

Raport - Dyskretyzacja cech ciągłych oraz redukacja wymiaru metodami PCA i MDS

Filip Michewicz 282239
Wiktor Niedźwiedzki 258882

30 kwietnia 2025 Anno Domini

Spis treści

1 Wprowadzenie	3
2 Zadanie 1: Dyskretyzacja (przedziałowanie) cech ciągłych	5
2.1 Własności danych	5
2.2 Korelacje między zmiennymi	5
2.3 Wykresy skrzypcowe	7
2.4 Porównanie nienadzorowanych metod dyskretyzacji	8
2.5 Dyskretyzacja za pomocą własnych przedziałów	16
2.6 * Stworzenie modelu predykcyjnego	16
2.7 Podsumowanie	19
3 Zadanie 2: Analiza składowych głównych (Principal Component Analysis (PCA))	20
3.1 Własności danych	20
3.2 Przygotowanie danych do analizy	21
3.3 Wyznaczanie składowych głównych	23
3.4 Korelacja zmiennych	25
3.5 Zmienność odpowiadająca poszczególnym składowym	26
3.6 Wizualizacja danych wielowymiarowych	30
3.7 Podsumowanie	36
3.8 Potrzeba standaryzacji	38
4 Zadanie 3: Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))	40
4.1 Przygotowanie danych	40
4.2 Redukcja wymiaru na bazie MDS	41
4.3 Wizualizacja danych	43
4.3.1 * Rodzina Sage – symbol nierówności klasowej na Titaniku	47
4.4 Podsumowanie	47

Spis wykresów

1	Macierz korelacji zmiennych numerycznych - irys	6
2	Macierz korelacji zmiennych numerycznych - irys - podział ze względu na gatunek	7
3	Wykresy skrzypcowe dla zmiennych numerycznych a) Sepal.Length b) Sepal.Width c) Petal.Length d) Petal.Width	9
4	Porównanie długości i szerokości płatków gatunków versicolor oraz virginica	10
5	Metody dyskretyzacji dla zmiennej Sepal.Length	12
6	Metody dyskretyzacji dla zmiennej Sepal.Width	13
7	Metody dyskretyzacji dla zmiennej Petal.Length	14
8	Metody dyskretyzacji dla zmiennej Petal.Width	15
9	Dyskretyzacja zmiennej Petal.Width - własne przedziały	17
10	Wykres pudełkowy zmiennych numerycznych (przed standaryzacją) - City Quality of Life . .	22
11	Wykres pudełkowy zmiennych numerycznych (po standaryzacji) - City Quality of Life . . .	23
12	Wykres pudełkowy wektorów ładunków składowych	24
13	Dwuwykres dla dwóch pierwszych składowych	27
14	Macierz korelacji zmiennych numerycznych	28
15	Skumulowana składowa wariancja	29
16	Wykres rozrzutu składowych głównych; Afryka	32
17	Wykres rozrzutu składowych głównych; Azja	33
18	Wykres rozrzutu składowych głównych; Azja - kraje	33
19	Wykres rozrzutu składowych głównych; Europa	34
20	Wykres rozrzutu składowych głównych; Ameryka Północna	35
21	Wykres rozrzutu składowych głównych; Australia i Oceania - kraje	36
22	Wykres rozrzutu składowych głównych; Ameryka Południowa	37
23	Dwuwykres dla dwóch pierwszych składowych - dane niestandaryzowane	39
24	Diagram Sheparda dla danych pasażerów Titanica po redukcji wymiarów do przestrzeni dwuwymiarowej metodą MDS	42
25	MDS - Rozmieszczenie pasażerów Titanica w przestrzeni dwuwymiarowej	43
26	MDS - Rozmieszczenie pasażerów Titanica w przestrzeni dwuwymiarowej z podziałem na a) Klasę b) Płeć c) Port zaokrętowania d) Grupę wiekową	45
27	MDS - Rozmieszczenie pasażerów Titanica w przestrzeni dwuwymiarowej z podziałem na status przeżycia	46

Spis tabel

1	Opis zmiennych w zbiorze danych iris	5
2	Porównanie skuteczności metod dyskretyzacji w zależności od zmiennej	16
3	Porównanie skuteczności metod dyskretyzacji	18
4	Opis zmiennych w zbiorze danych City Quality of Life	21
5	Wektory i odpowiadające im wartości dla PC1, PC2 i PC3	25
6	Opis zmiennych w zbiorze danych titanic_train	41

```
library(corrplot)
```

```
## Warning: pakiet 'corrplot' został zbudowany w wersji R 4.4.3
```

```
## corrplot 0.95 loaded
```

```
library(ggplot2)
```

```
## Warning: pakiet 'ggplot2' został zbudowany w wersji R 4.4.3
```

```
library(gridExtra)
library(e1071)
```

```
## Warning: pakiet 'e1071' został zbudowany w wersji R 4.4.3
```

```
library(xtable)
```

```
## Warning: pakiet 'xtable' został zbudowany w wersji R 4.4.2
```

```
library(knitr)
```

```
## Warning: pakiet 'knitr' został zbudowany w wersji R 4.4.3
```

```
library(plyr)
```

```
## Warning: pakiet 'plyr' został zbudowany w wersji R 4.4.3
```

```
library(dplyr)
```

```
##
```

```
## Dołączanie pakietu: 'dplyr'
```

```
## Następujące obiekty zostały zakryte z 'package:plyr':
```

```
##
```

```
##      arrange, count, desc, failwith, id, mutate, rename, summarise,
##      summarize
```

```

## Następujący obiekt został zakryty z 'package:gridExtra':
##
##      combine

## Następujące obiekty zostały zakryte z 'package:stats':
##
##      filter, lag

## Następujące obiekty zostały zakryte z 'package:base':
##
##      intersect, setdiff, setequal, union

library(GGally)

## Warning: pakiet 'GGally' został zbudowany w wersji R 4.4.2

## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2

library(MASS)

## Warning: pakiet 'MASS' został zbudowany w wersji R 4.4.3

##
## Dołączanie pakietu: 'MASS'

## Następujący obiekt został zakryty z 'package:dplyr':
##
##      select

library(arules)

## Warning: pakiet 'arules' został zbudowany w wersji R 4.4.3

## Ładowanie wymaganego pakietu: Matrix

##
## Dołączanie pakietu: 'arules'

## Następujący obiekt został zakryty z 'package:dplyr':
##
##      recode

## Następujące obiekty zostały zakryte z 'package:base':
##
##      abbreviate, write

```

```
library(cowplot)

## Warning: pakiet 'cowplot' został zbudowany w wersji R 4.4.2

library(scales)

## Warning: pakiet 'scales' został zbudowany w wersji R 4.4.3

library(factoextra)

## Warning: pakiet 'factoextra' został zbudowany w wersji R 4.4.3

## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

library(titanic)

## Warning: pakiet 'titanic' został zbudowany w wersji R 4.4.3

library(cluster)

data(iris)

data <- read.csv("uaScoresDataFrame.csv")

data("titanic_train")
```

1 Wprowadzenie

Celem niniejszego raportu jest przedstawienie wyników eksploracyjnej analizy danych przeprowadzonej w ramach trzech odrębnych zadań. Każde z nich koncentruje się na innej metodzie przetwarzania danych wielowymiarowych.

Zadanie 1 dotyczy dyskretyzacji cech ciągłych w zbiorze `iris`. Analiza obejmuje wybór cech o różnej zdolności dyskryminacyjnej oraz porównanie metod dyskretyzacji nienadzorowanej (równa szerokość, równa częstotliwość, metoda k-średnich, ręczne przedziały).

Zadanie 2 obejmuje zastosowanie analizy składowych głównych (PCA) na danych dotyczących jakości życia w miastach (`uaScoresDataFrame.csv`). Wykonano redukcję wymiarowości, oceniono zmienność głównych komponentów oraz przeanalizowano wkład poszczególnych zmiennych.

Zadanie 3 poświęcone jest skalowaniu wielowymiarowemu (MDS) w zbiorze `titanic_train`. Celem jest odwzorowanie danych w przestrzeni 2D, umożliwiające ocenę podobieństw i struktur skupisk wśród pasażerów Titanica, z uwzględnieniem zmiennej `Survived` oraz innych cech takich jak płeć, klasa, port zaokręgowania i grupa wiekowa.

2 Zadanie 1: Dyskretyzacja (przedziałowanie) cech ciągłych

W tym zadaniu przeprowadzona zostanie analiza zbioru danych *iris* z pakietu *datasets* w R i zawiera wyniki pomiarów dotyczących trzech gatunków irysów:

- **Setosa**
- **Versicolor**
- **Virginica**

Dane zostały udostępnione przez znanego brytyjskiego statystyka, Ronaldiego Fishera, w 1936 roku i od tego czasu stały się klasycznym przykładem w wielu analizach statystycznych oraz w kontekście algorytmów klasyfikacyjnych.

2.1 Własności danych

Zbiór danych *Iris* zawiera **150** przypadków, po 50 dla każdego z trzech gatunków oraz **5** cech. Liczba brakujących danych wynosi **0**.

Znaczenie poszczególnych cech oraz ich typ przedstawiono w tabeli 1.

```
df_table <- data.frame(  
  Typ = sapply(iris, class),  
  Opis = c("Długość działki kielicha (cm)",  
         "Szerokość działki kielicha (cm)",  
         "Długość płatka (cm)",  
         "Szerokość płatka (cm)",  
         "Gatunek (setosa, versicolor, virginica)")  
)  
  
kable(df_table, col.names = c("Zmienna", "Typ", "Opis"),  
      caption = "Opis zmiennych w zbiorze danych iris")
```

Tabela 1: Opis zmiennych w zbiorze danych iris

Zmienna	Typ	Opis
Sepal.Length	numeric	Długość działki kielicha (cm)
Sepal.Width	numeric	Szerokość działki kielicha (cm)
Petal.Length	numeric	Długość płatka (cm)
Petal.Width	numeric	Szerokość płatka (cm)
Species	factor	Gatunek (setosa, versicolor, virginica)

Wszystkie cechy dotyczące kwiatów są miarami ciągłymi i różnią się w zależności od gatunku irysa, który jest zmienną jakościową. Celem analizy będzie zrozumienie, w jaki sposób te cechy mogą pomóc w klasyfikacji irysów do odpowiednich gatunków oraz zastosowanie różnych technik dyskretyzacji i analizy, aby lepiej zrozumieć właściwości tych danych.

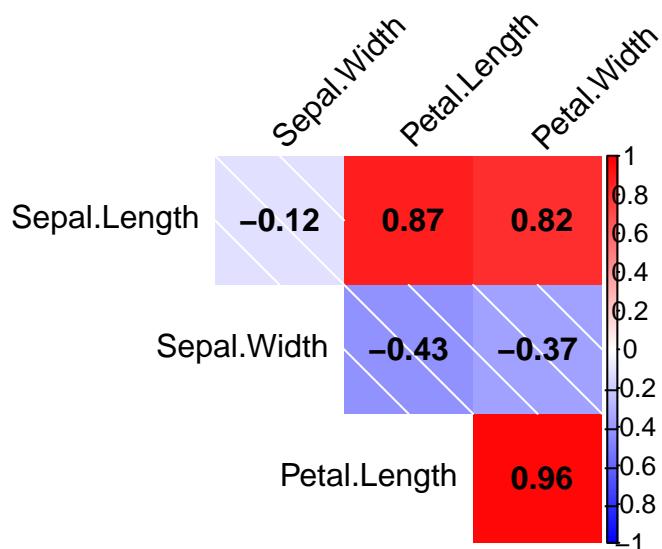
2.2 Korelacje między zmiennymi

W tej części sprawdzimy, czy istnieją korelacje między zmiennymi – zarówno w całym zbiorze danych, jak i z uwzględnieniem podziału na gatunki.

```

macierz <- cor(iris[, -5])
corrplot(macierz, method = "shade", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black",
         diag = FALSE,
         number.cex=1)

```



Wykres 1: Macierz korelacji zmiennych numerycznych - irys

Na podstawie macierzy korelacji z wykresu 1. można zauważać następujące zależności między cechami badanych kwiatów: - wzrost długości działki kielicha wiąże się ze wzrostem długości oraz szerokości płatków, - wzrost szerokości działki kielicha umiarkowanie koreluje ujemnie z długością i szerokością płatków, - bardzo silna dodatnia korelacja występuje między długością a szerokością płatków.

Sprawdzamy teraz, jak te korelacje wyglądają, gdy uwzględnimy podział na gatunki.

```

iris.setosa <- subset(iris, Species == "setosa")[, 1:4]
iris.versicolor <- subset(iris, Species == "versicolor")[, 1:4]
iris.virginica <- subset(iris, Species == "virginica")[, 1:4]

cor.setosa <- cor(iris.setosa)
cor.versicolor <- cor(iris.versicolor)
cor.virginica <- cor(iris.virginica)

par(mfrow = c(1, 3))

# Wykresy korelacji

```

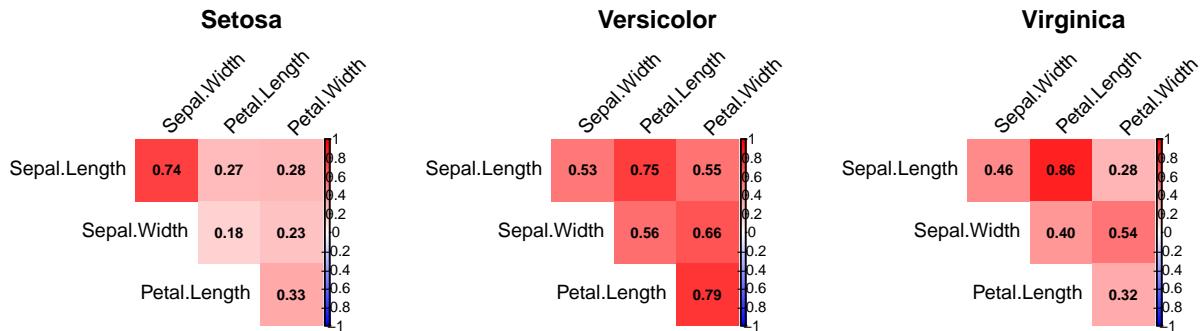
```

corrplot(cor_setosa, method = "shade", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", diag = FALSE, number.cex = 0.7)
title("Setosa")

corrplot(cor_versicolor, method = "shade", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", diag = FALSE, number.cex = 0.7)
title("Versicolor")

corrplot(cor_virginica, method = "shade", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black", diag = FALSE, number.cex = 0.7)
title("Virginica")

```



Wykres 2: Macierz korelacji zmiennych numerycznych - irys - podział ze względu na gatunek

Na podstawie macierzy korelacji z wykresu 2. można zauważyć, że:

- dla *setosa* występuje silna dodatnia korelacja między `sepal.length` a `sepal.width`,
- dla *versicolor* wszystkie zmienne są ze sobą umiarkowanie lub silnie skorelowane ($r > 0.5$), a najsilniejsze korelacje ($r > 0.75$) występują między `petal.length` a `sepal.length` oraz `petal.length` a `petal.width`,
- dla *virginica* największą korelację ($r = 0.86$) obserwuje się między `petal.length` a `sepal.length`.

Podział na gatunki pozwala na dokładniejsze uchwycenie specyficznych wzorców korelacji między cechami kwiatów. W ogólnej macierzy korelacji (bez podziału na gatunki) widać ogólnie tendencje, jednak dla poszczególnych gatunków korelacje różnią się, co wskazuje na większą jednorodność cech w obrębie niektórych gatunków. Podział ten pozwala na lepsze zrozumienie specyfiki zależności między cechami.

2.3 Wykresy skrzypcowe

W tej części tworzymy wykresy skrzypcowe dla poszczególnych zmiennych numerycznych, aby zobrazować ich rozkład w podziale na gatunki.

```

zmienne <- colnames(iris)[-5]

wykresy <- list()
for (i in seq_along(zmienne)) {
  wykres <- ggplot(iris, aes_string(x = "Species", y = zmienne[i],
                                    fill = "Species")) +
    geom_violin() +
    labs(title = paste0(letters[i], " ", zmienne[i]),
         x = "Gatunek",
         y = paste(zmienne[i], "(cm)")) +
    theme_minimal() +
    theme(legend.position = "none")

  wykresy <- c(wykresy, list(wykres))
}

## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## i Please use tidy evaluation idioms with `aes()`.
## i See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

grid.arrange(grobs = wykresy, ncol = 2)

```

Powysze wykresy wskazują, że spośród wszystkich zmiennych numerycznych to cechy płatków – a szczególnie ich długość i szerokość – są najbardziej przydatne do rozróżniania gatunków irysów. Widzimy, że *setosa* jest jednoznacznie odróżniała na podstawie tych cech, podczas gdy działki kielicha są znacznie mniej unikalne. W kontekście dalszej analizy, np. dyskretyzacji czy klasyfikacji, pojawia się naturalne pytanie: która z tych dwóch zmiennych – długość czy szerokość płatków – lepiej pozwala rozróżnić pozostałe dwa gatunki, *versicolor* i *virginica*?

```

ggpairs(filter(iris[, c(3, 4, 5)], Species != "setosa"),
        aes(colour=Species, alpha=0.5),
        columns=1:2)

```

Z wykresów można odczytać, że mniejsze „pokrycie” mamy dla zmiennej *Petal.Width*, co sprawia, że jest ona lepszą cechą dyskretyzacyjną niż *Petal.Length*.

