunki:

- Startups: -0.450
- Venture.Capital: -0.418
- Tolerance: +0.333
- Healthcare: +0.301
- Safety: +0.343

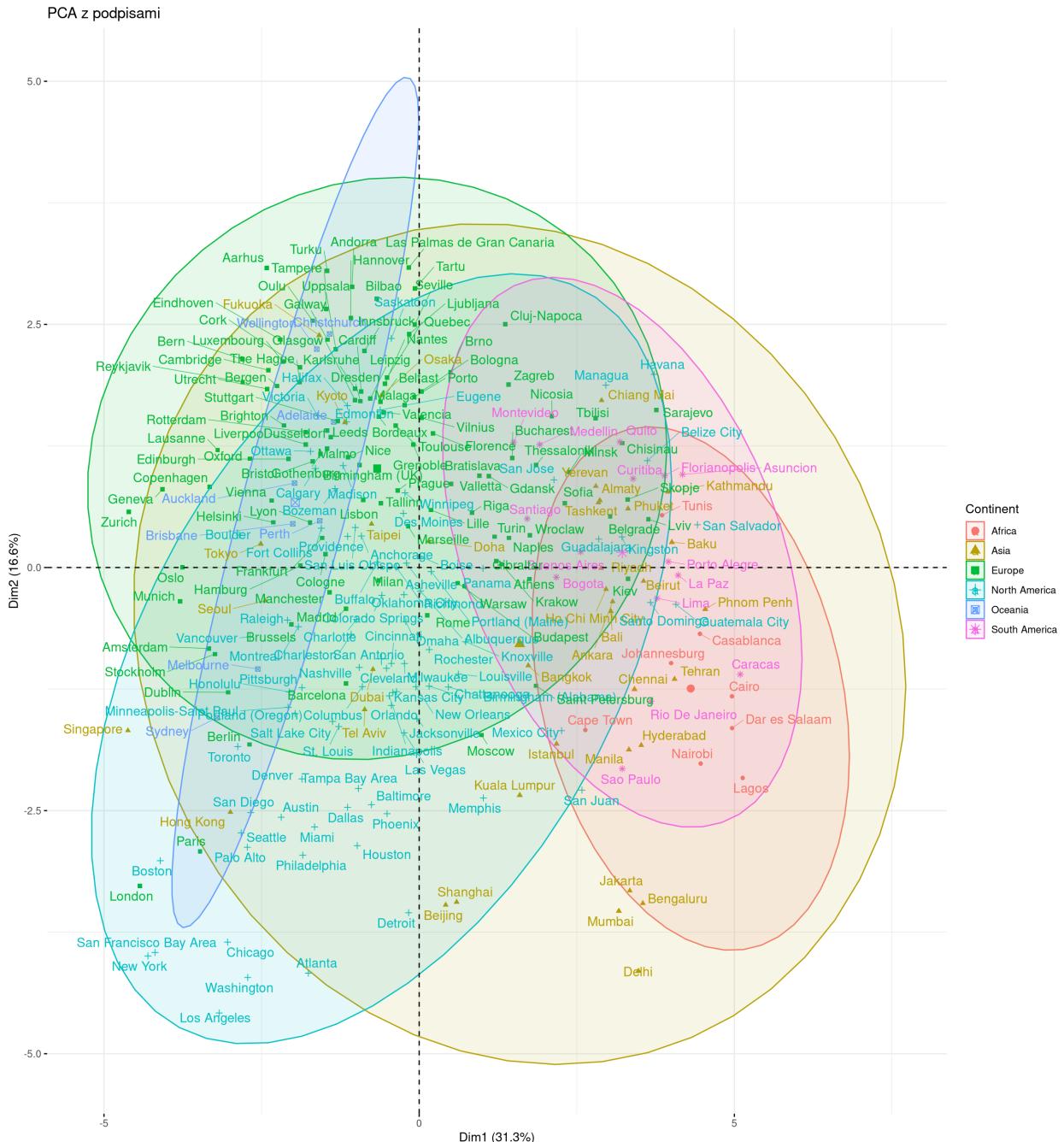
Interpretacja PC2: Na górze znajdują się miasta z lepszą opieką zdrowotną, bezpieczeństwem i komunikacją, a na dole miasta z większym potencjałem startupowym i dostępem do kapitału.

Europa i Ameryka Północna są do siebie stosunkowo zbliżone — miasta z tych regionów mają podobne profile cech. Ameryka północna jest trochę przesunięta wzdułóż osi PC2.

Azja jest rozproszona — duże zróżnicowanie miast.

Afryka i Ameryka Południowa raczej z jednej strony wykresu — sugeruje podobny typ profilu miast (np. niższe koszty życia, ale mniejszy rozwój edukacji, ekonomii i służby zdrowia).