2.4 Porównanie nienadzorowanych metod dyskretyzacji

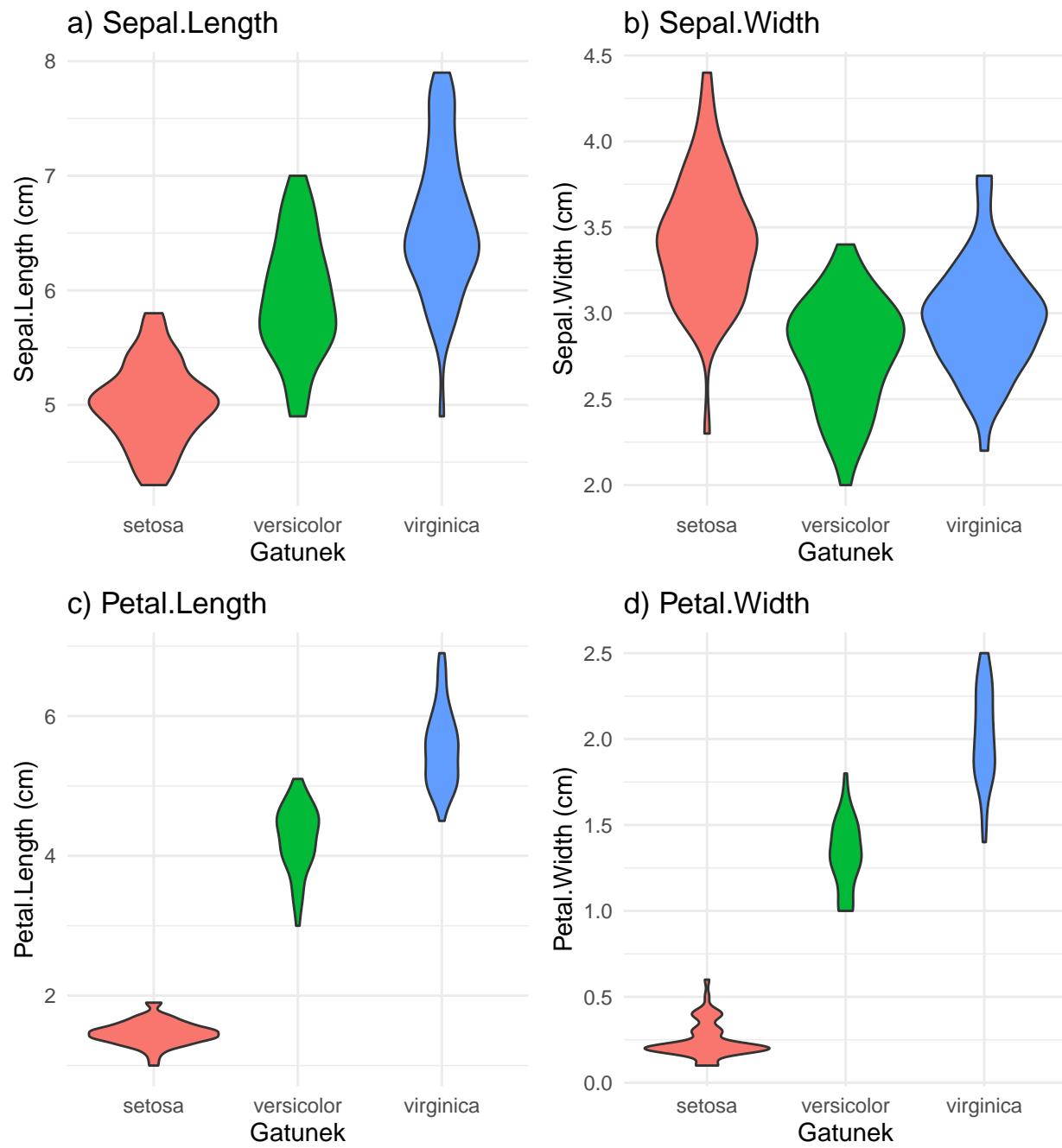
W tej części oceniać będziemy nienadzorowane metody dyskretyzacji *równych przedziałów, równych częstotliwości* i metodę *k-średnich* dla poszczególnych cech, co pozwoli nam ocenić nie tylko która zmienna, ale również który sposób podziału jest najlepszy do rozróżnienia gatunków.

```

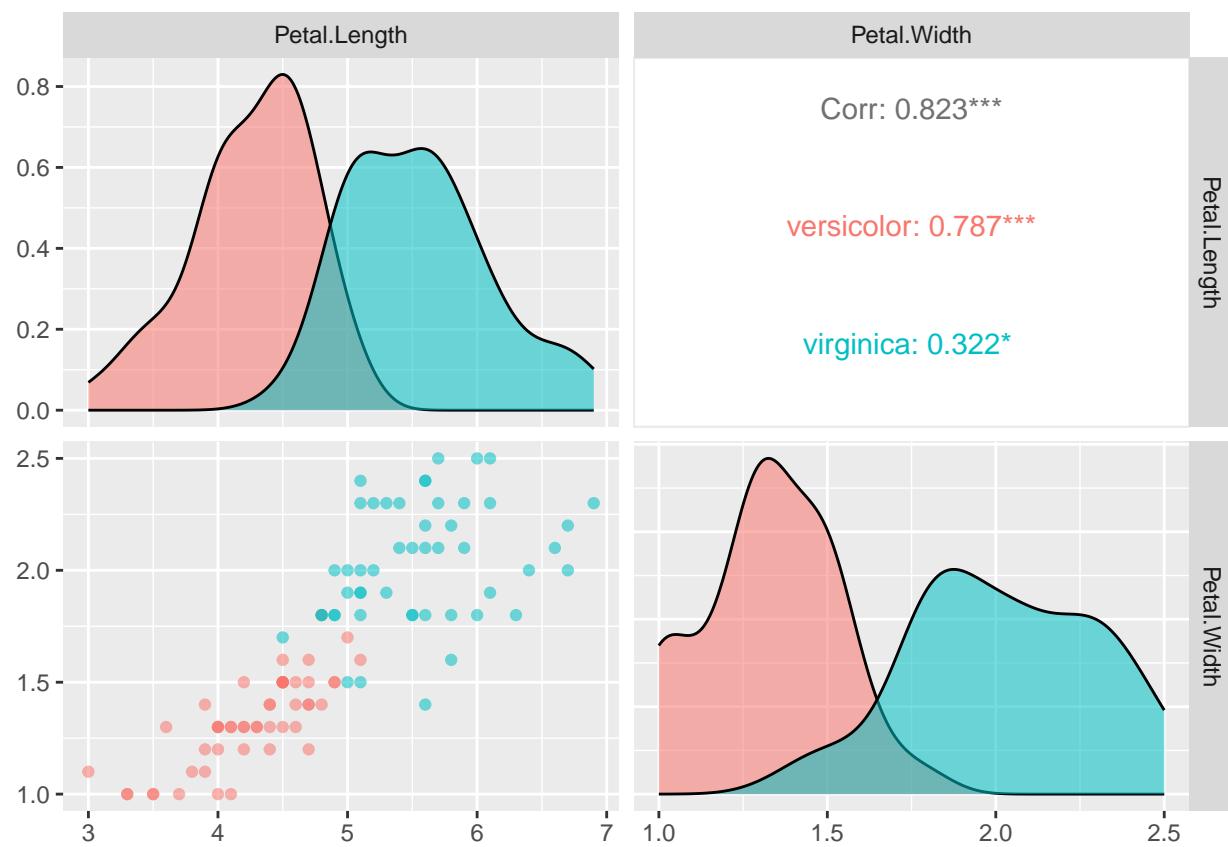
wyniki <- data.frame(matrix(ncol = 3, nrow = length(zmienne)))
colnames(wyniki) <- c("Równe przedziały", "Równa częstotliwość", "k-średnich")
rownames(wyniki) <- zmienne

wykresy <- list()

```



Wykres 3: Wykresy skrzypcowe dla zmiennych numerycznych
a) Sepal.Length b) Sepal.Width c) Petal.Length
d) Petal.Width



Wykres 4: Porównanie długości i szerokości płatków gatunków versicolor oraz virginica

```

for (zmienna in zmienne) {
  for (metoda in c("interval", "frequency", "cluster")) {
    dyskretyzacja <- discretize(iris[[zmienna]], method = metoda, breaks = 3)

    iris$dyskretyzacja <- dyskretyzacja

    wykres <- ggplot(iris, aes(x = dyskretyzacja, fill = Species)) +
      geom_bar(position = "fill") +
      ggtitle(paste0("Metoda ", ifelse(metoda == "interval",
                                         "równych przedziałów", ifelse(metoda == "frequency",
                                         "równych częstotliwości", "k-średnich")))) +
      labs(y = "Częstość", x = NULL) +
      theme_minimal() +
      theme(legend.position = "none")

    wykresy <- append(wykresy, list(wykres))

    match <- compareMatchedClasses(iris$Species, dyskretyzacja)$diag
    wyniki[zmienna, ifelse(metoda == "interval", "Równe przedziały",
                            ifelse(metoda == "frequency", "Równa częstotliwość",
                                   "k-średnich"))] <- match
  }
}

p <- ggplot(iris, aes(x = dyskretyzacja, fill = Species)) +
  geom_bar(position = "fill") +
  labs(fill = "Gatunek") +
  theme(direction = "horizontal")

legend <- get_legend(p)

## Warning in get_plot_component(plot, "guide-box"): Multiple components found;
## returning the first one. To return all, use 'return_all = TRUE'.

grid.arrange(
  grobs = c(wykresy[1:3], list(legend)),
  ncol = 1, heights = c(3, 3, 3, 1)
)

```

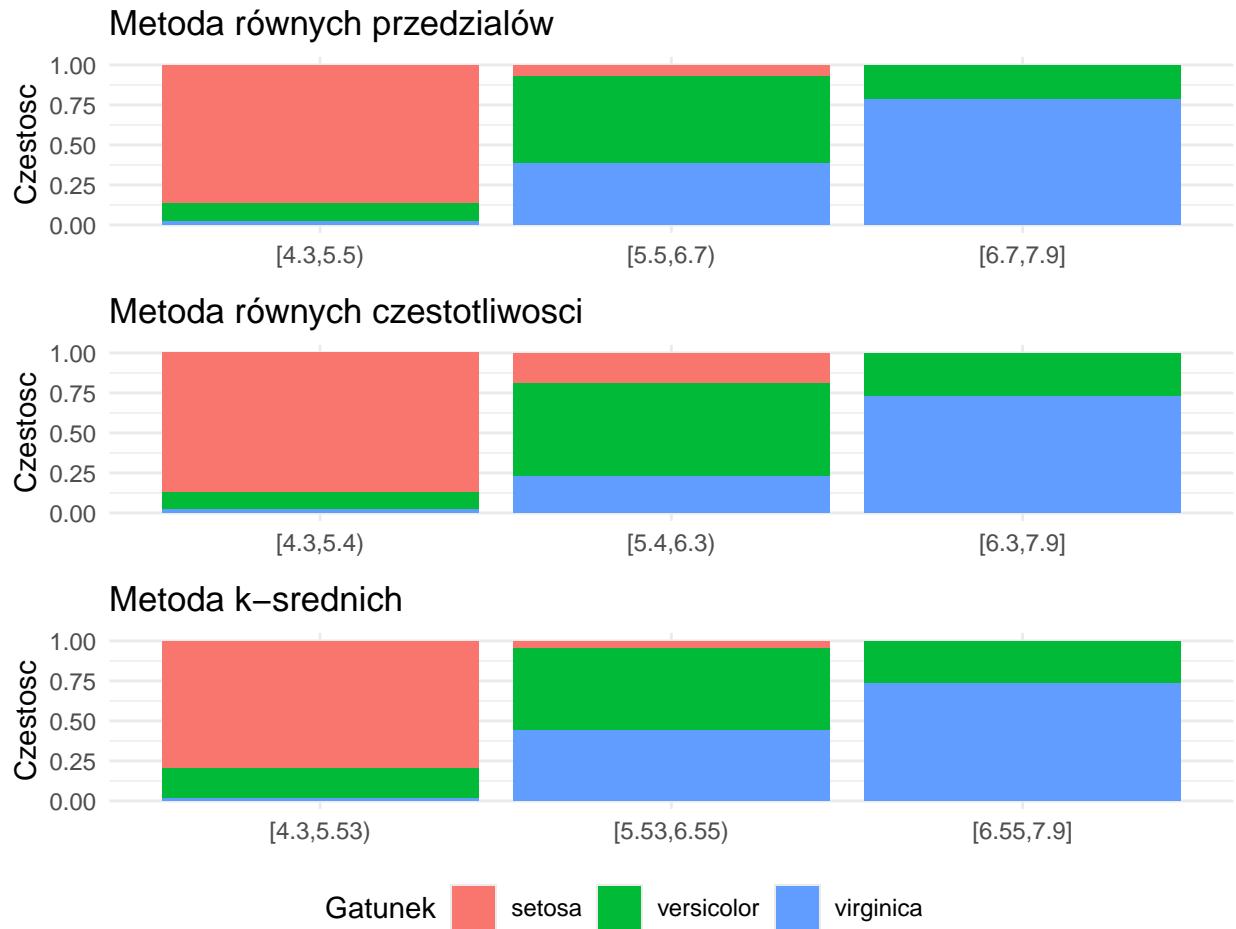
Z wykresu 5. wynika, że zmienna Sepal.Length wykazuje niewielką skuteczność w rozróżnianiu poszczególnych wyników. Porównując różne metody, można zauważyc, że ich wyniki są zbliżone do siebie.

```

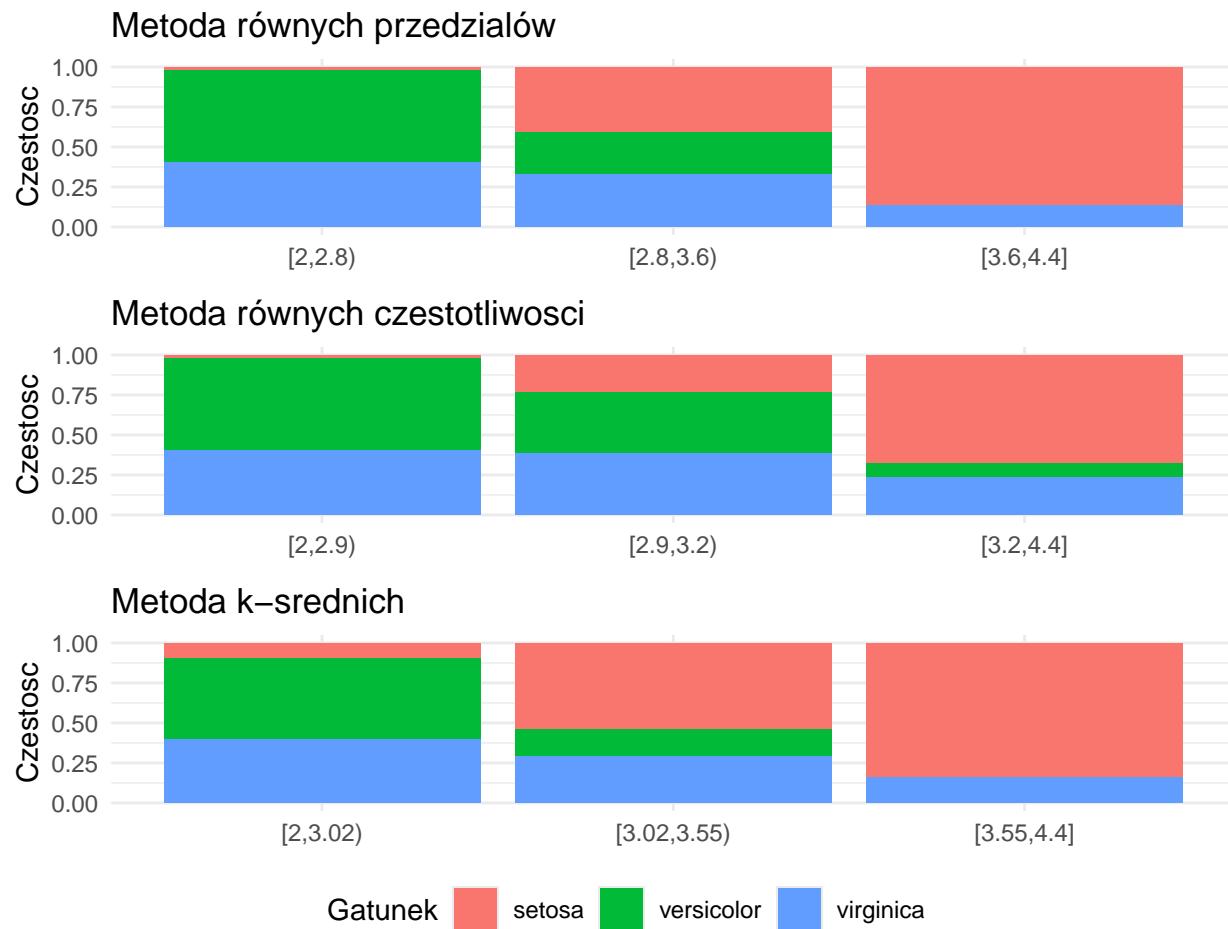
grid.arrange(
  grobs = c(wykresy[4:6], list(legend)),
  ncol = 1, heights = c(3, 3, 3, 1)
)

```

Na wykresie 6. widać, że podobnie jak wcześniej, zmienna Sepal.Width nie jest efektywna w rozróżnianiu gatunków irysów. Najgorsze wyniki uzyskała metoda równej częstotliwości, podczas gdy pozostałe metody dają zbliżone rezultaty.



Wykres 5: Metody dyskretyzacji dla zmiennej Sepal.Length

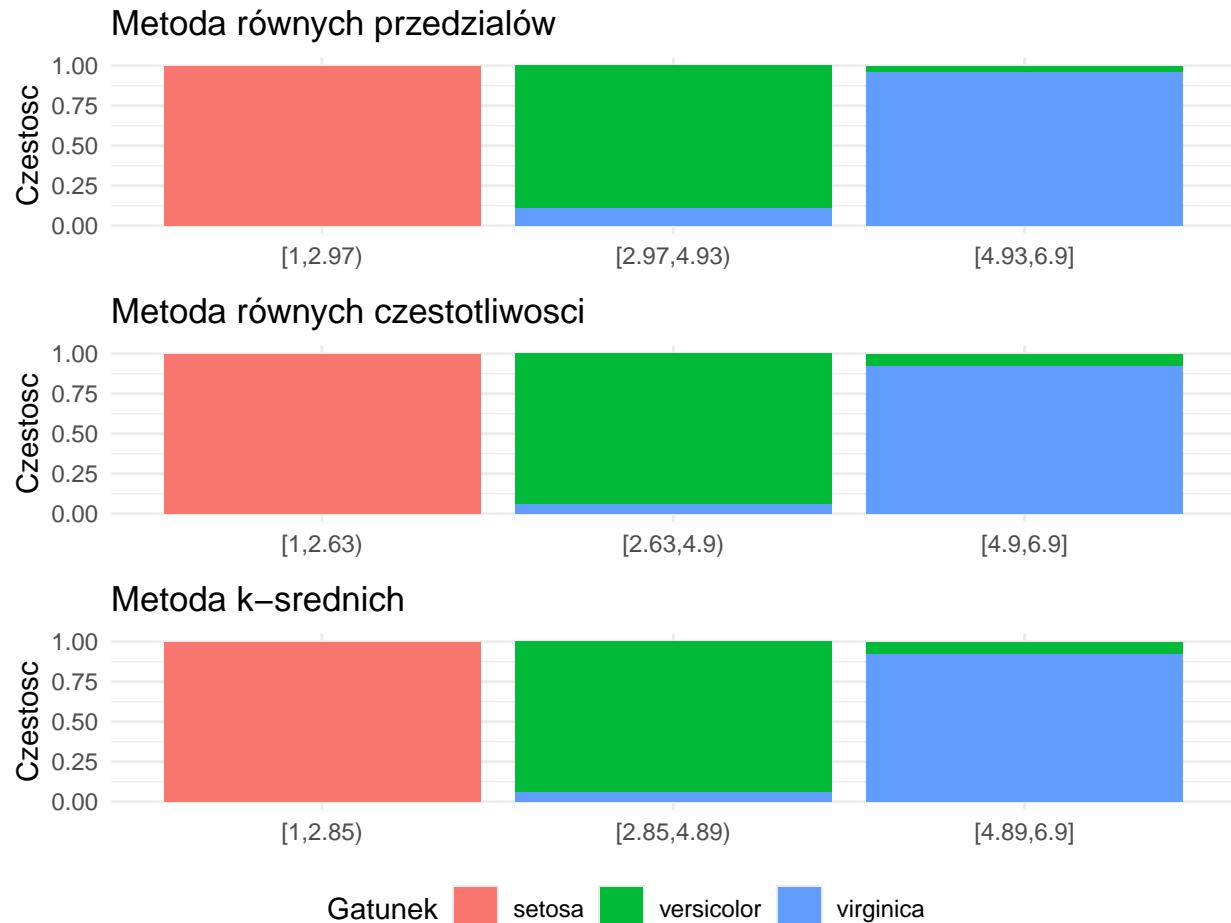


Wykres 6: Metody dyskretyzacji dla zmiennej Sepal.Width

```

grid.arrange(
  grobs = c(wykresy[7:9], list(legend)),
  ncol = 1, heights = c(3, 3, 3, 1)
)