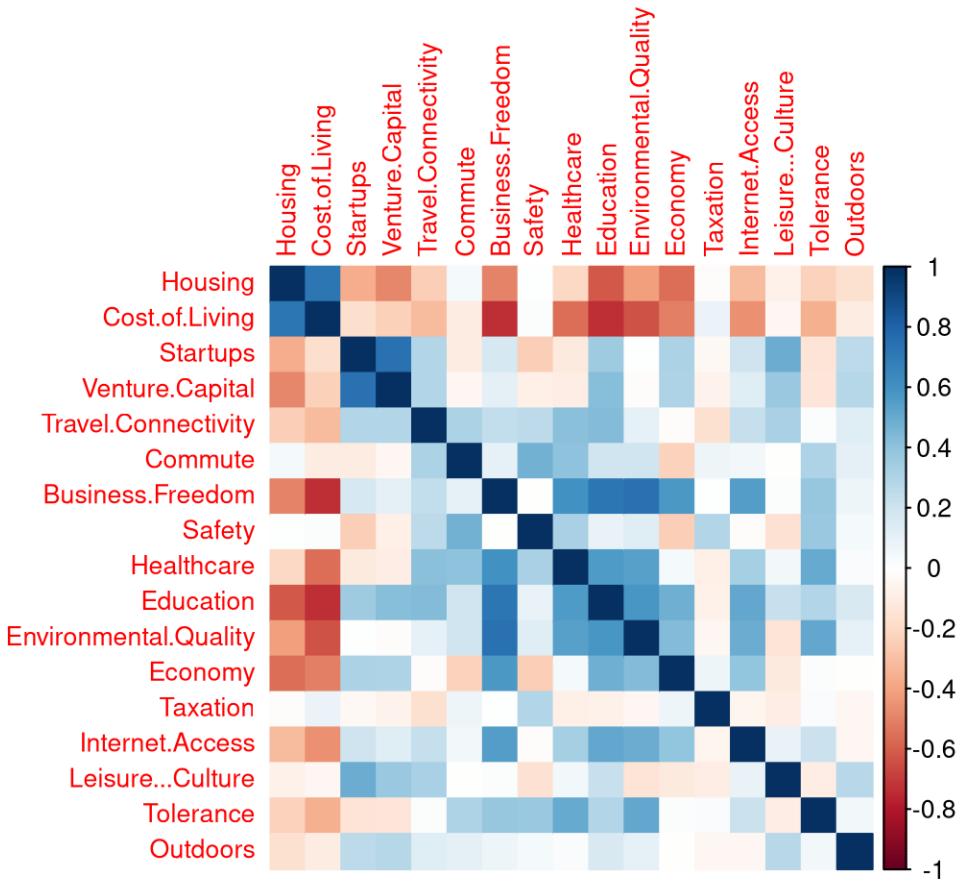
Oceania — bardzo mała liczba punktów, ale zróżnicowana głównie wzdułóż PC2.



Wykres 18: Podpisane obserwacje w celu identyfikacji miast najmocniej odstających.

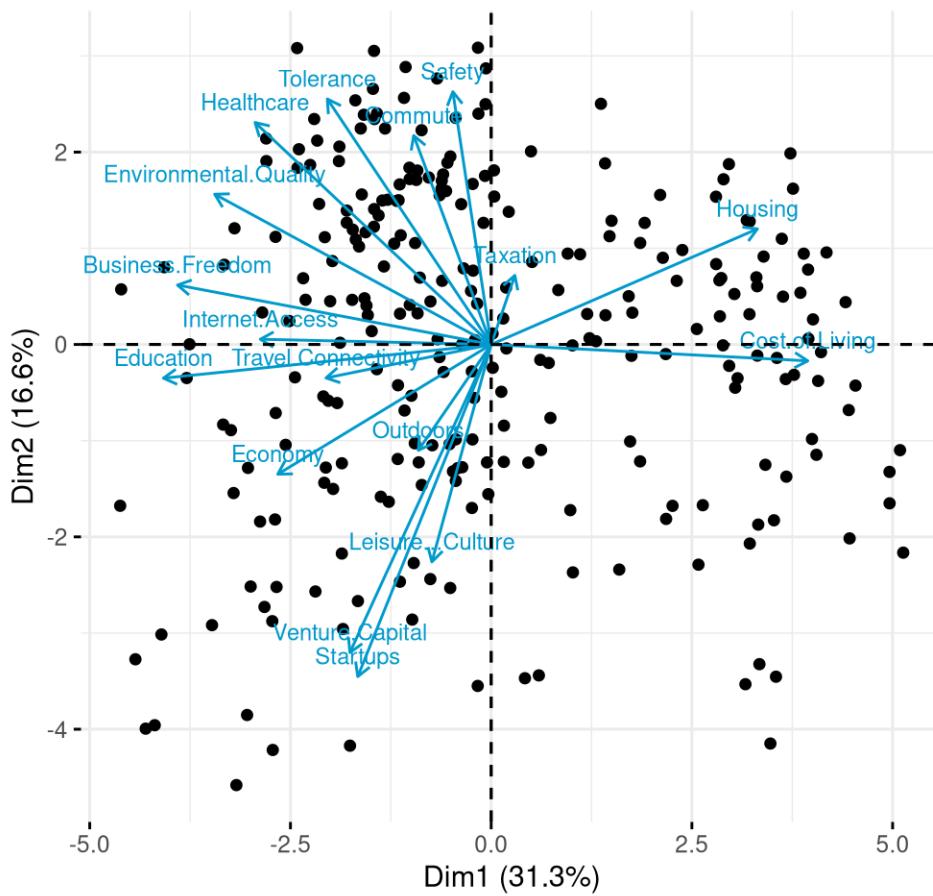
Z wykresu 18 odczytujemy, że najbardziej odstaje Los Angeles, New York i San Francisco Bay Area. Prawdopodobnie dlatego, że są to świetne miasta dla biznesu, inwestycji i startupów, ale za to nie są bezpieczne. Z drugiej strony odstaje również Delhi, bo w tym mieście koszty życia są niskie i przez to jest bliżej miast afrykańskich, ale jest nim dobrze rozwinięte Startups i Venture.capital, przez co wyróżnia się na tle miast z prawej strony wykresu.