```



Wykres 7: Metody dyskretyzacji dla zmiennej Petal.Length

Z wykresu 7. wynika, że zmienna `Petal.Length` umożliwia bardzo skuteczne rozróżnienie kwiatów na odpowiednie gatunki.

```

grid.arrange(
  grobs = c(wykresy[10:12], list(legend)),
  ncol = 1, heights = c(3, 3, 3, 1)
)

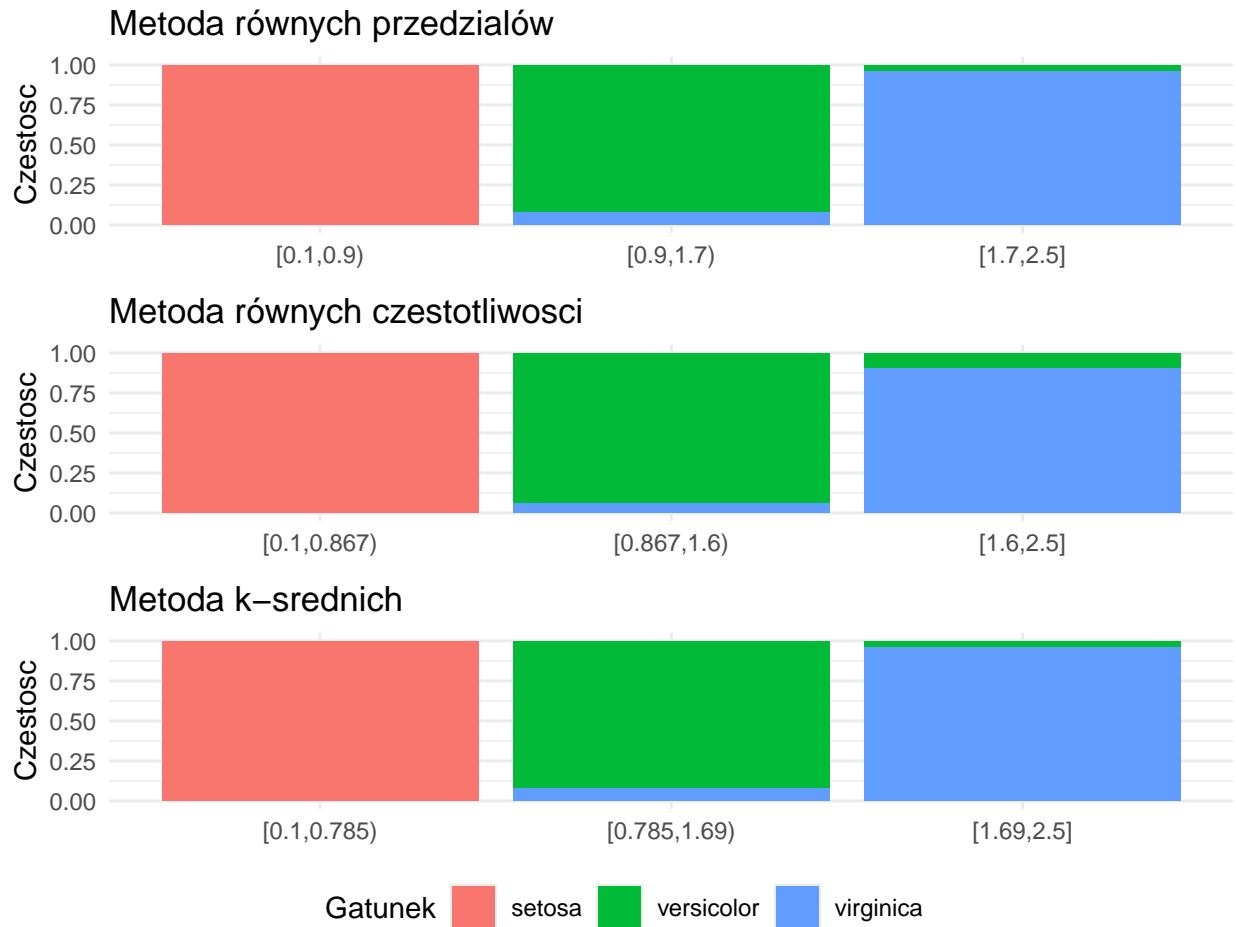
```

Z wykresu 8. wynika, że zmienna `Petal.Width`, podobnie jak `Petal.Length`, umożliwia bardzo skuteczne rozróżnienie kwiatów na odpowiednie gatunki.

```

średnie <- colMeans(wyniki)
wyniki_procentowe <- wyniki * 100

```



Wykres 8: Metody dyskretyzacji dla zmiennej Petal.Width

```

wyniki_procentowe <- rbind(wyniki_procentowe, średnie * 100)

rownames(wyniki_procentowe)[nrow(wyniki_procentowe)] <- "Średnia"

kable(wyniki_procentowe, caption = "Porównanie skuteczności metod dyskretyzacji w zależności od zmiennej")

```

Tabela 2: Porównanie skuteczności metod dyskretyzacji w zależności od zmiennej

	Równe przedziały	Równa częstotliwość	k-srednich
Sepal.Length	57.29	72.00	58.01
Sepal.Width	41.26	56.00	47.20
Petal.Length	94.67	95.33	95.33
Petal.Width	96.00	94.67	96.00
Średnia	72.30	79.50	74.14

Powyższa tabela pokazuje, jaką część kwiatów poprawnie sklasyfikowała dana metoda dla każdej zmiennej. Widać, że uzyskane wyniki różnią się w zależności od zmiennej ($>90\%$ dla płatków oraz $<75\%$ dla kielichów). Dla “najlepszej” cechy, czyli szerokości płatków, najlepszymi metodami są *k-srednich* oraz *równych przedziałów* i to pomimo, że średnio obie te metody wypadają gorzej od *równych częstotliwości*.

2.5 Dyskretyzacja za pomocą własnych przedziałów

Już wiemy, że zmienna Petal.Width najlepiej nadaje się do przeprowadzenia nienadzorowanej dyskretyzacji. Przyjmijmy teraz arbitralnie dwa punkty podziału: 0.75 oraz 1.65, aby utworzyć przedziały klas.

```

min <- min(iris$Petal.Width)
max <- max(iris$Petal.Width)
iris$dysk <- cut(iris$Petal.Width, breaks=c(min, 0.75, 1.65, max), right=FALSE, include.lowest=TRUE)

ggplot(iris, aes(x=dysk, fill=Species)) +
  geom_bar(position="fill") +
  labs(y="Częstość", x=NULL) +
  theme_minimal() +
  theme(legend.position = "bottom")

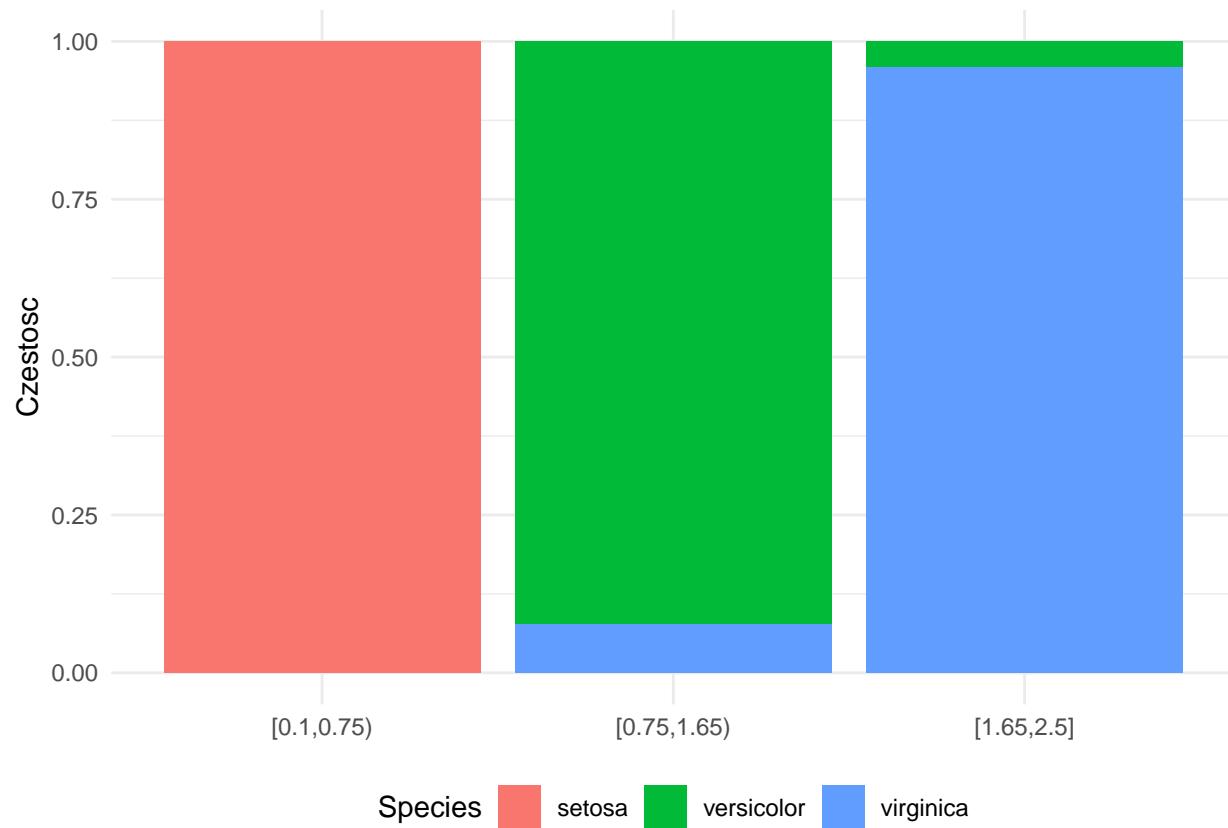
match <- compareMatchedClasses(iris$Species, iris$dysk)$diag

```

Zgodność własnych przedziałów wynosi **0.96**, jednak ta metoda jest mniej efektywna, ponieważ polega na ręcznym ustalaniu przedziałów przez człowieka. W przeciwieństwie do metod nienadzorowanych, które są realizowane przez komputer, automatycznie dopasowując optymalne granice, ta metoda może być subiektywna i nie zawsze precyzyjna, szczególnie przy dużych zbiorach danych.

2.6 * Stworzenie modelu predykcyjnego

W poprzednich zadaniach napotkaliśmy problem, który wynika z faktu, że metoda *k-srednich* była stosowana na całym zbiorze danych. Taki sposób działania może prowadzić do zjawiska overfittingu, czyli nadmiernego dopasowania modelu do danych treningowych, co skutkuje słabszą generalizacją na nowych, nieznanych



Wykres 9: Dyskretyzacja zmiennej Petal.Width - własne przedziały

danych. Aby temu zapobiec, w tej dodatkowej części przeprowadzimy podział zbioru danych na część treningową i testową. Dzięki temu będziemy mogli przeprowadzić trening modelu na zbiorze treningowym, a następnie ocenić jego skuteczność na zbiorze testowym, co pozwoli nam uzyskać bardziej wiarygodną ocenę jakości modelu.

Zbiór danych został podzielony w proporcji 75% na zbiór treningowy i 25% na zbiór testowy. Ponieważ analizowany zbiór (iris) zawiera jedynie 150 obserwacji, wyniki pojedynczego podziału mogłyby być obarczone dużą losowością. Aby zwiększyć wiarygodność oceny skuteczności metod dyskretyzacji, procedurę losowego podziału powtórzono 1000 razy. Dla każdego podziału obliczono wartość dopasowania, a uzyskane wyniki uśredniono i zaprezentowano w poniższej tabeli.

```

zmienna <- "Petal.Width"
wyniki <- data.frame()
liczba_iteracji <- 1000

for (i in 1:liczba_iteracji) {
  trening_set <- iris[sample(nrow(iris), 3/4 * nrow(iris)), ]
  test_set    <- iris[!rownames(iris) %in% rownames(trening_set), ]

  for (metoda in c("interval", "frequency", "cluster")) {
    dyskretyzacja <- discretize(trening_set[[zmienna]],
                                  method = metoda, breaks = 3)
    breaks <- attr(dyskretyzacja, "discretized:breaks")

    test_set$dyskretyzacja <- cut(test_set[[zmienna]],
                                    breaks = breaks,
                                    include.lowest = TRUE,
                                    labels = FALSE)

    match <- compareMatchedClasses(test_set$Species,
                                   test_set$dyskretyzacja)$diag
    wyniki[metoda, i] <- match
  }
}

średnie_wyniki <- rowMeans(wyniki)

wyniki_procentowe <- średnie_wyniki * 100
wyniki_procentowe <- as.data.frame(t(wyniki_procentowe))
rownames(wyniki_procentowe) <- "Średnia"

kable(wyniki_procentowe, caption = "Porównanie skuteczności metod dyskretyzacji", digits = 2, col.names

```

Tabela 3: Porównanie skuteczności metod dyskretyzacji

Równe przedziały	Równa częstotliwość	k-średnich
Średnia	96.14	91.59
		94.76

Metoda *równych przedziałów* osiągnęła najwyższą średnią skuteczność, przewyższając skuteczność metody *k-średnich* oraz *równych częstotliwości*. Sugeruje to, że dyskretyzacja oparta na równych przedziałach najlepiej odwzorowuje strukturę klas w danych dla zmiennej *Petal.Width*.

2.7 Podsumowanie

- Szerokość płatka (`Petal.Width`) stanowi najlepszy czynnik rozróżniający gatunki irysa.
- Długość płatka (`Petal.Length`) również skutecznie separuje poszczególne gatunki.
- Cechy działek kielicha (`Sepal.Length`, `Sepal.Width`) wykazują najmniejszą zdolność klasyfikacyjną, poprawnie klasyfikując mniej niż 75% próbek.
- Dyskretyzacja za pomocą klastrowania *k-srednich* osiąga najwyższą skuteczność w przypadku cech płatków.
- Największą moc predykcyjną wykazuje *metoda równych przedziałów*.
- *Metoda równoczęstotliwościowa* jest najsłabsza spośród nienadzorowanych metod, mimo że średnio daje najlepsze wyniki.
- Ręczna dyskretyzacja szerokości płatka (`Petal.Width`) przynosi wyniki zbliżone do innych metod, jednak jest subiektywna i trudna do uogólnienia.

3 Zadanie 2: Analiza składowych głównych (Principal Component Analysis (PCA))

Celem tej analizy jest zastosowanie metody analizy składowych głównych (PCA) do zbioru danych *City Quality of Life* (2020), który zawiera wskaźniki jakości życia w wybranych miastach, takie jak warunki mieszkaniowe, koszty utrzymania, bezpieczeństwo i opieka zdrowotna. Na początku przeanalizujemy zmienność cech, aby zdecydować, czy konieczna jest ich standaryzacja. Następnie, wyznaczmy składowe główne, ocenimy ich wkład w wyjaśnianie zmienności danych oraz zwizualizujemy wyniki, by lepiej zrozumieć podobieństwa między miastami. Na końcu zbadamy korelacje między zmiennymi oraz podsumujemy główne wnioski z przeprowadzonej analizy.

3.1 Własności danych

Zbiór danych *City Quality of Life* zawiera **266** przypadków oraz **21** cech. Liczba brakujących danych wynosi **0**.

Potencjalne znaczenie poszczególnych cech oraz ich typ przedstawiono w tabeli 4. Problemem jest jednak brak opisu zmiennych w zbiorze danych, co uniemożliwia ich jednoznaczną interpretację.

```
df_table <- data.frame(
  Typ = sapply(data, class),
  Opis = c("Indeks obserwacji",
  "Nazwa miasta",
  "Kraj miasta",
  "Kontynent miasta",
  "Wskaźnik jakości warunków mieszkaniowych (0-10)",
  "Wskaźnik kosztów życia (0-10)",
  "Wskaźnik liczby startupów (0-10)",
  "Wskaźnik dostępności kapitału venture (0-10)",
  "Wskaźnik dostępności transportu i łączności (0-10)",
  "Wskaźnik czasu potrzebnego na dojazd do pracy (0-10)",
  "Wskaźnik wolności podejmowania działalności gospodarczej (0-10)",
  "Wskaźnik poziomu bezpieczeństwa (0-10)",
  "Wskaźnik jakości opieki zdrowotnej (0-10)",
  "Wskaźnik jakości systemu edukacji (0-10)",
  "Wskaźnik jakości środowiska naturalnego (0-10)",
  "Wskaźnik kondycji gospodarki miejskiej (0-10)",
  "Wskaźnik wysokości opodatkowania (0-10)",
  "Wskaźnik dostępności internetu (0-10)",
  "Wskaźnik oferty kulturalnej i rozrywkowej (0-10)",
  "Wskaźnik poziomu tolerancji społecznej (0-10)",
  "Wskaźnik jakości warunków do sportu i wypoczynku (0-10)")
)

library(knitr)
kable(df_table, col.names = c("Zmienna", "Typ", "Opis"),
  caption = "Opis zmiennych w zbiorze danych City Quality of Life")
```

Tabela 4: Opis zmiennych w zbiorze danych City Quality of Life

Zmienna	Typ	Opis
X	integer	Indeks obserwacji
UA_Name	character	Nazwa miasta
UA_Country	character	Kraj miasta
UA_Continent	character	Kontynent miasta
Housing	numeric	Wskaźnik jakości warunków mieszkaniowych (0-10)
Cost.of.Living	numeric	Wskaźnik kosztów życia (0-10)
Startups	numeric	Wskaźnik liczby startupów (0-10)
Venture.Capital	numeric	Wskaźnik dostępności kapitału venture (0-10)
Travel.Connectivity	numeric	Wskaźnik dostępności transportu i łączności (0-10)
Commute	numeric	Wskaźnik czasu potrzebnego na dojazd do pracy (0-10)
Business.Freedom	numeric	Wskaźnik wolności podejmowania działalności gospodarczej (0-10)
Safety	numeric	Wskaźnik poziomu bezpieczeństwa (0-10)
Healthcare	numeric	Wskaźnik jakości opieki zdrowotnej (0-10)
Education	numeric	Wskaźnik jakości systemu edukacji (0-10)
Environmental.Quality	numeric	Wskaźnik jakości środowiska naturalnego (0-10)
Economy	numeric	Wskaźnik kondycji gospodarki miejskiej (0-10)
Taxation	numeric	Wskaźnik wysokości opodatkowania (0-10)
Internet.Access	numeric	Wskaźnik dostępności internetu (0-10)
Leisure... Culture	numeric	Wskaźnik oferty kulturalnej i rozrywkowej (0-10)
Tolerance	numeric	Wskaźnik poziomu tolerancji społecznej (0-10)
Outdoors	numeric	Wskaźnik jakości warunków do sportu i wypoczynku (0-10)

3.2 Przygotowanie danych do analizy

Tworzymy podzbiór zawierający wyłącznie dane numeryczne.

```
data_num <- data[, sapply(data, is.numeric)][-1]
#Pierwsza kolumna zawiera dane typu int, jednak są one tylko unikalnymi identyfikatorami

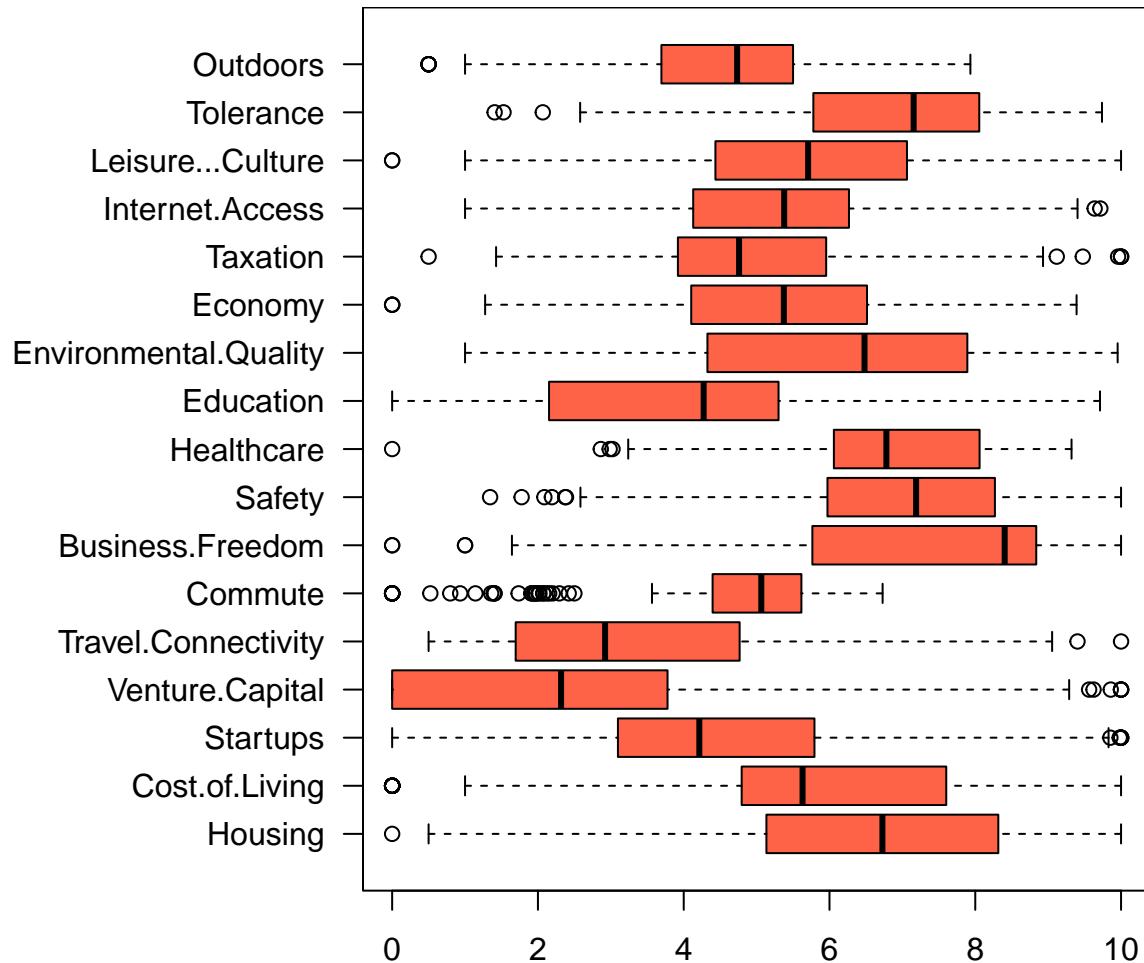
wariancje <- apply(data_num, MARGIN=2, FUN=function(x) sqrt(sum((x-mean(x))^2)/length(x)))

op <- par(no.readonly = TRUE)
par(mar = c(2, 10, 0, 2))
boxplot(data_num, las = 1, col = "tomato", horizontal = TRUE)

par(op)
```

Z wykresu 10. można zauważyć dużą różnorodność rozstępów zmiennych. Przykładem jest zmienna `Venture.Capital`, która ma niskie wartości, z tym że 25% najmniejszych danych ma wartość zero. Z kolei zmienna `Healthcare` ma dość wysokie wartości, przy czym dolna ćwiartka zaczyna się już od około 3. Co ważne, dane wymagają standaryzacji, co wynika z dużego zakresu zmienności, który w przypadku niektórych zmiennych wynosi od 1.48 do 2.55.

```
standarized_data_num <- scale(data_num, center=TRUE, scale=TRUE)
```

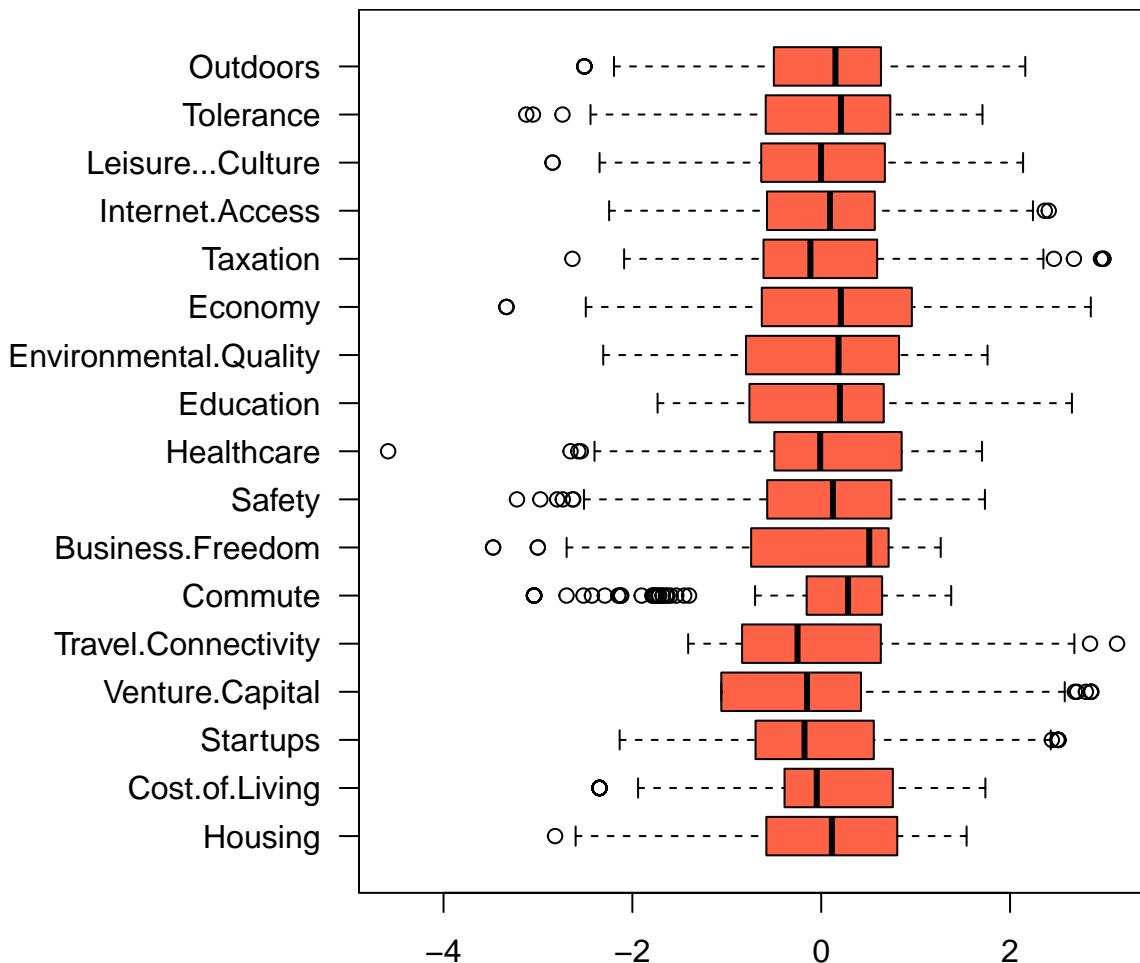


Wykres 10: Wykres pudelkowy zmiennych numerycznych (przed standaryzacją) - City Quality of Life

```

op <- par(no.readonly = TRUE)
par(mar = c(2, 10, 0, 2))
boxplot(standarized_data_num, las = 1, col = "tomato", horizontal = TRUE)

```



Wykres 11: Wykres pudelkowy zmiennych numerycznych (po standaryzacji) - City Quality of Life

```
par(op)
```

Na wykresie 11. widać, że wartości zmiennych są ujednolicione pod względem skali, a ich średnie i odchylenia standardowe mieszczą się w zbliżonych przedziałach. Odchylenie standardowe wynosi teraz 1 dla każdej zmiennej.

Mimo wyrównanych zakresów nadal można dostrzec dane odstające, co sugeruje znaczące różnice w wybranych aspektach jakości życia między niektórymi miastami.

3.3 Wyznaczanie składowych głównych

W celu dalszej analizy danych, przeprowadzimy wyznaczenie składowych głównych na podstawie standaryzowanych zmiennych. Metoda PCA pozwoli nam zredukować wymiarowość danych, zachowując jak największej informacji o ich zmienności.

```
pca_result <- prcomp(standarized_data_num, scale=TRUE)
```

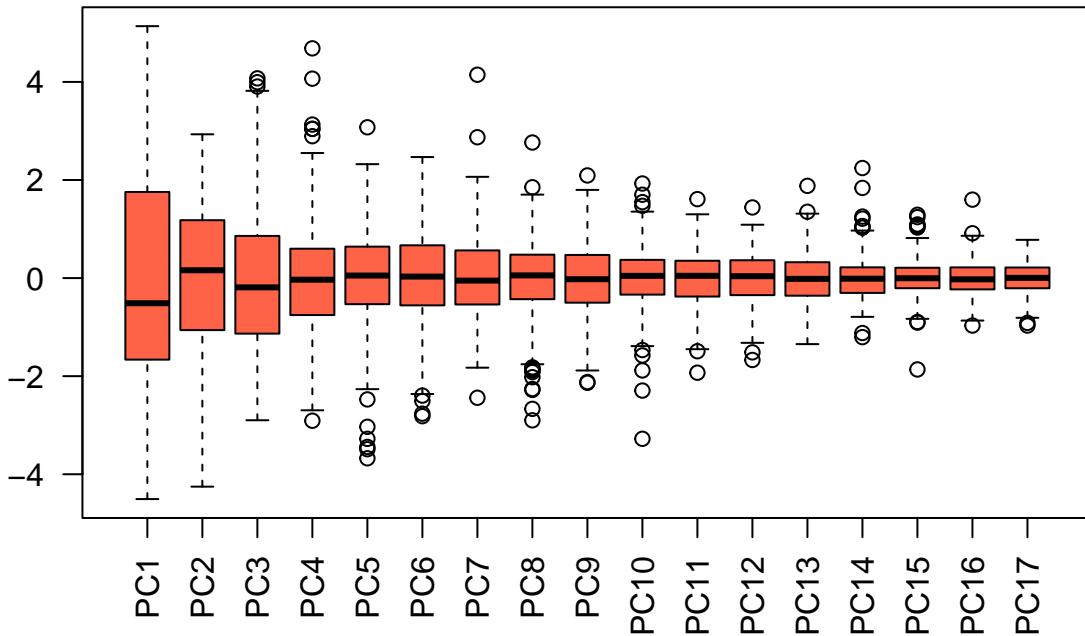
```
pca_summary <- summary(pca_result)$importance
```

#Do wykresów

```
xlab1 <- paste0("PC1 (", round(pca_summary[2, 1]*100, 2), "%)")
```

```
ylab1 <- paste0("PC2 (", round(pca_summary[2, 2]*100, 2), "%)")
```

```
boxplot(pca_result$x, las=2, col="tomato")
```



Wykres 12: Wykres pudełkowy wektorów ładunków składowych

Na wykresie 12. widać, że szerokość kolejnych pudełek (odpowiadających kolejnym składowym głównym) zmniejsza się. Oznacza to, że każda następna składowa główna wyjaśnia coraz mniejszą część wariancji danych. Co więcej, spadek ten nie jest liniowy — różnica procentowa pomiędzy kolejnymi składowymi staje się coraz mniejsza.

```
PC123 <- pca_result$rotation[, 1:3]
```

```
tbl_123 <- PC123[order(abs(PC123[, 1]), decreasing = TRUE), 1, drop = FALSE]
```

```
tabela2 <- PC123[order(abs(PC123[, 2]), decreasing = TRUE), 2, drop = FALSE]
```

```
tabela3 <- PC123[order(abs(PC123[, 3]), decreasing = TRUE), 3, drop = FALSE]
```

```
final_table <- data.frame(
```

```

Wektor_PC1 = rownames(tabela1),
Wartosc_PC1 = tabela1[, 1],
Wektor_PC2 = rownames(tabela2),
Wartosc_PC2 = tabela2[, 1],
Wektor_PC3 = rownames(tabela3),
Wartosc_PC3 = tabela3[, 1]
)

rownames(final_table) <- NULL
colnames(final_table) <- c("Wektor PC1", "Wartość PC1", "Wektor PC2", "Wartość PC2", "Wektor PC3", "Wartość PC3")

kable(final_table, caption = "Wektory i odpowiadające im wartości dla PC1, PC2 i PC3", digits = 2)

```

Tabela 5: Wektory i odpowiadające im wartości dla PC1, PC2 i PC3

Wektor PC1	Wartość PC1	Wektor PC2	Wartość PC2	Wektor PC3	Wartość PC3
Education	-0.40	Startups	-0.48	Commute	-0.51
Business.Freedom	-0.38	Venture.Capital	-0.43	Travel.Connectivity	-0.34
Environmental.Quality	-0.33	Leisure... Culture	-0.36	Safety	-0.33
Housing	0.31	Tolerance	0.36	Cost.of.Living	-0.33
Healthcare	-0.28	Safety	0.29	Housing	-0.31
Internet.Access	-0.28	Environmental.Quality	0.25	Economy	0.31
Economy	-0.27	Healthcare	0.24	Leisure... Culture	-0.31
Cost.of.Living	0.26	Outdoors	-0.19	Healthcare	-0.28
Venture.Capital	-0.24	Cost.of.Living	-0.18	Outdoors	-0.15
Travel.Connectivity	-0.21	Travel.Connectivity	-0.14	Tolerance	-0.10
Tolerance	-0.19	Taxation	0.11	Education	-0.07
Startups	-0.18	Business.Freedom	0.10	Environmental.Quality	0.05
Commute	-0.11	Economy	-0.07	Internet.Access	0.03
Outdoors	-0.09	Housing	0.05	Business.Freedom	0.02
Leisure... Culture	-0.07	Education	-0.05	Taxation	-0.02
Safety	-0.04	Commute	0.03	Venture.Capital	0.01
Taxation	0.03	Internet.Access	0.02	Startups	0.01

Z Tabeli 5. wynika, że pierwsza składowa główna (PC1) odzwierciedla przede wszystkim różnice w zakresie infrastruktury społecznej i środowiskowej oraz warunków mieszkaniowych. Miasta o wysokich wartościach PC1 charakteryzuje lepsza jakość mieszkań i niższe koszty życia, podczas gdy ujemne ładunki wskazują na słabsze wyniki w edukacji, wolności biznesu i jakości środowiska.

Druga składowa główna (PC2) wyodrębnia oś innowacyjności przeciwstawioną otwartości społecznej. Ujemne wartości PC2 skupią się na rozbudowanym ekosystemie startupowym i dostępie do kapitału venture, natomiast dodatnie ładunki odpowiadają za wyższy poziom tolerancji społecznej oraz bezpieczeństwa.

Trzecia składowa główna (PC3) kładzie nacisk na logistykę i gospodarkę. Najsilniejsze ujemne ładunki wskazują na długi czas dojazdów i słabą dostępność transportu, zaś dodatnie wartości PC3 wiążą się z lepszą kondycją gospodarczą miasta.

3.4 Korelacja zmiennych

Na dwuwymiarowym wykresie przedstawiono wektory ładunków poszczególnych zmiennych względem pierwszej i drugiej składowej głównej. Kierunek każdego wektora wskazuje, w jaki sposób dana zmienna wpływa

na obie składowe. Wektory o podobnym kierunku sugerują podobny wpływ na strukturę danych, przy czym silniejszy wpływ na PC1 ma ten, który tworzy mniejszy kąt z osią poziomą – ponieważ PC1 tłumaczy większą część całkowitej wariancji.

Kąty między wektorami odzwierciedlają natomiast zależności koreacyjne między zmiennymi. Im mniejszy kąt, tym silniejsza korelacja dodatnia; im bliżej kąta rozwartego, tym silniejsza korelacja ujemna. Kąt prosty sugeruje brak korelacji – zmienne są niezależne.

```
fviz_pca_var(pca_result,
              col.var = "contrib",
              gradient.cols = c("steelblue", "yellowgreen", "tomato"),
              labelsize = 4,
              repel = TRUE,
              xlab = xlab1,
              ylab = ylab1,
              title = NULL)
```

Na wykresie 13. biplot (PC1 vs. PC2) wyróżnić można cztery grupy zmiennych, które różnią się pod względem wpływu na wyjaśnianą wariancję oraz wzajemnych korelacji.

Z wykresu wynika, że **Taxation** i **Commute** mają minimalny wpływ na wspólną wariancję (najkrótsze wektory), niewiele silniej – **Outdoors** i **Safety**. Natomiast kluczową rolę pełnią **Education**, **Venture.Capital**, **Startups**, **Business.Freedom** oraz **Environmental.Quality**, których wektory są najdłuższe, co świadczy o ich dominującym udziale w kształtowaniu struktury danych.

Zmienne ulokowane w tej samej ćwiartce wykresu wykazują dodatnie korelacje. Najbardziej wyraźne skupiska pojawiają się w drugiej i trzeciej ćwiartce, gdzie wektory są niemal równoległe i blisko osi poziomej. W przeciwnieństwie do nich **Cost.of.Living** i **Housing** zwrócone są przeciwne do większości pozostałych wskaźników. Jednocześnie między nimi zachodzi dodatnia korelacja – wyższe oceny kosztów życia współwystępują z wyższymi ocenami warunków mieszkaniowych, kosztem pozostałych czynników.

Interpretacja przestrzeni biplota w ćwiartkach pozwala zmapować profile miast. W górnym lewym rogu znajdują się wskaźniki **Business.Freedom**, **Environmental.Quality**, **Tolerance**, **Healthcare** oraz **Internet.Access**. Po lewej stronie dominuje **Education**, w dolnej lewej – **Venture.Capital** i **Startups**, w dolnej prawej – **Cost.of.Living**, a w prawej części wykresu – **Housing**. Taki układ obrazowo odzwierciedla trzy główne wymiary zróżnicowania: jakość usług społecznych i środowiskowych, innowacyjność i biznes oraz warunki mieszkaniowo-transportowe.