2.6 Korelacje zmiennych i biplot



Wykres 19: Macierz korelacji i biplot dla wszystkich zmiennych ilościowych

PCA - Biplot



Wykres 20: Macierz korelacji i biplot dla wszystkich zmiennych ilościowych

Występuje dodatnia korelacja między:

- Venture.capital i Startups,
- Housing i Cost.of.Living,
- Environmental.Quality i Business.Freedom,
- Education i Business.Freedom
- Tolerance i Environmental.Quality.

Występuje ujemna korelacja między:

- Business.Freedom i Cost.of.Living,
- Education i Cost.of.Living,
- Education i Housing.

2.7 Wnioski końcowe

Ciekawe obserwacje:

- PC1 rozróżnia miasta pod względem taniego życia i mieszkań (prawa strona) vs. jakości edukacji, środowiska, zdrowia (lewa strona).

- PC2 oddziela **miasta startupowe i kulturalne (dół)** od **miast bezpiecznych i dobrze skomunikowanych (góra)**.
- Miasta takie jak **Los Angeles, San Francisco, New York** odstają na dole wykresu, co wynika z ich wysokich wartości w *Startups, Venture Capital* i *Leisure & Culture*.
- **Delhi i Mumbaj** odstają w prawy dół wykresu — czyli są stosunkowo tanie, ale mają cechy rozwijających się metropolii z dużą aktywnością gospodarczą.
- **Zurich** jest silnie przesunięty w lewo — to bardzo drogie, ale wysokiej jakości miasto.
- **Lagos** to miasto najbardziej wysunięte w prawo, czyli tanie koszty życia, ale bardzo niski poziom udogodnień.
- Kontynenty są rozróżnialne — Europa i Ameryka Północna podobne, Azja rozproszona, Afryka i Ameryka Południowa zbliżone.

Do zadowalającej wizualizacji danych wystarczają nam dwie (około 50% wariancji) lub trzy (około 60%) składowe, jednakże zdecydowałem się wybrać dwie ponieważ wykresy 2d są czytelniejsze.

Zastosowanie standaryzacji miało istotny wpływ na wyniki analizy PCA. Zmiennych użytych w analizie nie można było bezpośrednio porównywać, ponieważ miały one różne rozrzuty. Brak standaryzacji skutkowałby tym, że zmienne o największej wariancji zdominowałyby pierwsze składowe główne, co zniekształciłoby strukturę danych.

3 Zadanie 3

3.1 Przygotowanie danych

Celem zadania jest przeprowadzenie skalowania wielowymiarowego (MDS) na zbiorze `titanic_train` w celu znalezienia potencjalnego klarownego podziału danych na skupiska i znalezienia cech, które będą wyróżniać owe grupy.

Tabela 7: Struktura zbioru danych `titanic train`

Zmienna	Opis	Typ	Przykładowe_wartości
PassengerId	Identyfikator pasażera	integer	1, 2, 3
Survived	Czy pasażer przeżył katastrofę	integer	0, 1, 1
Pclass	Klasa biletu	integer	3, 1, 3
Name	Nazwisko pasażera	character	Braund, Mr. Owen Harris, Cumings, Mrs. John Bradley (Florence Briggs Thayer), Heikkinen, Miss. Laina
Sex	Płeć pasażera	character	male, female, female
Age	Wiek pasażera	numeric	22, 38, 26
SibSp	Czy na pokładzie było rodzeństwo lub współmałżonek (orzaz liczba)	integer	1, 1, 0
Parch	Czy na pokładzie były dzieci lub rodzice pasażera (orzaz liczba)	integer	0, 0, 0
Ticket	Numer biletu	character	A/5 21171, PC 17599, STON/O2. 3101282
Fare	Oplata pasażera	numeric	7.25, 71.2833, 7.925
Cabin	Numer kabiny	character	, C85,
Embarked	Port zaokrętowania	character	S, C, S

Zbiór `titanic_train` zawiera **891** obserwacji oraz **12** cech.

Tabela 8: Wartości brakujące w zbiorze Titanic

zmienna	liczba NA
PassengerId	0
Survived	0
Pclass	0
Name	0
Sex	0
Age	177
SibSp	0
Parch	0
Ticket	0
Fare	0
Cabin	0
Embarked	0

Na podstawie tabeli 8 zawiera **177** wartości brakujących. W oryginalnym zbiorze typy zmiennych `Survived`, `Pclass`, `Sex` oraz `Embarked` zostały nieprawidłowo wczytane. Zamieniamy je na zmienne czynnikowe. Dodatkowo wartości brakujące w zmiennej `Age` uzupełniamy poprzez zastosowanie metody k-sąsiadów. Ponadto, cechy takie jak `PassengerId`, `Name`, `Ticket` otaz `Cabin` są identyfikatorami pasażerów, nie są potrzebne w dalszej analizie, dlatego możemy je pominać.