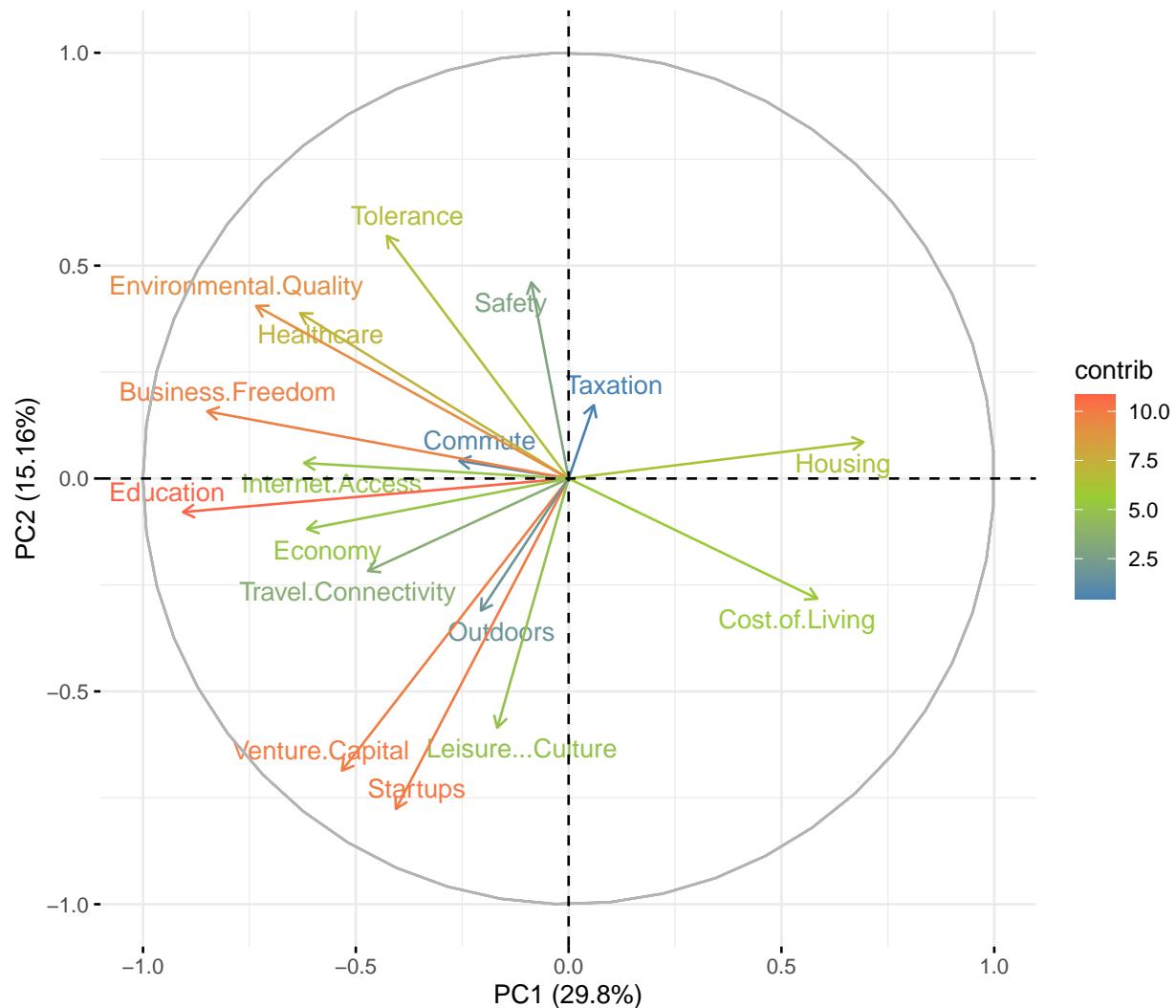
```
corrplot(cor(standarized_data_num), method = "shade", type = "upper",
         col = colorRampPalette(c("blue", "white", "red"))(200),
         tl.col = "black", tl.srt = 45,
         addCoef.col = "black",
         diag = FALSE,
         number.cex=0.7)
```

Powyższa macierz korelacji potwierdza najważniejsze wnioski z biplotu. Wartości korelacji między zmiennymi w macierzy odzwierciedlają zależności, które można zaobserwować na wykresie biplota, zwłaszcza w odniesieniu do zmiennych umiejscowionych w tej samej ćwiartce, które wykazują silne dodatnie korelacje.

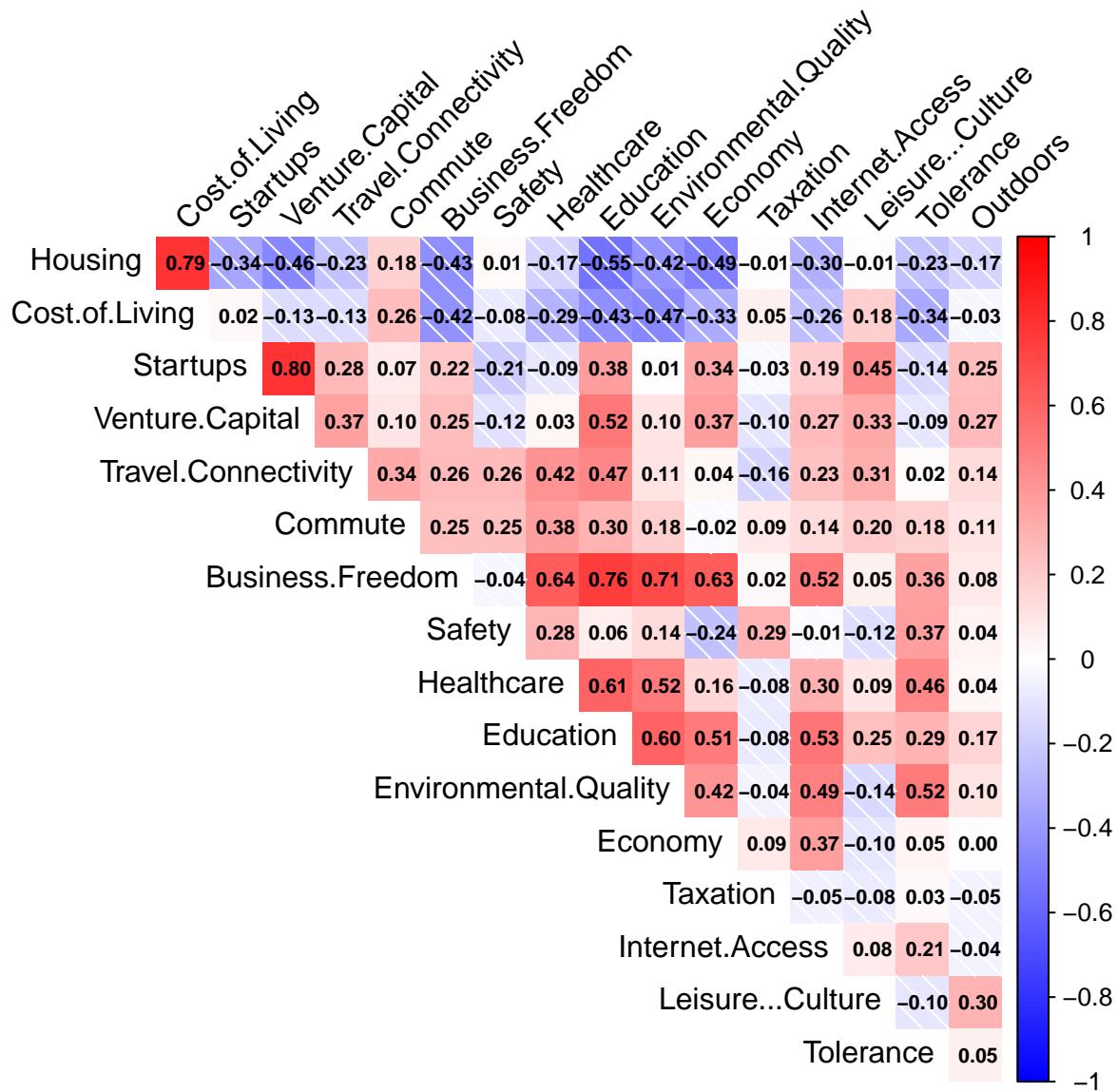
3.5 Zmienna odpowiadająca poszczególnym składowym

W tej części sprawdzimy, jaką część skumulowanej wariancji tłumaczą kolejne składowe główne (PCA). W tym celu będziemy analizować wartości skumulowanej wariancji dla poszczególnych składowych i sprawdzić, jak każda z nich przyczynia się do wyjaśniania ogólnej wariancji danych.

Variables – PCA



Wykres 13: Dwuwykres dla dwóch pierwszych składowych



Wykres 14: Macierz korelacji zmiennych numerycznych

```

pca_summary <- summary(pca_result)

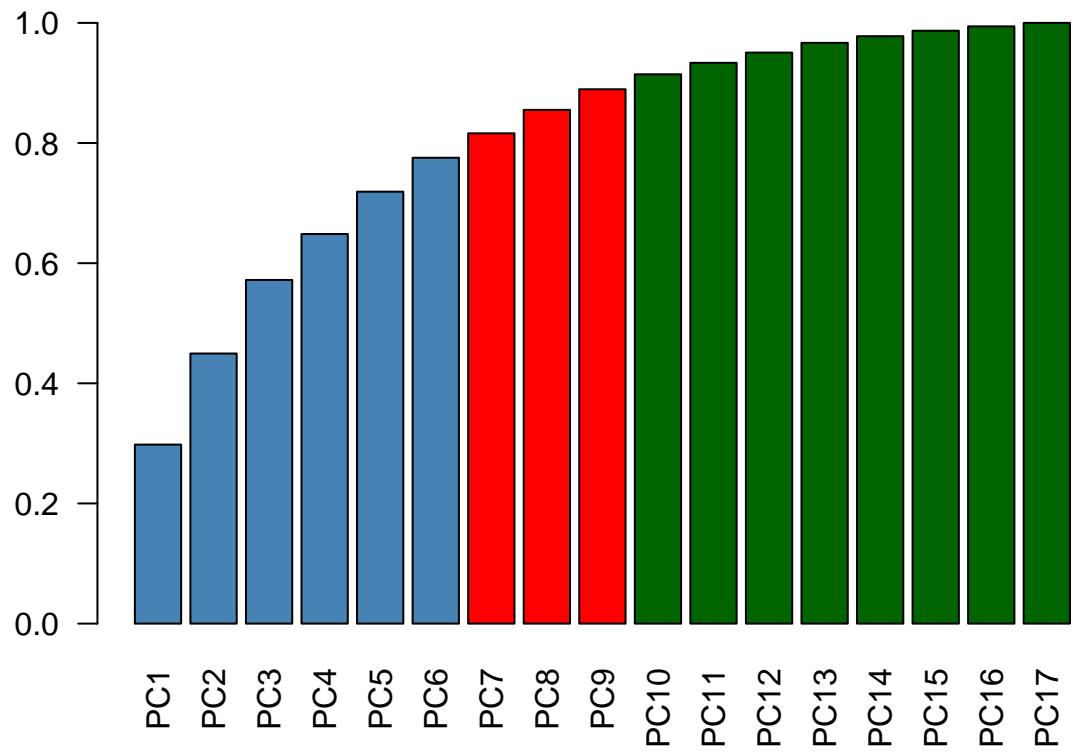
which_80_cumulative <- which(pca_summary$importance["Cumulative Proportion", ] >= 0.80)
which_90_cumulative <- which(pca_summary$importance["Cumulative Proportion", ] >= 0.90)
ile_pca <- ncol(pca_summary$importance)

cumulative_50 <- min(which(pca_summary$importance["Cumulative Proportion", ] >= 0.50))
ile_PC12 <- pca_summary$importance["Cumulative Proportion", 2]

colors <- rep("steelblue", length(pca_summary$importance["Cumulative Proportion", ]))
colors[which_80_cumulative] <- "red"
colors[which_90_cumulative] <- "darkgreen"

barplot(
  pca_summary$importance["Cumulative Proportion", ],
  las = 2,
  col = colors)

```



Wykres 15: Skumulowana składowa wariancja

Na powyższym wykresie kolory oznaczają, jaką część skumulowanej wariancji tłumaczą:

- *Niebieski* - poniżej 80%,
- *Czerwony* - równo lub więcej niż 80%, ale poniżej 90%,
- *Zielony* - równo lub więcej niż 90%.

Można zauważyć, że aby wytłumaczyć co najmniej 80% wariancji danych, wystarczy nam 7 z 17 składowych, a dla co najmniej 90% - 10. Jednakże do ich wizualizacji możemy użyć maksymalnie 3 składowych ze względu na ograniczenie trójwymiarowości prawdziwego świata. Na szczeble PC1 i PC2 odpowiadają za 44.96% wariancji, co daje nam dostateczną reprezentację struktury danych w dwóch wymiarach.

3.6 Wizualizacja danych wielowymiarowych

W tej części skupimy się na wizualizacji wyników dla miast, wykorzystując wykres pokazujący relację między pierwszą (PC1) a drugą (PC2) składową główną. W celu dokładniejszej analizy, dane będą podzielone według kontynentów lub krajów, co pozwoli na identyfikację grup miast oraz ewentualnych punktów odstających. Celem jest próba zrozumienia struktury tych danych i odpowiedzenie na pytanie, dlaczego poszczególne miasta są rozmieszczone w taki sposób, a nie inaczej. Na wykresie w tle będzie również można zauważać miasta należące do innych grup, co umożliwia szersze spojrzenie na całą analizowaną przestrzeń.

```
main_components <- pca_result$x[, c(1, 2)]

punkty <- function(kontynent, uwzgledniaj_kraje = FALSE){
  df <- data %>%
    mutate(
      PC1 = main_components[, 1],
      PC2 = main_components[, 2]
    )

  continent_levels <- unique(df$UA_Continent)
  continent_colors <- rainbow(length(continent_levels))
  names(continent_colors) <- continent_levels

  df_cont <- df %>% filter(UA_Continent == kontynent)
  country_levels <- unique(df_cont$UA_Country)
  country_colors <- rainbow(length(country_levels))
  names(country_colors) <- country_levels

  p <- ggplot(df, aes(x = PC1, y = PC2)) +
    geom_point(
      colour = adjustcolor(continent_colors[df$UA_Continent], alpha.f = 0.2),
      size = 3
    ) +
    geom_text(
      aes(label = UA_Name),
      nudge_y = 0.3,
      colour = adjustcolor("black", alpha.f = 0.2),
      size = 2
    ) +
    geom_point(
      data = df_cont,
      colour = continent_colors[kontynent],
      size = 3
    )
}
```

```

) +
geom_text(
  data = df_cont,
  aes(label = UA_Name),
  nudge_y = 0.3,
  colour= "black",
  size = 3
) +
labs(x = xlab1, y = ylab1) +
theme_minimal()

if (uwzgledniaj_kraje) {
  country_colors <- setNames(rainbow(length(unique(df_cont$UA_Country))), s = 0.7, v = 0.9),
  unique(df_cont$UA_Country))

  p <- p +
    geom_point(
      data = df_cont,
      aes(colour = UA_Country),
      size = 3
    ) +
    scale_colour_manual(
      name = "Kraje",
      values = country_colors
    ) +
    guides(
      colour = guide_legend(
        nrow = 3,
        byrow = TRUE
      )
    ) +
    theme(
      legend.position = "bottom",
      legend.box = "vertical",
      legend.title = element_text(size = 8),
      legend.text = element_text(size = 7),
      legend.key.size = unit(0.5, "cm"),
      legend.spacing.y = unit(0.2, "cm"),
      legend.margin = margin(t = 0, b = 0)
    )
  }

  print(p)
}

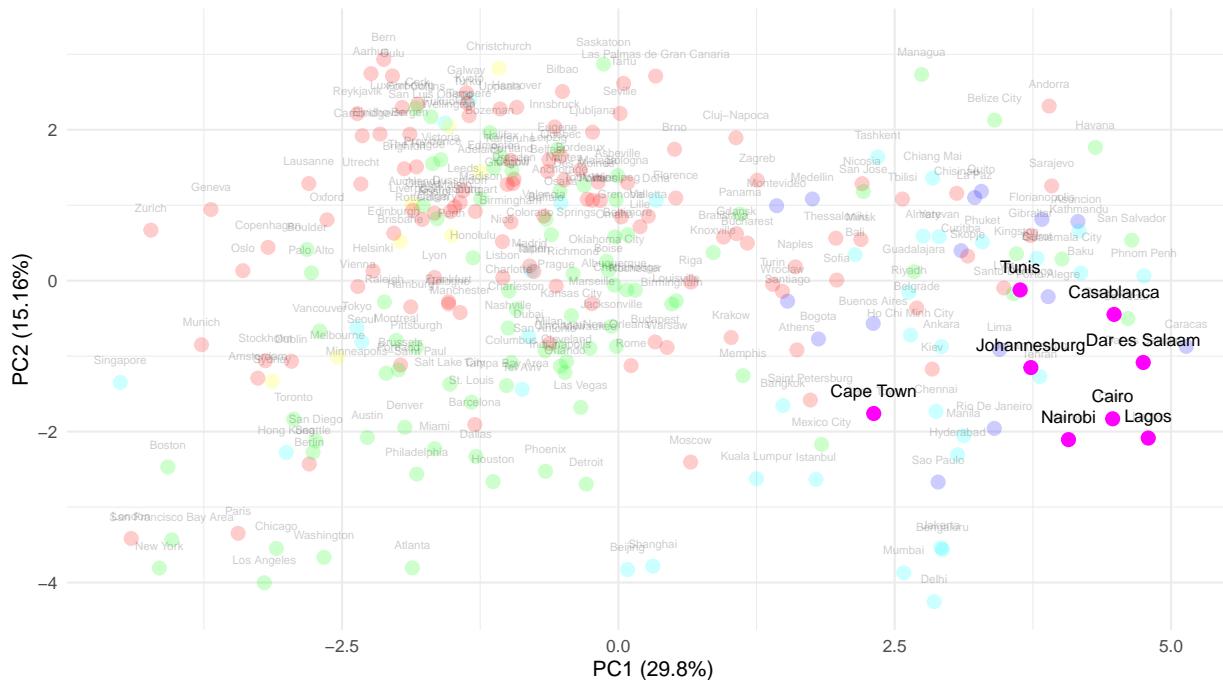
continents <- sort(unique(data$UA_Continent))
j <- 1

punkty(continents[j])

j <- j+1

```

Na wykresie 16. miasta afrykańskie tworzą jedną grupę, umiejscowioną po prawej stronie wykresu, co wska-



Wykres 16: Wykres rozrzutu składowych głównych; Afryka

zuje, że różnią się one od większości miast w zbiorze danych. Ich położenie w tej części wykresu, z mniejszą liczbą innych miast w tle, sugeruje, że miasta afrykańskie mają specyficzne cechy, które wyróżniają je na tle globalnym. Ich umiejscowienie trochę w dół na wykresie może wskazywać na niższe wyniki w niektórych aspektach, takich jak infrastruktura, usługi społeczne czy rozwój gospodarczy, w porównaniu do miast z innych regionów.

Wartością odstającą jest Południowoafrykański **Kapsztad** (Cape Town), które znajduje się nieco w lewo od reszty miast. To może wynikać z jego specyficznych cech, takich jak lepsza infrastruktura, wyższy poziom rozwoju gospodarczego, większe inwestycje w turystykę i biznes, a także lepsze wyniki w zakresie jakości życia i usług społecznych. Kapsztad jest jednym z bardziej rozwiniętych miast w Afryce, co powoduje, że odstaje od innych miast afrykańskich na wykresie.