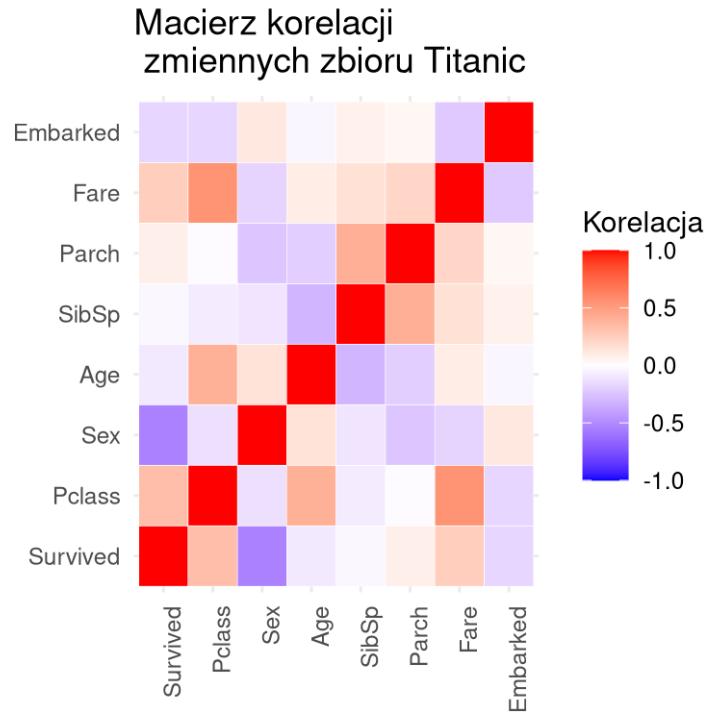
```

Titanic$Survived <- as.factor(Titanic$Survived)
Titanic$Pclass <- as.ordered(Titanic$Pclass)
Titanic$Pclass <- factor(Titanic$Pclass, levels = c("3",
    "2", "1"), ordered = TRUE)
Titanic$Sex <- as.factor(Titanic$Sex)
Titanic$Embarked <- as.factor(Titanic$Embarked)

Titanic_clean <- dplyr::select(Titanic, -PassengerId,
    -Name, -Ticket, -Cabin)
# Uzupełniamy wartości brakujące metodą
# k-sąsiadów
Titanic_clean <- kNN(Titanic_clean, k = 5, imp_var = FALSE)

```

3.2 Wstępna analiza danych



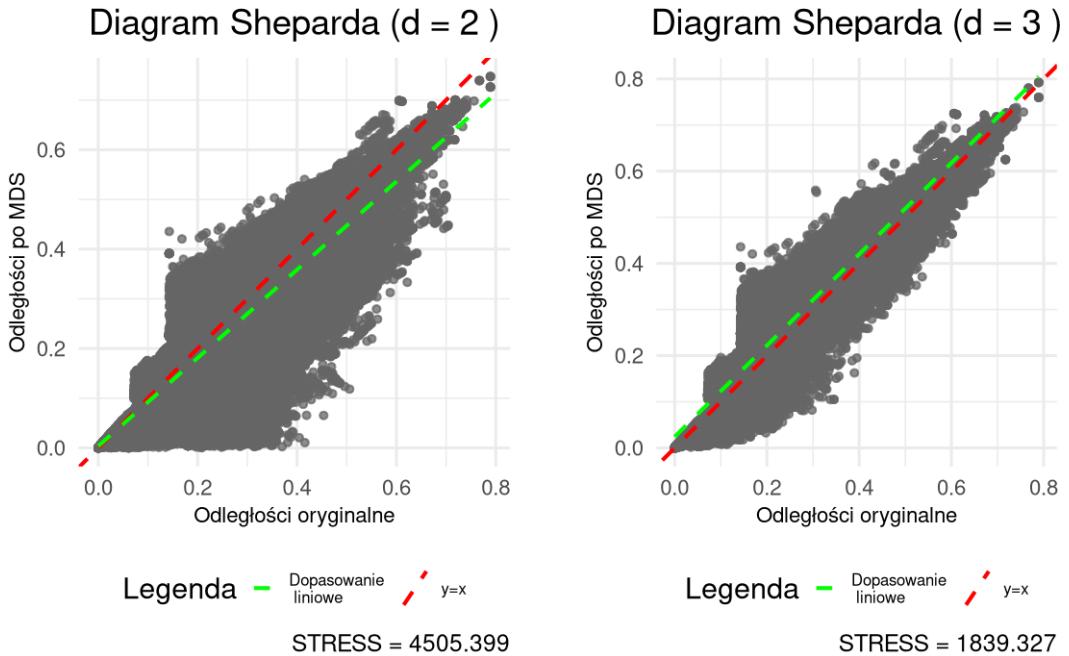
Wykres 21: Macierz korelacji wybranych cech zbioru titanic train

Wykres 21 przedstawia macierz korelacji wybranych zmiennych zbioru `titanic_train`. Możemy zaobserwować następujące zależności:

- Silna dodatnia korelacja zmiennych `Fare` i `Pclass` - im lepsza klasa, tym wyższa cena biletu.
- Dodatnia korelacja między `Survived` i `Fare` oraz `Survived` i `Pclass` - osoby które zapłaciły więcej za bilet i miały lepszą klasę, miały większe szanse na przeżycie katastrofy.
- Silna ujemna korelacja między `Survived` i `Sex` - silny związek między płcią, a szansą na przeżycie.
- Dodatnia korelacja między `Age` a `Pclass` - osoby starsze miały bilety z lepszą klasą.

3.3 Redukcja wymiaru na bazie MDS

Przeprowadzamy redukcję wymiarów metodą skalowania wielowymiarowego, poprzez wyznaczenie macierzy niepodobieństwa/odmiенноści i korzystamy z podstawowych parametrów funkcji `cmdscale()` sprowadzając dane do 2 oraz do 3 wymiarów.