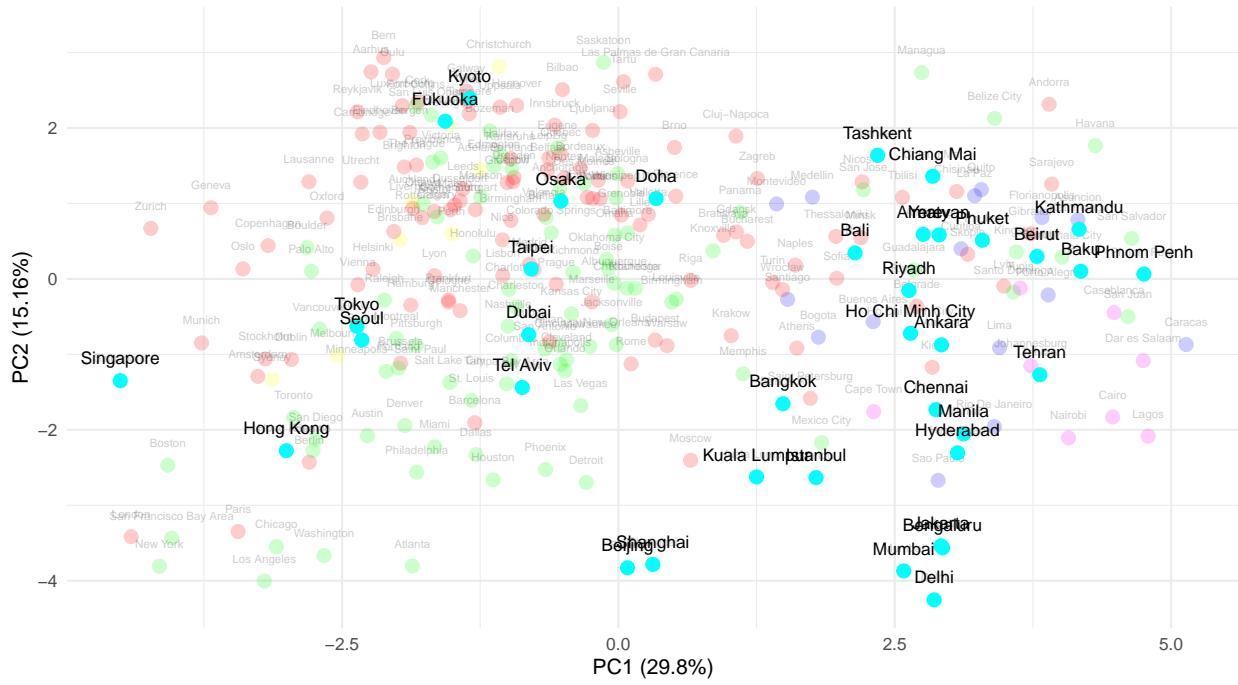
```
punkty(continents[j])
```

```
j <- j+1
```

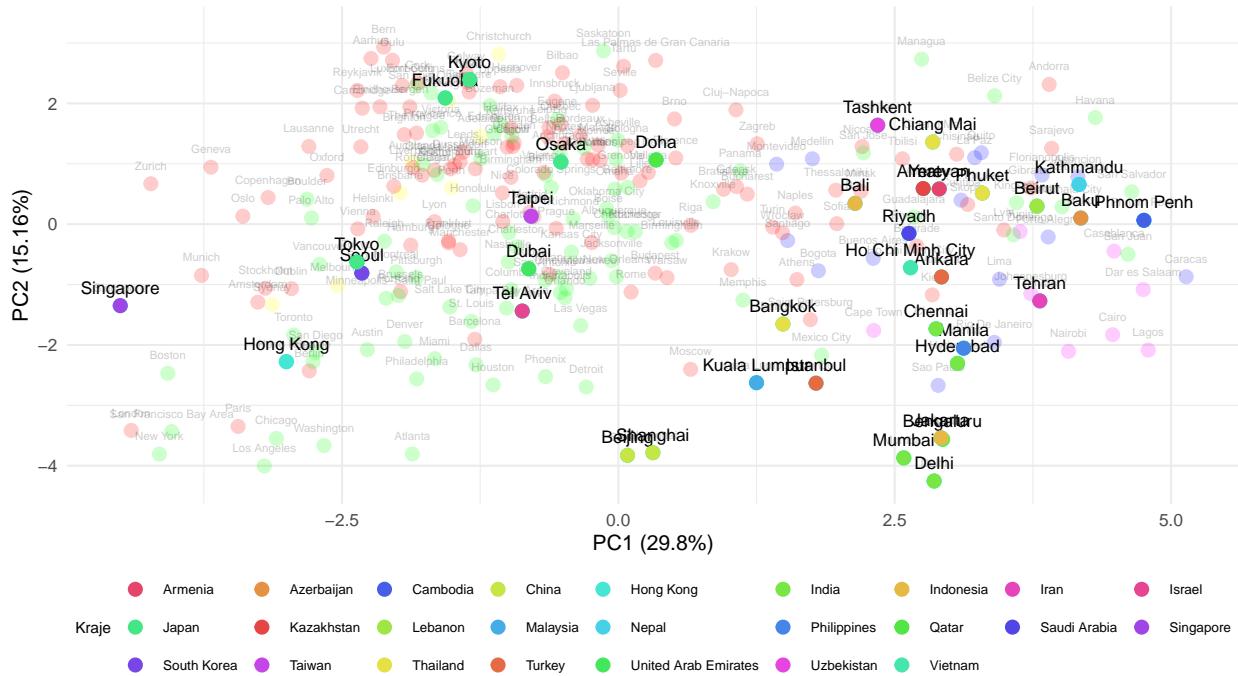
Na wykresie 17. wiele miast azjatyckich jest zgrupowanych po prawej stronie, co wskazuje na ich wspólne cechy. Widać także wiele miast po lewej stronie, które znajdują się na tle większej liczby miast. Dokonajmy podziału wykresu ze względu na kraje, aby spróbować zidentyfikować podobieństwa wśród nich.

```
punkty(continents[j-1], uwzględnij_kraje = TRUE)
```

Na wykresie 18. miasta po lewej stronie – czyli te o najwyższych wartościach wskaźników społeczeństwowo-środowiskowych – pochodzą głównie z najbogatszych państw Azji, takich jak Japonia, Singapur, Tajwan, Zjednoczone Emiraty Arabskie czy Katar. Spośród nich najbardziej wysunięty w lewo jest Singapur, co odzwierciedla jego niezwykle silny profil biznesowo-innowacyjny, natomiast w lewym dolnym rogu plasuje się Hongkong – finansowo-biznesowe centrum Chin z własną giełdą, co idealnie koresponduje z jego regionalnym znaczeniem.



Wykres 17: Wykres rozrzutu składowych głównych; Azja

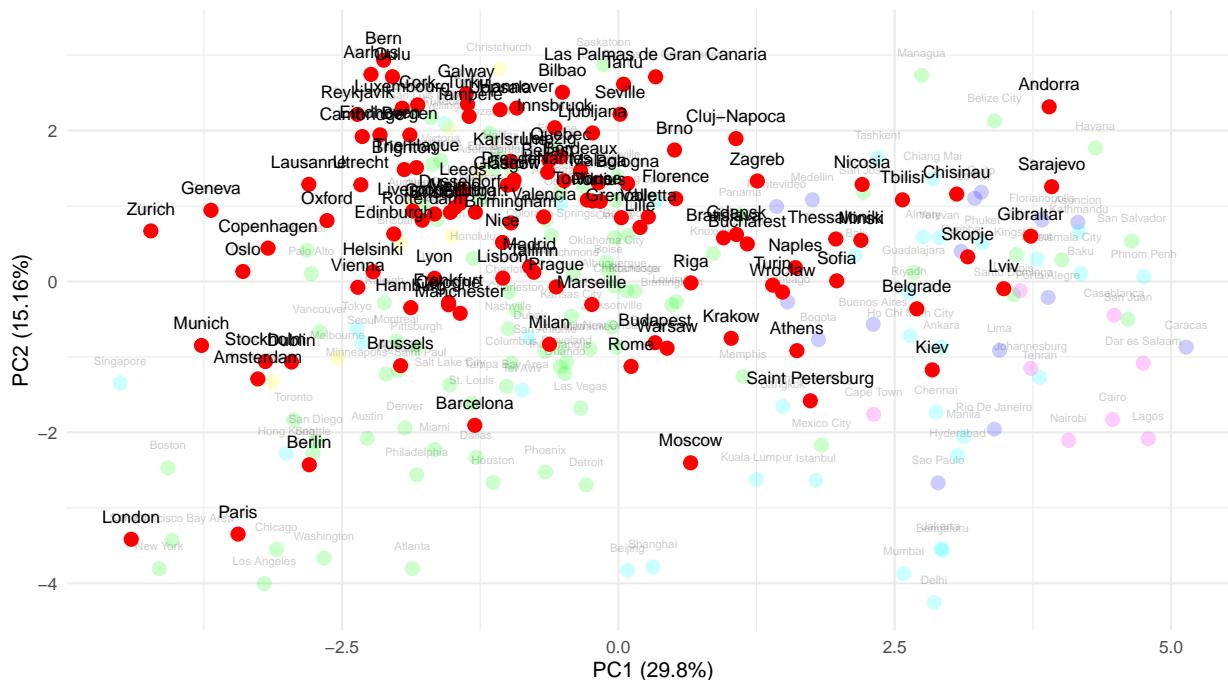


Wykres 18: Wykres rozrzutu składowych głównych; Azja - kraje

Z ciekawostek warto zwrócić uwagę na chińskie aglomeracje znajdujące się w dolnej części wykresu, czyli o niskich wartościach PC2, co świadczy o relatywnie słabszym poziomie usług społecznych i jakości życia.

Po prawej stronie wykresu znajdują się miasta z azjatyckich krajów rozwijających się, wśród których występuje duże zróżnicowanie wzdłuż osi PC2. Oznacza to, że choć łączy je relatywnie słabszy profil inwestycyjno-biznesowy, to znacznie różnią się pod względem jakości usług społecznych i środowiskowych — część z nich cechują wyższe koszty życia.

punkty(continents[j])



Wykres 19: Wykres rozrzutu składowych głównych; Europa

j <- j+1

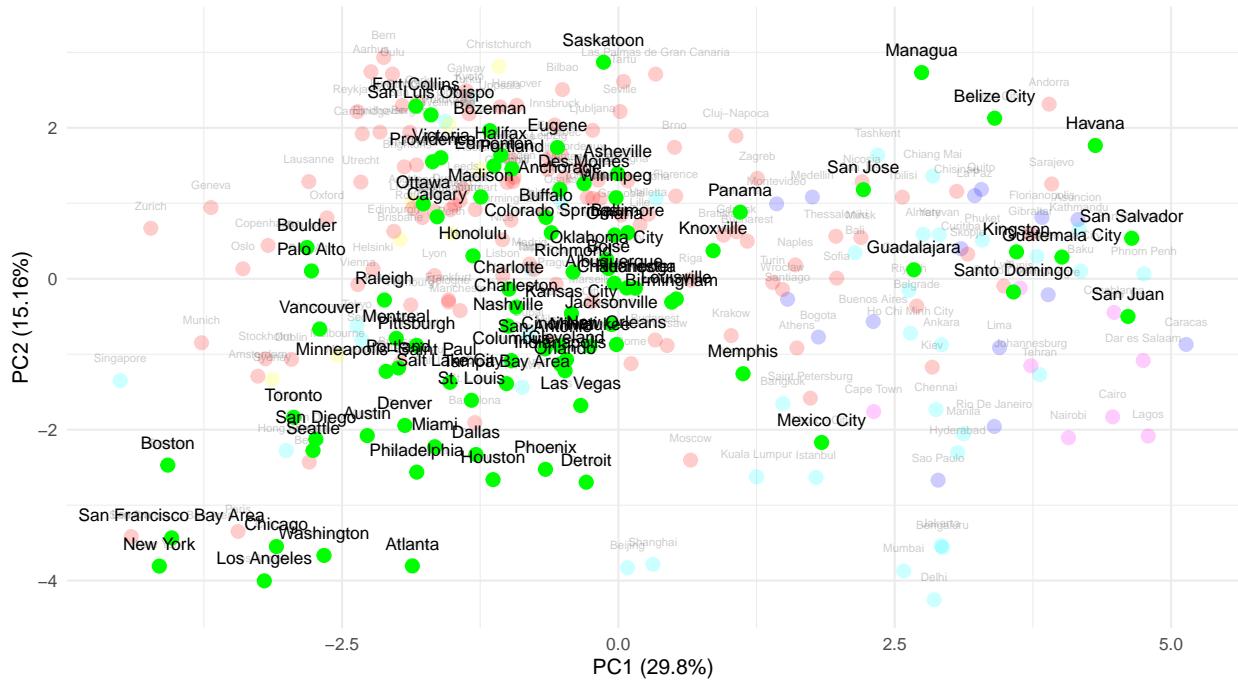
Na wykresie 19. widać, że miasta europejskie są najbardziej zróżnicowane spośród wszystkich analizowanych kontynentów — choć częściowo może to wynikać z ich dużej liczby w zbiorze danych. W centralnej części wykresu, z przesunięciem w kierunku lewego górnego rogu, znajduje się skupisko miast o zrównoważonych profilach społeczno-środowiskowych i umiarkowanym potencjale biznesowym. Świadczy to o względnej równowadze między jakością życia a możliwościami gospodarczymi.

Z kolei po prawej stronie, choć nadal blisko centrum, ulokowane są miasta z byłego bloku sowieckiego, takie jak Lwów, Nikozja, Wrocław czy Tbilisi. Ich pozycja może wskazywać na umiarkowany poziom kosztów życia i mieszkańców przy niższych wskaźnikach społecznych i ograniczonym potencjale inwestycyjnym. Na tym tle wyróżniają się Andora i Gibraltar — ich odstające pozycje mogą wynikać ze specyfiki ustrojowej i fiskalnej.

Wśród wyraźnie odstających punktów znajduje się także Moskwa, która uzyskuje stosunkowo niską wartość PC2 (niższa jakość usług społecznych), ale dodatnią PC1, co może wskazywać na pewien potencjał gospodarczy i inwestycyjny.

W lewym dolnym rogu wykresu zauważalne są stolice największych gospodarek Unii Europejskiej — Londyn, Paryż i Berlin. Ich pozycja sugeruje silny profil innowacyjno-inwestycyjny, wysoki poziom wolności gospodarczej i rozwinięte ekosystemy startupowe, co czyni je atrakcyjnymi ośrodkami dla kapitału i przedsiębiorczości.

```
punkty(continents[j])
```



Wykres 20: Wykres rozrzutu składowych głównych; Ameryka Północna

```
j <- j+1
```

Na wykresie 20. przedstawiającym miasta północnoamerykańskie można wyróżnić trzy główne grupy. Najliczniejsza z nich znajduje się w centralnej części wykresu, podobnie jak w przypadku miast europejskich. Obejmuje ona przede wszystkim miasta ze Stanów Zjednoczonych i Kanady, co świadczy o ich wysokim i zrównoważonym poziomie rozwoju społeczno-gospodarczego, łączącym dobre warunki życia z umiarkowanym potencjałem inwestycyjnym.

W lewym dolnym rogu widoczne są największe i najbardziej rozpoznawalne miasta amerykańskie, takie jak Nowy Jork, San Francisco, Chicago, Los Angeles, Waszyngton, Boston czy Atlanta. Ich pozycja odzwierciedla silny profil biznesowo-inwestycyjny — wysokie wartości zmiennych takich jak Venture.Capital, Startups czy Business.Freedom — oraz rolę, jaką pełnią jako centra gospodarcze i innowacyjne kraju.

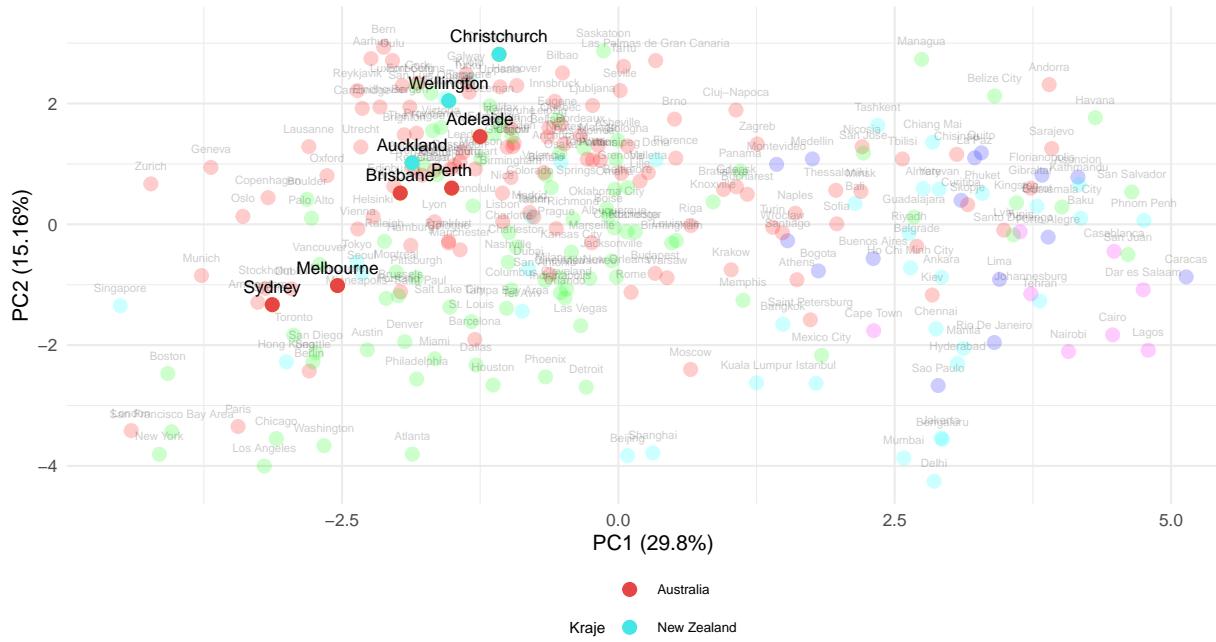
Z kolei po prawej stronie wykresu znajdują się miasta z Ameryki Środkowej i wysp karaibskich, co wskazuje na ich niższy poziom wskaźników gospodarczych i społecznych, a także relatywnie wyższe koszty życia.

Wśród miast odstających znajduje się Meksyk, który lokuje się nieco bardziej w lewym dolnym rogu w porównaniu do pozostałych miast regionu. Zawdzięcza to lepszym wynikom w zakresie aktywności gospodarczej i dostępności dla inwestorów, co wyróżnia go na tle sąsiadów z Ameryką Łacińską.

```
punkty(continents[j], uwzgledniaj_kraje = TRUE)
```

```
j <- j+1
```

Na wykresie 21. przedstawiono miasta z Australii i Oceanii, pochodzące wyłącznie z dwóch państw – Australii i Nowej Zelandii. Widać wyraźne podobieństwo tych miast między sobą, jak i ich ogólną zbieżność z miastami



Wykres 21: Wykres rozrzutu składowych głównych; Australia i Oceania - kraje

europejskimi i północnoamerykańskimi, co świadczy o wysokim poziomie rozwoju usług publicznych, jakości życia oraz stabilności instytucjonalnej.

Spośród wszystkich miast regionu najbardziej wyróżniają się Sydney i Melbourne, które są przesunięte w lewy dolny róg wykresu. Taka pozycja wskazuje na ich silny profil inwestycyjno-biznesowy.

Odzwierciedla to rzeczywistość: Sydney to główne centrum finansowe Australii, siedziba giełdy ASX oraz wielu międzynarodowych instytucji bankowych i korporacyjnych. Melbourne z kolei wyróżnia się dynamicznym środowiskiem startowym, rozwiniętą infrastrukturą badawczo-rozwojową i obecnością licznych firm z sektora usług profesjonalnych. Oba miasta pełnią rolę kluczowych hubów gospodarczych w regionie Azji i Pacyfiku.

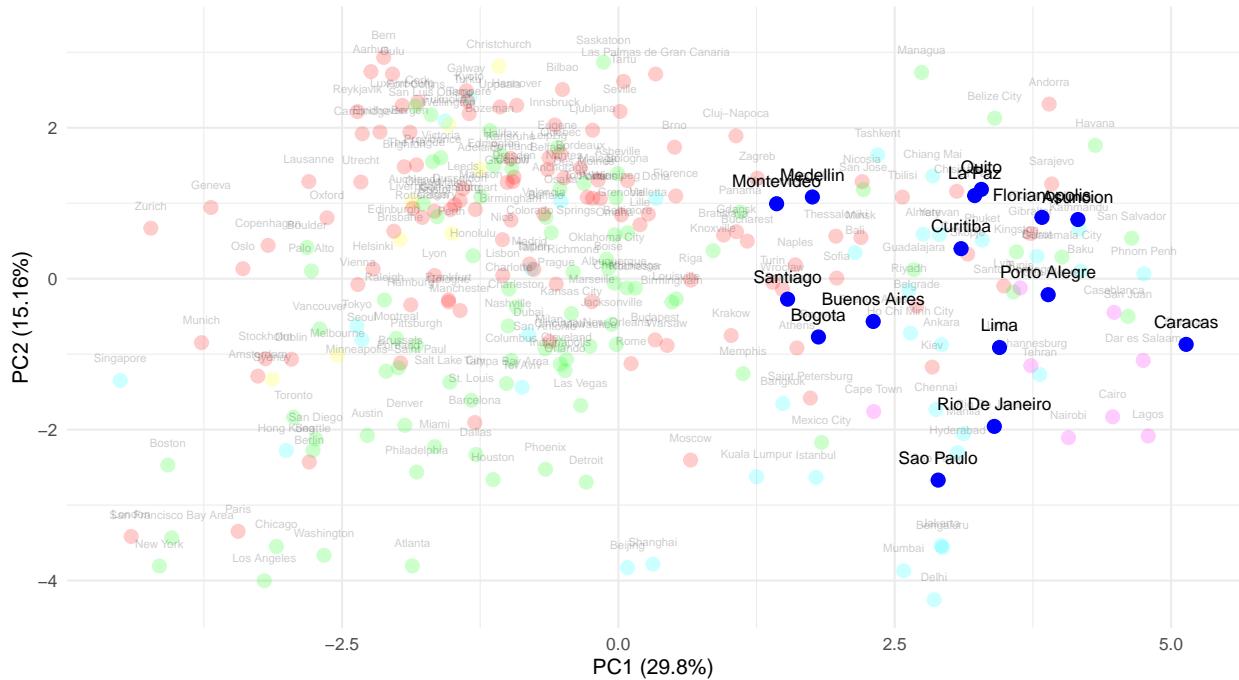
```
punkty(continents[j])
```

Na wykresie 22. miasta z Ameryki Południowej skupione są głównie po prawej stronie, co świadczy o relatywnie niższych wartościach wskaźników innowacyjno-biznesowych oraz umiarkowanym lub niskim poziomie jakości usług społecznych. Taka pozycja sugeruje ograniczony dostęp do kapitału inwestycyjnego, mniejsze otoczenie dla rozwoju startupów oraz wyzwanie w zakresie infrastruktury i swobód gospodarczych.

Najbardziej odstającym miastem na prawo jest **Caracas** – stolica Wenezueli – która od lat zmaga się z głębokim kryzysem gospodarczym, hiperinflacją, niestabilnością polityczną oraz odpływem ludności. Tego rodzaju problemy mają istotny wpływ na klimat inwestycyjny i jakość życia mieszkańców, co zostało bezpośrednio odzwierciedlone w analizie PCA i wizualizacji na biplocie.

3.7 Podsumowanie

Przeprowadzona analiza głównych składowych (PCA) umożliwiła redukcję wymiarowości zbioru danych przy jednoczesnym zachowaniu możliwie dużej części informacji zawartej w oryginalnych zmiennych. Dane wejściowe zostały uprzednio wystandardyzowane, co pozwoliło na eliminację wpływu różnic w jednostkach pomiarowych i skalach.



Wykres 22: Wykres rozrzutu składowych głównych; Ameryka Południowa

Pierwsze dwie główne składowe (PC1 i PC2) wyjaśniają łącznie około 45% całkowitej wariancji zbioru danych. Choć nie jest to dominująca część zmienności, wynik ten umożliwia interpretację ogólnych tendencji i struktur obecnych w danych.

Składowa główna PC1 odzwierciedla przede wszystkim warunki życia i środowisko ekonomiczne. Wysokie wartości tej składowej są skorelowane z korzystnymi warunkami mieszkaniowymi i niższymi kosztami życia, natomiast wartości ujemne wskazują na większe obciążenia społeczne i ekonomiczne, takie jak ograniczony dostęp do edukacji czy niższa wolność gospodarcza.

Składowa PC2 różnicuje miasta pod względem innowacyjności i rozwoju technologicznego (ujemne wartości) oraz aspektów społecznych i środowiskowych, takich jak bezpieczeństwo, jakość powietrza czy dostępność przestrzeni publicznych (wartości dodatnie).

Rozmieszczenie miast w przestrzeni pierwszych dwóch głównych składowych wskazuje na istnienie silnych zróżnicowań, przy czym niektóre miasta (np. Singapur, Caracas, Kapsztad) wykazują odmienny profil od miast regionu.

Kolejne składowe wyjaśniają coraz mniejszą część wariancji danych, jednak ich udział nie maleje w sposób liniowy. Co więcej, poszczególne składowe mogą mieć różny wpływ na różne cechy – ich interpretacja wymaga zatem ostrożności i uwzględnienia kontekstu analitycznego.

Wnioski analizy biplotów:

- W Afryce klaster wyraźnie odseparowany od pozostałych miast na biplocie, z niższymi wynikami infrastruktury i usług społecznych; Kapsztad stanowi wyjątek z relatywnie silnym profilem. -W Azji dwa główne klastry – metropolie państw bogatszych (Singapur, ZEA, Japonia) oraz miasta rozwijających się państw o zróżnicowanej jakości usług społecznych i kosztach życia. Odstające są miasta Zachodniego Tajwanu.
- W Europie w centrum biplotu dominują główne miasta europejskie o zrównoważonych profilach; na prawo od nich skupisko miast byłego bloku sowieckiego, a poza nimi wyraźnie odstający liderzy biznesowi (Londyn, Paryż, Berlin).

- W Ameryce Północnej trzy klastry – zrównoważone miasta USA i Kanady, globalne huby biznesowe (NY, SF, Chicago), oraz regiony Ameryki Środkowej i Karaibów; Meksyk wyraźnie wyróżnia się na tle miast Ameryki Środkowej.
- W Australii i Oceanii jednorodny klaster o wysokich standardach życia i silnym potencjale inwestycyjnym, z Sydney i Melbourne na czele.
- W Ameryce Południowej większość miast osiąga niższe wskaźniki innowacyjno-biznesowe i usługowe; Caracas wyraźnie odstaje negatywnie.

Obecne dwie składowe dają nam dobry wgląd na wpływ danych na ogólne wyniki, jednak odpowiadają one za jedynie 44.96. Jest to zadowalająca część, jednak aby były bardziej poprawne, powinniśmy posiadać większy udział skumulowanej wariancji, np. zasugerowane wcześniej 80%.

3.8 Potrzeba standaryzacji

Zbadajmy jeszcze, czy aby na pewno potrzebna jest standaryzacja danych.

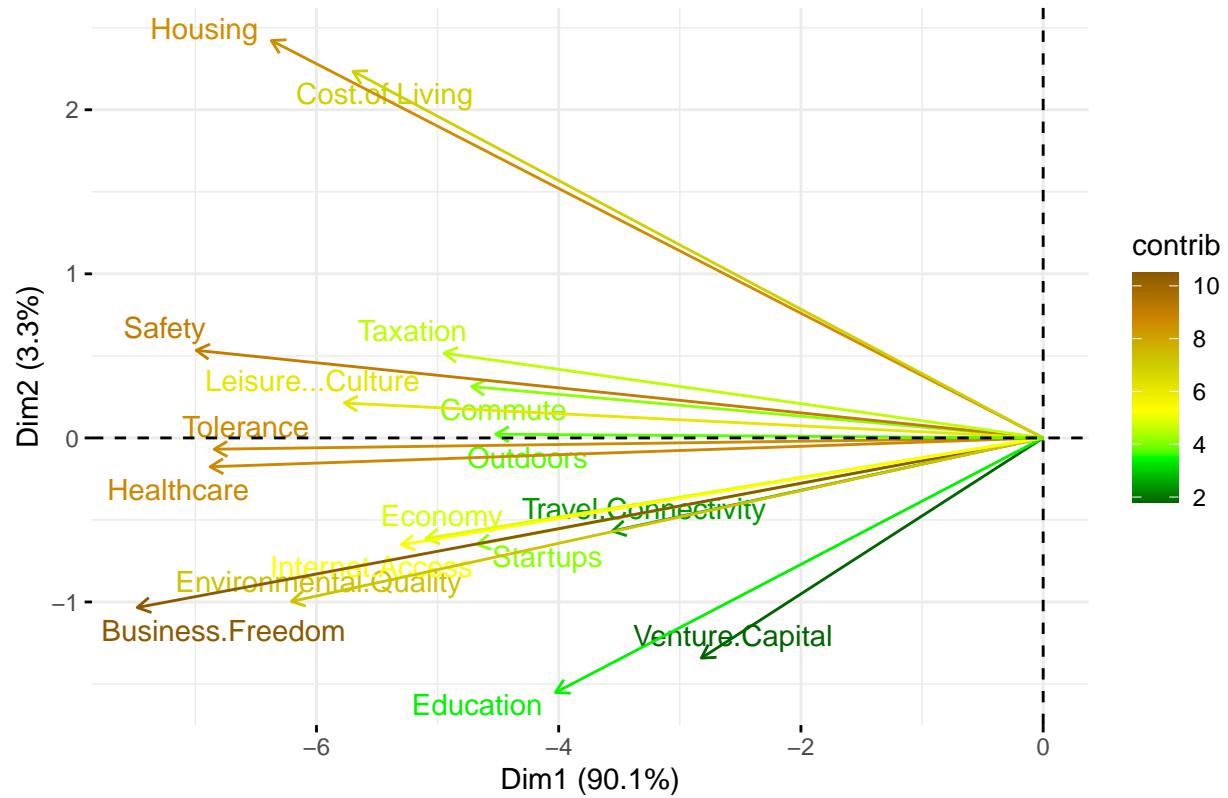
```
pca_result1 <- prcomp(data_num, scale=FALSE, center=FALSE)

fviz_pca_var(pca_result1, col.var = "contrib",
             gradient.cols = c("darkgreen", "green", "yellow", "yellow3",
                               "orange3", "orange4"),
             labelszie = 4, repel = TRUE)
```

Jak widać po wykresie 23. standaryzacja danych jest potrzebna, w przeciwnym wypadku pierwsza składowa odpowiada za zbyt dużą część wariancji (ponad 90%!), co nie pozwala na całkowicie poprawną analizę. Dodatkowo dwuwykres różni się od poprzedniego, co może nas prowadzić do innych i niekoniecznie dobrych wniosków.

Podsumowując, analiza głównych składowych pozwala wygodnie zobrazować wzajemne zależności pomiędzy cechami oraz wyodrębnić zbiory przypadków o zbliżonych profilach bądź te wyróżniające się na tle pozostałych. Dzięki temu możemy skuteczniej przyporządkować im sensowne etykiety klas, często niewidoczne w surowych danych.

Variables – PCA



Wykres 23: Dwuwykres dla dwóch pierwszych składowych - dane nieustandardyzowane

4 Zadanie 3: Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))

W tym zadaniu będziemy przeprowadzać skalowanie wielowymiarowe na zbiorze danych `titanic_train`, który zawiera dane na temat części pasażerów niesławnego Titanica, takie jak cenę, kod i klasę biletu, imiona i nazwiska, płeć czy informację, czy przetrwali katastrofę. Na początku wczytamy dane, i w razie potrzeby dokonamy poprawy poszczególnych zmiennych. Następnie dokonamy na nim operację MDS, i za pomocą diagramu Shephera ocenimy zgodność z danymi wyjściowymi. Wreszcie, ocenimy czy zastosowana na danych metoda pozwala nam na znalezienie klastry, składające się z ludzi, i czy jesteśmy w stanie znaleźć cechy, które wyróżniają członków tych grup.

4.1 Przygotowanie danych

Zbiór danych `titanic_train` zawiera **891** przypadków oraz **12** cech. Liczba brakujących danych wynosi **177**.

W oryginalnym zbiorze zmienne `Sex`, `Embarked`, `Survived` oraz `Pclass` zostały wczytane z nieprawidłowymi typami danych.

```
titanic_train <- titanic_train[titanic_train$Embarked != "", ]  
  
titanic_train$Sex <- as.factor(titanic_train$Sex)  
titanic_train$Embarked <- as.factor(titanic_train$Embarked)  
titanic_train$Survived <- factor(titanic_train$Survived,  
                                    levels = c(0, 1),  
                                    ordered = TRUE)  
  
titanic_train$Pclass <- factor(titanic_train$Pclass,  
                                 levels = c(3L, 2L, 1L),  
                                 ordered = TRUE)
```

W poniższej tabeli przedstawiono opis oraz typ zmiennych zbioru `titanic_train`.

```
df_table <- data.frame(  
  Typ = sapply(sapply(titanic_train, class), function(x) x[1]),  
  Opis = c("Unikalny identyfikator pasażera",  
          "Czy pasażer przeżył? (0 - nie, 1 - tak)",  
          "Klasa podróży (1 = najlepsza, 3 = najgorsza)",  
          "Imię i nazwisko pasażera",  
          "Płeć pasażera",  
          "Wiek pasażera w latach",  
          "Liczba rodzeństwa i/lub małżonków na pokładzie",  
          "Liczba rodziców i/lub dzieci na pokładzie",  
          "Numer biletu",  
          "Opłata za bilet",  
          "Kabina pasażera (jeśli znana)",  
          "Port zaokrętowania (C = Cherbourg, Q = Queenstown, S = Southampton)")  
  
kable(df_table, col.names = c("Zmienna", "Typ", "Opis"),  
      caption = "Opis zmiennych w zbiorze danych titanic_train")
```

Tabela 6: Opis zmiennych w zbiorze danych titanic_train

Zmienna	Typ	Opis
PassengerId	integer	Unikalny identyfikator pasażera
Survived	ordered	Czy pasażer przeżył? (0 - nie, 1 - tak)
Pclass	ordered	Klasa podróży (1 = najlepsza, 3 = najgorsza)
Name	character	Imię i nazwisko pasażera
Sex	factor	Płeć pasażera
Age	numeric	Wiek pasażera w latach
SibSp	integer	Liczba rodzeństwa i/lub małżonków na pokładzie
Parch	integer	Liczba rodziców i/lub dzieci na pokładzie
Ticket	character	Numer biletu
Fare	numeric	Opłata za bilet
Cabin	character	Kabina pasażera (jeśli znana)
Embarked	factor	Port zaokrętowania (C = Cherbourg, Q = Queenstown, S = Southampton)

Do dalszej analizy usuwamy cechy, które służą do identyfikowania pasażerów.

```
titanic_train$PassengerId <- NULL
titanic_train>Name <- NULL
titanic_train$Ticket <- NULL
titanic_train$Cabin <- NULL
```

4.2 Redukcja wymiaru na bazie MDS

W tej części wykorzystano metodę MDS do odwzorowania danych pasażerów Titanica w przestrzeni dwuwymiarowej. Celem było zobrazowanie struktury danych oraz identyfikacja klastrów przy zachowaniu możliwie wiernych odległości między przypadkami.

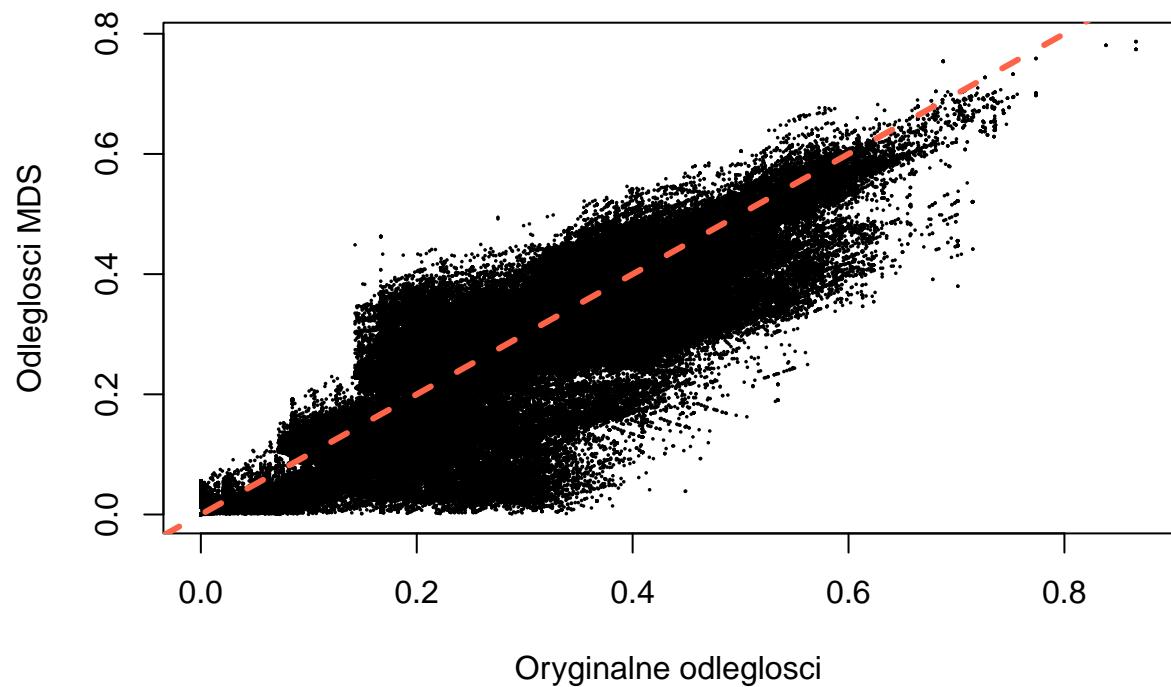
```
diss_matrix <- daisy(titanic_train[, names(titanic_train) != "Survived"],
                      metric = "gower")

wymiary <- 2
mds_result <- cmdscale(diss_matrix, k = wymiary, eig = TRUE)
points_mds <- as.data.frame(mds_result$points)
colnames(points_mds) <- c("Wymiar1", "Wymiar2")

# Wykres Sheparda
plot(diss_matrix, dist(points_mds),
      xlab = "Oryginalne odległości", ylab = "Odległości MDS", pch = 16, cex = 0.25)
abline(a=0, b=1, col = "tomato", lty = 2, lwd = 3, cex = 0.25)
```

```
odl <- max(abs(as.vector(diss_matrix)-as.vector(dist(points_mds))))
min <- min(as.vector(diss_matrix))
max <- max(as.vector(diss_matrix))
```

Wykres 24. przedstawia odwzorowanie oryginalnych, wielowymiarowych odległości w przestrzeni 2-wymiarowej. Jak widać, wartości te mieszczą się w przedziale od 0 do 0.87. Odległość punktów od prostej $y = x$ pozwala ocenić jakość odwzorowania. W tym przypadku rozbieżności są znaczne – większość punktów



Wykres 24: Diagram Sheparda dla danych pasażerów Titanica po redukcji wymiarów do przestrzeni dwuwymiarowej metodą MDS

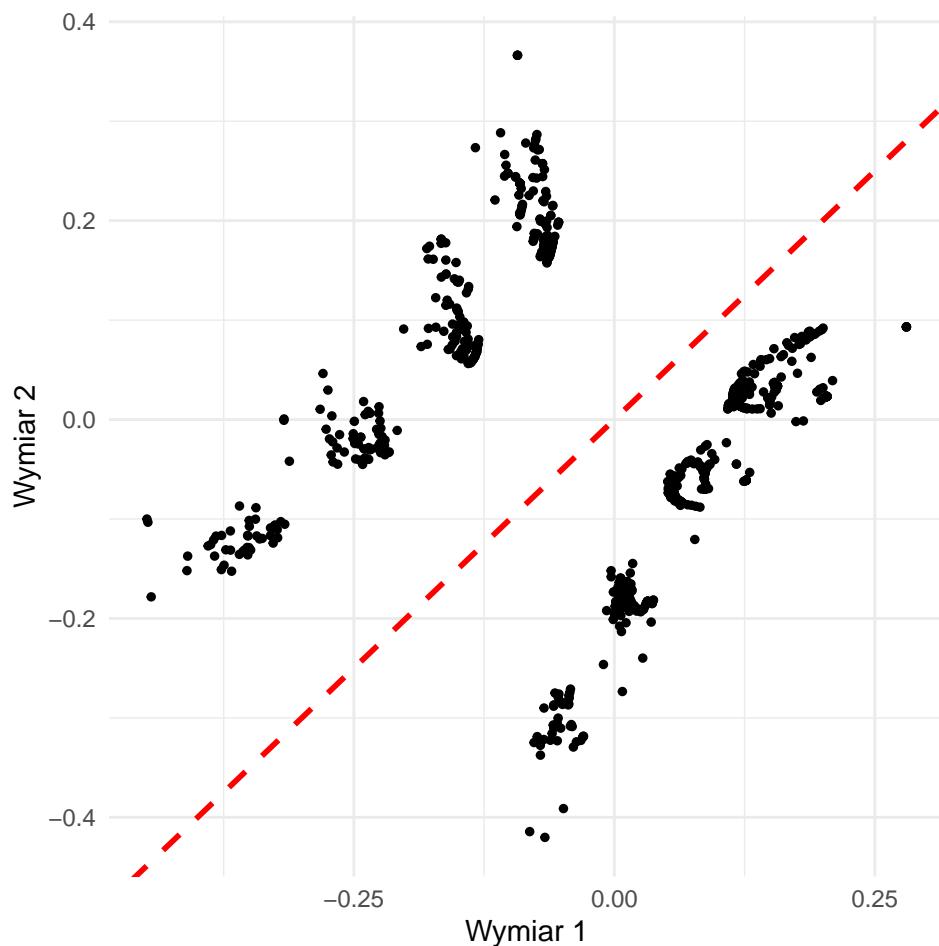
znajduje się daleko od idealnej linii odwzorowania, a maksymalne odchylenie wynosi aż 0.41, co stanowi 47.36% względem największej wartości.