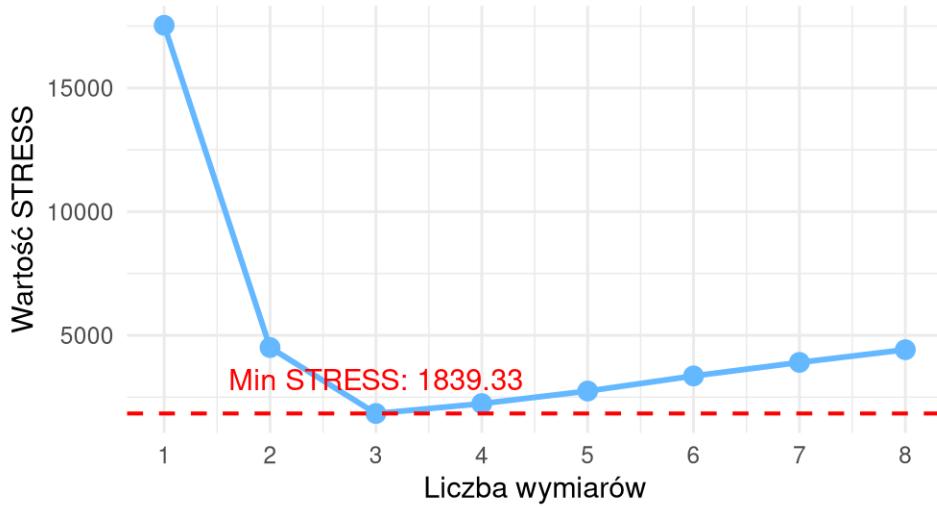
Wykres 22: Wykresy Sheparda zbioru titanic train dla wybranych wymiarów

Wykres 22 przedstawia rozrzułt odległości po zastosowaniu MDS (sprowadzenie do 2 oraz 3 wymiarów) względem oryginalnych odległości. Czerwona linia przedstawia prostą $y = x$, natomiast zielona regresję liniową. Możemy zauważyć, że w przypadku 3 wymiarów, proste niemal się pokrywają, a punkty są bardziej skupione wokół prostej. Oznacza to, że odległości po zastosowaniu MDS dla 3 wymiarów są zachowane znacznie lepiej, niż te uzyskane dla 2 wymiarów. Możemy zauważyć, że w przypadku sprowadzenia danych do 3 wymiarów, wartość kryterium STRESS, którą definiujemy jako:

$$S(z_1, z_2, \dots, z_n) = \sum_{i=1}^n \sum_{j=1}^n a_{ij} (d_{ij} - \|z_i - z_j\|)^2, \text{ gdzie}$$

a_{ij} - wagi (wartość zależna od algorytmu), d_{ij} - elementy macierzy odmiенноścii $D = [d_{ij}]_{i,j=1,\dots,n}$, $z_1, \dots, z_n \in \mathbb{R}^d$, jest znacznie niższa niż dla dwóch wymiarów. Potwierdza to wykres 23 na którym widoczna jest zależność wartości funkcji STRESS od liczby wymiarów. Zauważać można, że wartość tego kryterium jest najniższa (**1839.33**) dla właśnie trzech wymiarów (na wykresie punkt "łokciowy" znajduje się właśnie w tym miejscu). Oznacza to, że dla tej przestrzeni, odległości są najlepiej zachowane.

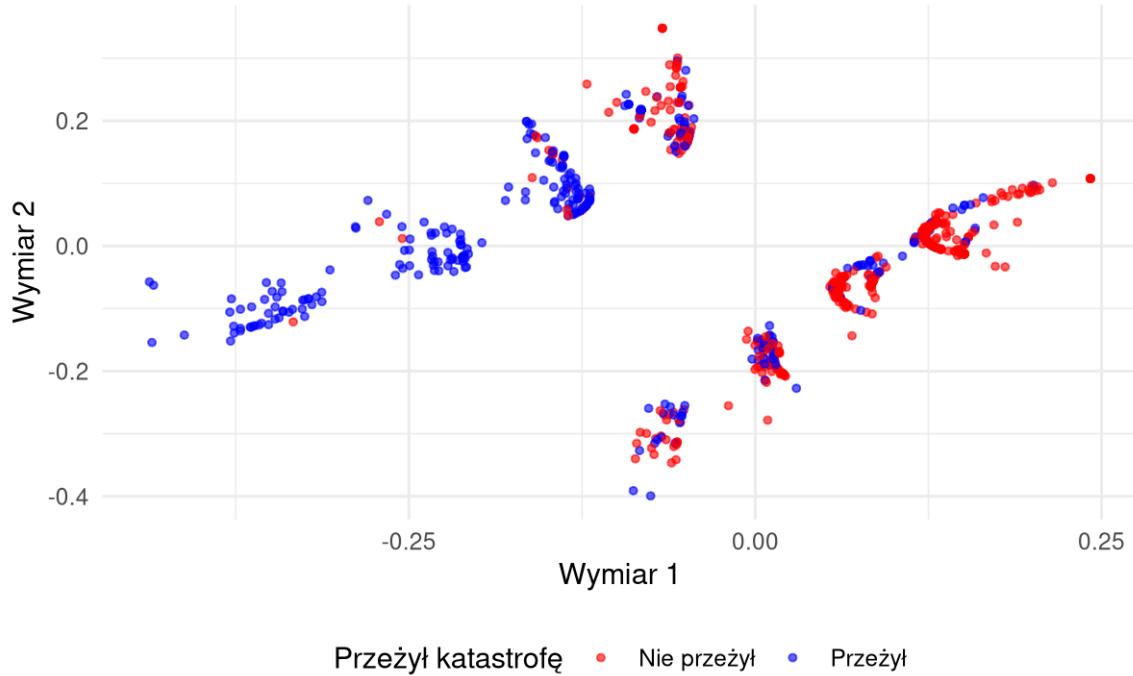
Funkcja STRESS w zależności od liczby wymiarów



Wykres 23: Wykres wartości zmiennej STRESS w zależności od liczby wymiarów

3.4 Wizualizacja danych

Wyniki MDS dla danych Titanic



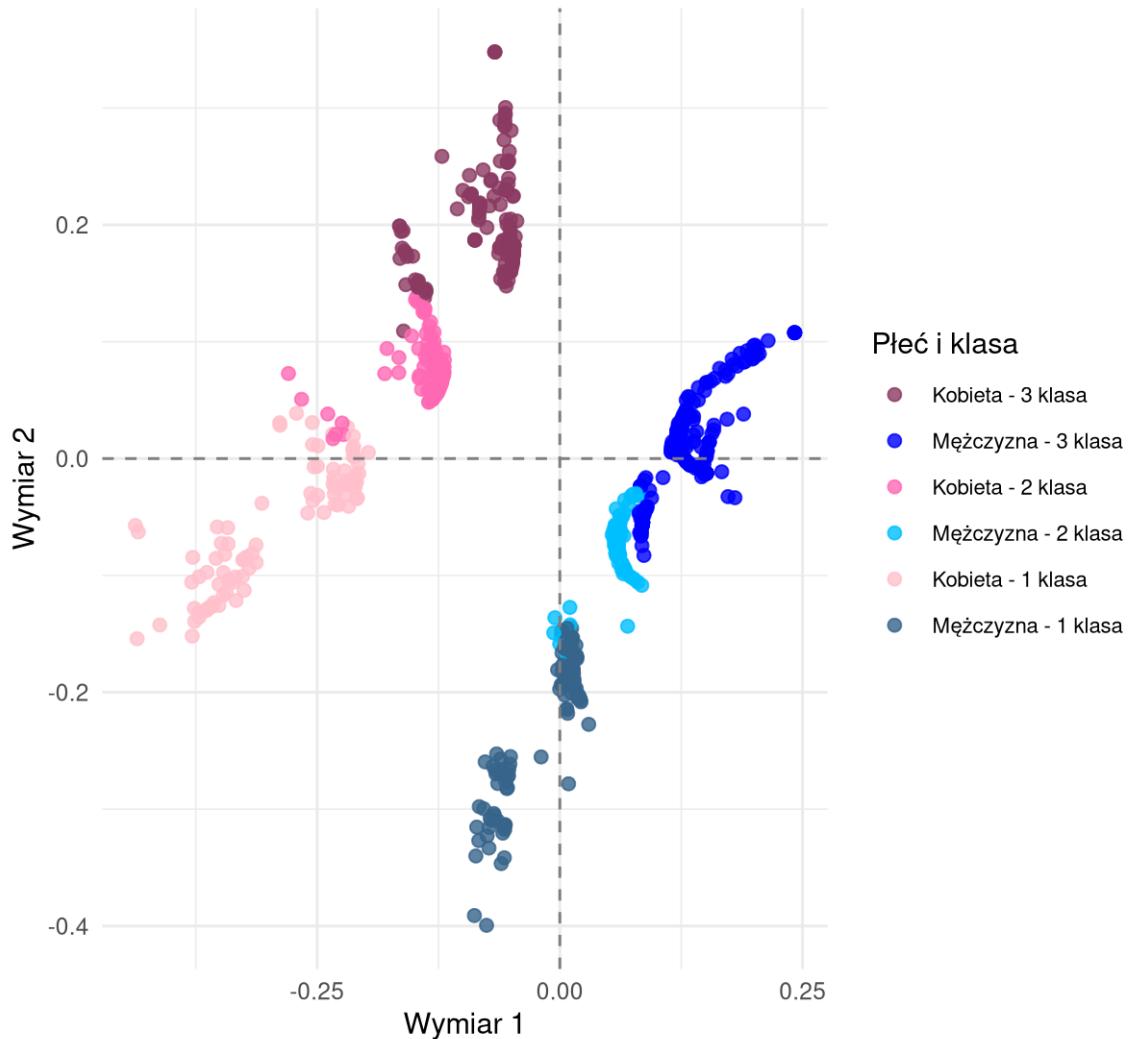
Wykres 24: Dwuwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Survived)

Wykres 24 przedstawia rozmieszczenia danych `titanic_train` po sprowadzeniu danych, poprzez MDS, do dwóch wymiarów. Możemy zauważyc podział na dwie główne grupy. Natomiast każda z tych grup zawiera

podział na 4 podgrupy. Dane są rozłożone w dość symetryczny sposób względem prostej $y = x$. W grupie “powyżej” tej prostej znaczna większość osób przeżyła katastrofę, natomiast w drugim klastrze większość zginęła. Na obrzeżach głównych klastrówauważalne są pewne obserwacje odstające. W III ćwiartce układu współrzędnych, obserwacjami odstającymi są te osoby, które przeżyły, natomiast w I i II, wszystkie obserwacje odstające to osoby, które zginęły.

Wyniki MDS 2D dla danych Titanic

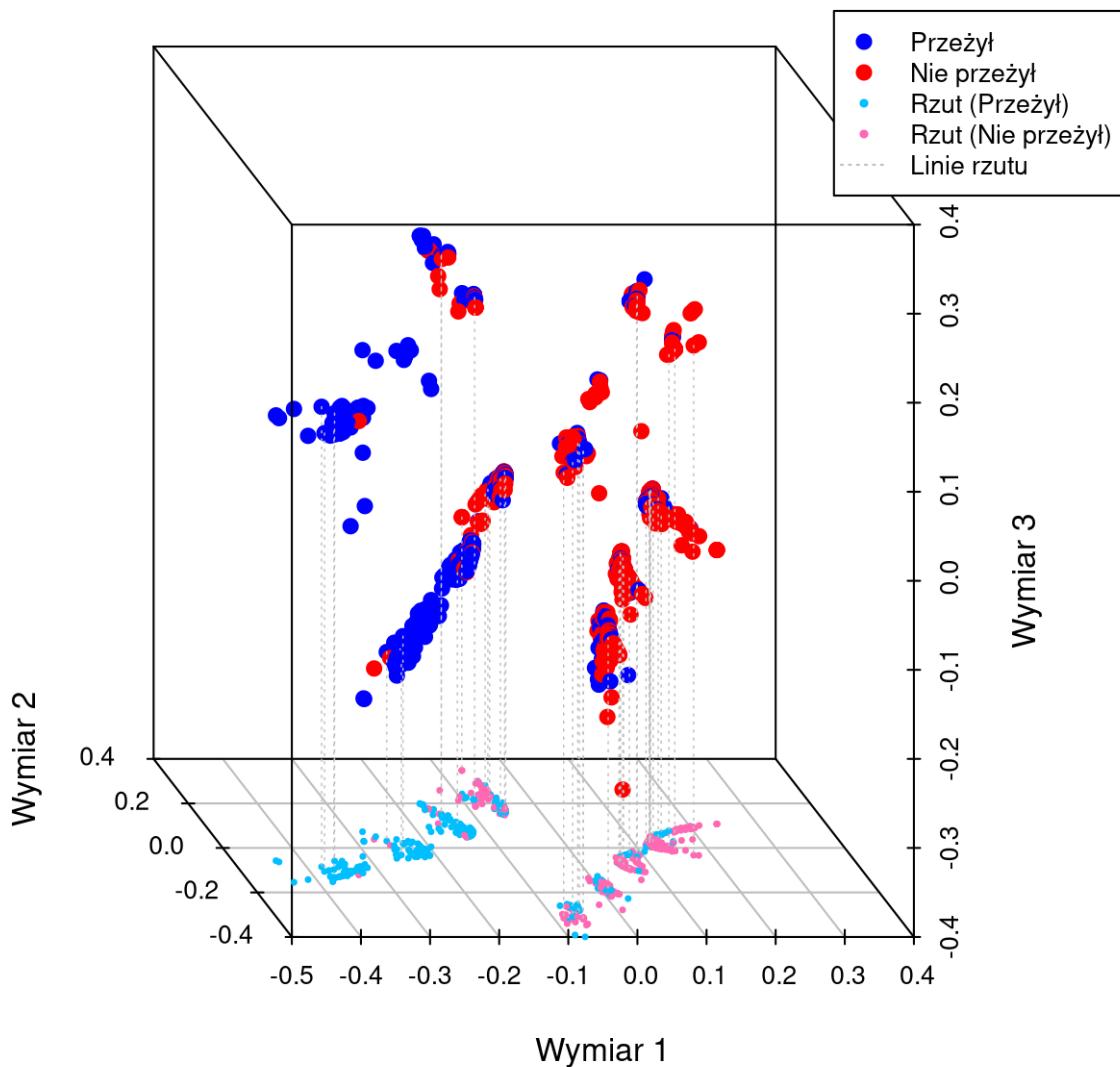
Z podziałem na płeć, klasę



Wykres 25: Dwuwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Sex i Pclass)

Wykres 25 przedstawia podział pasażerów ze względu na płeć i klasę biletu. Możemy zauważyć, że główne klastry całkowicie separują płeć. Natomiast podgrupy są wyraźnie podzielone względem klasy. Dodatkowo, odnosząc się do wykresu 24, odczytujemy, że osobami, które przeżyły katastrofę, były głównie kobiety. Również klasa biletu miała wpływ na przeżycie (osoby, które były w wyższej klasie, miały większą szansę przeżycia).

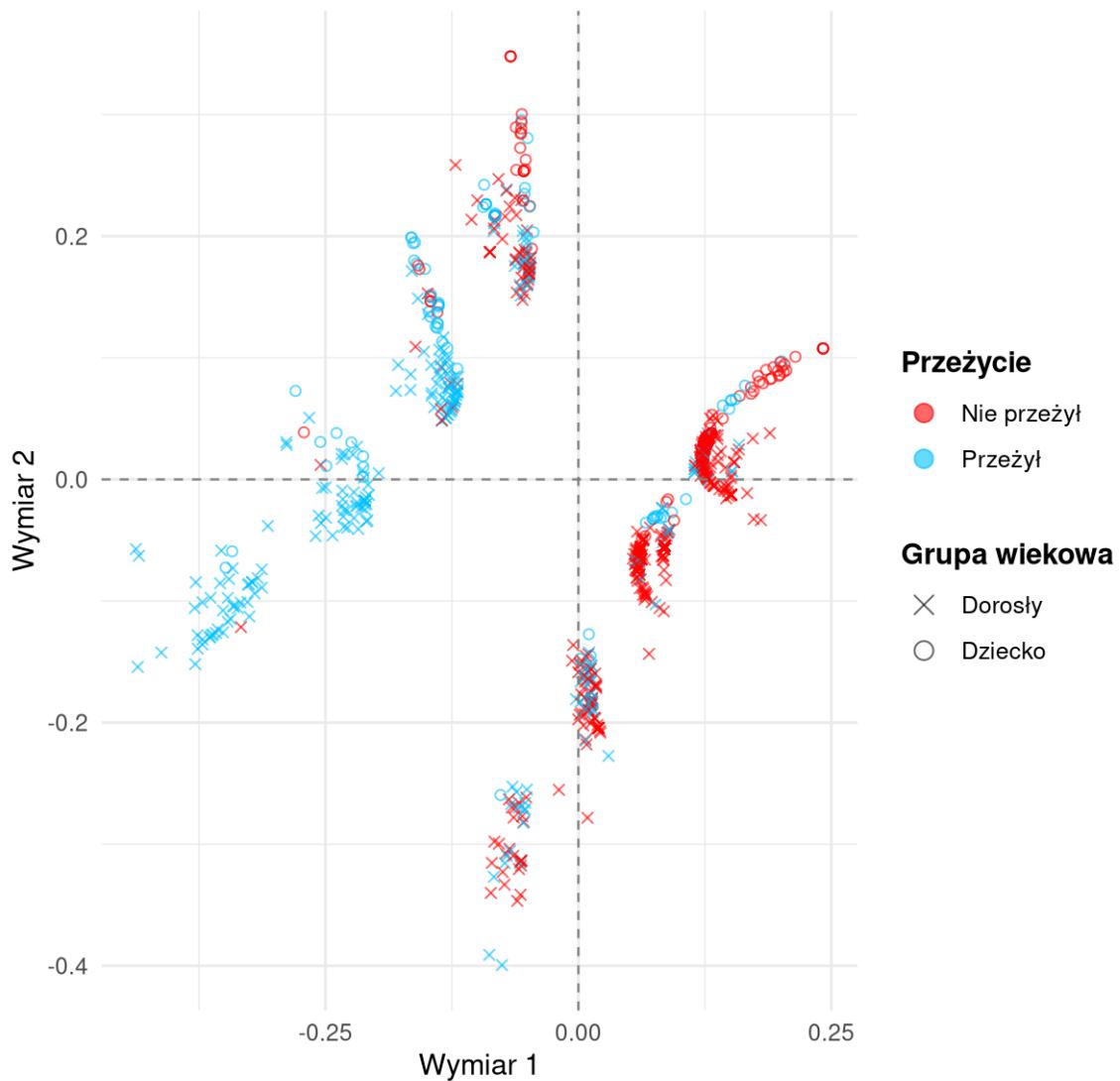
MDS 3D Titanic z rzutowaniem na podstawę



Wykres 26: Trójwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Survived)

Wykres 26 pokazuje rozmieszczenie obserwacji w trzech wymiarach. Interesującym spostrzeżeniem jest, że mężczyźni i kobiety z klasy pierwszej są umieszczeni na tej samej wysokości względem osi Wymiaru 3.

Wyniki MDS 2D dla danych Titanic



Wykres 27: Dwuwymiarowa reprezentacja MDS danych Titanic (grupowanie względem zmiennej Age i Survived)

Na podstawie wykresu 27 zauważamy, że obserwacje odstające, które są w III ćwiartce, to dorosli (mężczyźni i kobiety) z pierwszej klasy. Natomiast pozostałe, wcześniej wspomniane obserwacje odstające, to dzieci z klasy trzeciej. Dodatkowo każda z ośmiu podgrup ma dość wyraźne rozgraniczenie pomiędzy dorosłych a dziećmi. W przypadku kobiet z klasy pierwszej i drugiej, grupa wiekowa nie miała wpływu na przeżycie, co może sugerować, że dzieci ewakuowały się wraz z matkami. Inna sytuacja ma miejsce w pozostałych skupiskach, nie można zbytnio stwierdzić czy wiek miał wpływ na przeżycie katastrofy.

3.5 Podsumowanie

- Zbiór danych `titanic.train` po zastosowaniu MDS został podzielony na dwie główne grupy (kobiety i mężczyźni). Równie dobrze zostały odseparowane obserwacje ze względu na klasę biletu.
- W katastrofie zginęło zdecydowanie więcej mężczyzn niż kobiet. Wyjątkiem są nieliczni mężczyźni z klasy 1, którzy przeżyli (prawdopodobnie członkowie załogi lub wpływowe osoby).

- Skalowanie wielowymiarowe najlepiej odwzorowuje oryginalne odległości dla 3 wymiarów.
- Największy procent ofiar w katastrofie znajdowało się w 3 klasie.
- Czynniki społeczne (płeć, klasa) były głównymi wyznacznikami przeżycia
- Zmienne użyte w analizie (wiek, klasa, płeć, itp.) bardzo dobrze determinują przeżycie, jednak nie w pełni, co może wynikać z czynników losowych lub innych niewziętych pod uwagę zmiennych (np. miejsce na statku w momencie kolizji).