Świadczy to o tym, że uzyskane odwzorowanie za pomocą MDS jest dość słabe i może nie oddawać wiernie struktury oryginalnych danych. Możliwe, że dane te mają złożoną, nielinową strukturę, której klasyczny MDS nie jest w stanie uchwycić.

4.3 Wizualizacja danych

Przejdziemy teraz do wizualizacji danych za pomocą wykresów opartych na redukcji wymiarów metodą MDS. Celem tej analizy jest próba uchwycenia i zidentyfikowania ukrytej struktury w danych.

```
ggplot(points_mds, aes(x = Wymiar1, y = Wymiar2)) +  
  geom_point(size = 1) +  
  labs(x = "Wymiar 1", y = "Wymiar 2") +  
  theme_minimal() +  
  geom_abline(slope = 1, intercept = 0, lty = 2, lwd = 1, color="red")
```



Wykres 25: MDS - Rozmieszczenie pasażerów Titanica w przestrzeni dwuwymiarowej

Na wykresie 25. widoczny jest wyraźny podział przypadków na dwa duże klastry, które znajdują się po przeciwnych stronach przekątnej $y = x$. Jeden z nich, położony powyżej przekątnej, będziemy nazywać **górnym**,

a drugi — znajdujący się poniżej — **dolnym**. Każdy z klastrów dzieli się dodatkowo na cztery mniejsze grupy, ułożone w sposób przypominający liniowy wzrost. Układ ten może sugerować istnienie ukrytych, uporządkowanych struktur w zbiorze danych. Grupy są uporządkowane sekwencyjnie — od pierwszej, znajdującej się w lewym dolnym rogu dużego klastra, do czwartej, położonej w jego prawym górnym rogu tego klastra.

```
p1 <- ggplot(points_mds, aes(x = Wymiar1, y = Wymiar2,
                               color = factor(titanic_train$Pclass))) +
  geom_point(size = 1) +
  labs(x = "Wymiar 1", y = "Wymiar 2", color = "Klasa", title = "a) Klasa") +
  scale_color_manual(values = c("1" = "tomato", "2" = "steelblue", "3" = "darkgreen"),
                      labels = c("Trzecia", "Druga", "Pierwsza")) +
  theme_minimal() +
  theme(legend.position = "bottom")

p2 <- ggplot(points_mds, aes(x = Wymiar1, y = Wymiar2,
                               color = factor(titanic_train$Sex))) +
  geom_point(size = 1) +
  labs(x = "Wymiar 1", y = "Wymiar 2", color = "Płeć", title = "b) Płeć") +
  scale_color_manual(values = c("male" = "tomato", "female" = "steelblue"),
                      labels = c("Kobieta", "Mężczyzna")) +
  theme_minimal() +
  theme(legend.position = "bottom")

p3 <- ggplot(points_mds, aes(x = Wymiar1, y = Wymiar2,
                               color = factor(titanic_train$Embarked))) +
  geom_point(size = 1) +
  labs(x = "Wymiar 1", y = "Wymiar 2", color = "Miasto", title = "c) Port zaokrętowania") +
  scale_color_manual(values = c("C" = "tomato", "S" = "steelblue", "Q" = "darkgreen"),
                      labels = c("Cherbourg", "Queenstown", "Southampton")) +
  theme_minimal() +
  theme(legend.position = "bottom")

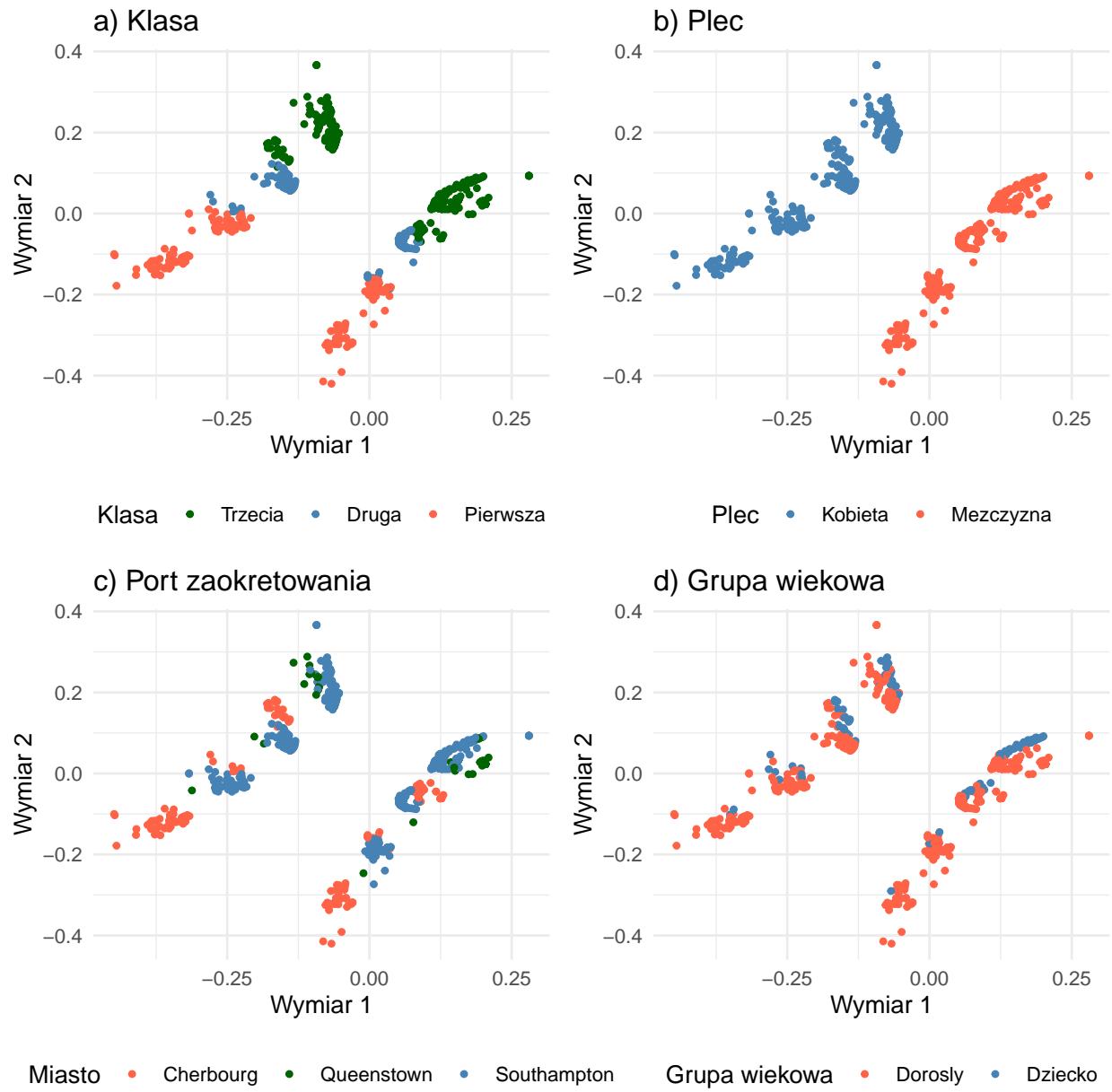
titanic.train$Child <- ifelse(!is.na(titanic_train$Age) & titanic_train$Age < 18, 1, 0)

p4 <- ggplot(points_mds, aes(x = Wymiar1, y = Wymiar2,
                               color = factor(titanic_train$Child))) +
  geom_point(size = 1) +
  labs(x = "Wymiar 1", y = "Wymiar 2", color = "Grupa wiekowa", title = "d) Grupa wiekowa") +
  scale_color_manual(values = c("0" = "tomato", "1" = "steelblue"),
                      labels = c("Dorosły", "Dziecko")) +
  theme_minimal() +
  theme(legend.position = "bottom")

grid.arrange(p1, p2, p3, p4, nrow = 2)
```

Na wykresie 26a) przedstawiono udział klas biletów w poszczególnych grupach. Pierwsze grupy składają się wyłącznie z pasażerów podróżujących klasą pierwszą. W drugich grupach dominują bilety klasy pierwszej, lecz pojawiają się także nieliczne przypadki klasy drugiej. Trzecie grupy cechuje bardziej zrównoważony udział klas drugiej i trzeciej, z lekką przewagą klasy drugiej. Natomiast ostatnia grupa obejmuje wyłącznie pasażerów z biletami klasy trzeciej.

Wykres 27b) przedstawia idealny podział ze względu na płeć. Górnny klaster obejmuje wyłącznie kobiety, natomiast dolny — wyłącznie mężczyzn. Taki układ sugeruje, że płeć jest czynnikiem różnicującym strukturę danych na klaster **górnny** i **dolny**.

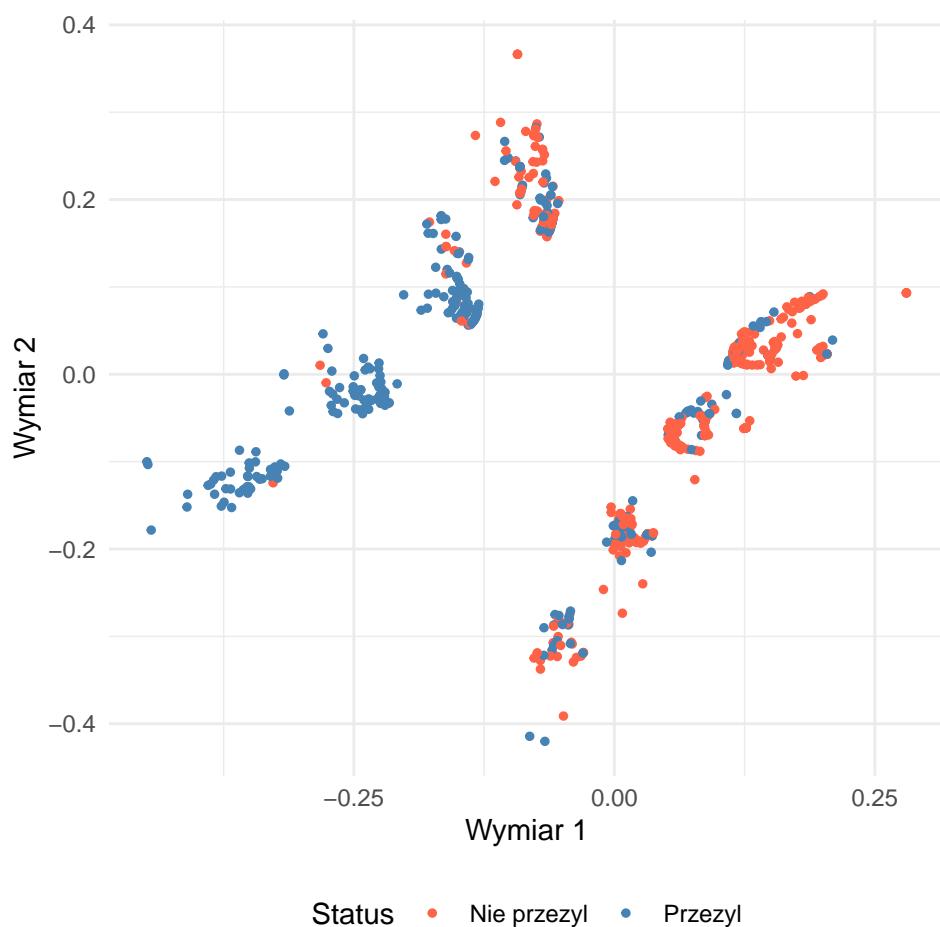


Wykres 26: MDS - Rozmieszczenie pasażerów Titanica w przestrzeni dwuwymiarowej z podziałem na a) Klasę b) Płeć c) Port zaokrętowania d) Grupę wiekową

Wykres 27c) przedstawia miejsca, z których pasażerowie dołączali na pokład Titanica. Pierwsze grupy obejmują głównie osoby, które wsiadły w Cherbourg we Francji. Grupy drugie i trzecie zdominowane są przez pasażerów z angielskiego Southampton, z mniejszym udziałem osób z Cherbourgą oraz pojedynczymi przypadkami z Queenstown. Natomiast czwarte grupy zawierają głównie pasażerów z Southampton oraz w mniejszym stopniu z irlandzkiego Queenstown.

Wykres 27d) pokazuje, że w pierwszych grupach występują jedynie pojedyncze przypadki dzieci, a ich liczba stopniowo rośnie w kolejnych grupach. Mimo tego w żadnej z grup dzieci nie stanowią większości, co wskazuje, że zdecydowana większość danych dotyczy osób dorosłych. Zauważalna jest jednak wyraźna tendencja wzrostu liczby dzieci w miarę przechodzenia do późniejszych grup.

```
ggplot(points_mds, aes(x = Wymiar1, y = Wymiar2,
                        color = factor(titanic_train$Survived))) +
  geom_point(size = 1) +
  labs(x = "Wymiar 1", y = "Wymiar 2", color = "Status") +
  scale_color_manual(values = c("0" = "tomato", "1" = "steelblue"),
                      labels = c("Nie przeżył", "Przeżył")) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Wykres 27: MDS - Rozmieszczenie pasażerów Titanica w przestrzeni dwuwymiarowej z podziałem na status przeżycia

Na wykresie 27. przedstawiono informacje dotyczące ofiar katastrofy Titanica. W pierwszym klastrze dominują osoby, które przeżyły, jednak w kolejnych grupach ich liczba systematycznie maleje na rzecz wzrostu udziału ofiar. Drugi klaszter również wykazuje rosnący udział osób zmarłych, lecz struktura grup jest nieco inna — pierwsza grupa zawiera zbliżoną liczbę ocalałych i ofiar, a w każdej następnej coraz większą część stanowią osoby, które nie przeżyły.

Na podstawie wykresów można wskazać kluczowe czynniki różnicujące osoby ocalone i ofiary katastrofy:

- Zginęło znacznie więcej mężczyzn niż kobiet, co potwierdza znane fakty historyczne.
- Kobiety z klasy pierwszej i drugiej w większości przeżyły. Kobiety z klasy trzeciej w większości zginęły.
- Osoby z wyższych klas miały większe szanse na przetrwanie.
- Osoby z klasy drugiej i trzeciej, które przeżyły stanowiły w większości dzieci.
- Wśród pasażerów pierwszej klasy zaobserwowano większe różnice w zależności od miejsca wejścia na pokład – francuskiego Cherbourga lub angielskiego Southampton – niż miało to miejsce w przypadku pasażerów klas niższych.

4.3.1 * Rodzina Sage – symbol nierówności klasowej na Titanicu

Najbardziej odstające punkty na wykresie – te położone najdalej na prawo i w górę – należą do interesującego przypadku rodziny Sage.

John i Annie Sage wraz z dziewięciorgiem dzieci w wieku od kilku do kilkunastu lat wywodzili się z Peterborough w Anglii i stanowili największą rodzinę na pokładzie Titanica. Ich podróż nie była zwykłym urlopem, lecz przeprowadzką całego gospodarstwa przez Ocean Atlantycki: planowali osiedlić się w Jacksonville na Florydzie i prowadzić tam pensjonat.

Niestety żaden z członków rodziny nie przeżył katastrofy. Jako pasażerowie trzeciej klasy nie mieli równych szans dostępu do wyższych pokładów i szalup ratunkowych, co znacząco zmniejszyło ich szanse na uratowanie życia.

Rodzina Sage bywa przywoływana w literaturze historycznej jako symbol niesprawiedliwości klasowej na Titanicu – ich los doskonale ukazuje, jak struktura okrętu i procedury ewakuacyjne faworyzowały podróżnych z wyższych szczebli społecznych.

4.4 Podsumowanie

W analizie danych ze zbioru *titanic_train* zastosowano metodę MDS w celu redukcji wymiarowości oraz wizualizacji danych w dwóch wymiarach. Przed przekształceniem przeprowadzono wstępne czyszczenie — usunięto dwa przypadki z brakującą informacją o miejscu zaokrętowania (*Embarked*). Stanowiły one jedynie 0.22% całego zbioru, więc ich usunięcie nie wpłynęło istotnie na ogólny obraz danych. Ponadto odpowiednie zmienne zostały poprawnie zaklasyfikowane jako kategoryczne.

Wyniki odwzorowania nie były idealne — dla wielu punktów zaobserwowano znaczące różnice między oryginalnymi a odwzorowanymi odległościami, co może świadczyć o ograniczeniach metody w oddaniu rzeczywistej, nieliniowej struktury danych. Niemniej jednak w przekształconej przestrzeni ujawniły się wyraźne wzorce: dwa główne klastry, z których każdy dzielił się na cztery mniejsze grupy.

Analiza rozmieszczenia punktów pozwoliła wyodrębnić cechy charakterystyczne dla ofiar katastrofy. Kobiety, dzieci oraz pasażerowie wyższych klas częściej należeli do grup, które przeżyły — co jest zgodne z dobrze udokumentowanym przebiegiem wydarzeń na Titanicu.

Wśród przypadków szczególnie odstających uwagę zwraca historia rodziny Sage — najliczniejszej rodziny na pokładzie, która w całości zginęła. Ich dramatyczna historia stała się symbolem nierówności społecznych, jakie ujawniły się podczas katastrofy.

5 Wnioski

Podsumowując, w raporcie rozważane były następujące pojęcia:

- Dyskretyzacja cech ciągłych
- Analiza składowych głównych (PCA)
- Skalowanie wielowymiarowe (MDS)

Dyskretyzacja zastosowana w analizie zbioru *iris* umożliwiła identyfikację cech najlepiej odróżniających poszczególne gatunki kwiatów. Wykorzystane metody — *k-srednich*, *równej częstotliwości* oraz *równych przedziałów* — wykazały różną skuteczność klasyfikacyjną, przy czym dwie pierwsze dały najlepsze wyniki. Przetestowano również ręczne definiowanie przedziałów, jednak podejście to okazuje się mało praktyczne przy większej liczbie zmiennych lub przypadków.

W przypadku danych dotyczących jakości życia w miastach analiza **głównych składowych (PCA)** pozwoliła na wyjaśnienie znacznej części wariancji oryginalnych danych, umożliwiając jednocześnie redukcję wymiarowości. Dzięki temu możliwe było wyróżnienie grup miast o podobnych cechach — zarówno na poziomie kontynentów, jak i niekiedy konkretnych krajów. Interpretacja biplotu ukazała również wzajemne zależności między zmiennymi, a zmniejszenie wymiaru z 17 do 2 znacznie ułatwiło analizę klastrów.

Z kolei **skalowanie wielowymiarowe (MDS)**, podobnie jak PCA, umożliwiło wizualizację danych pasażerów Titanica w dwóch wymiarach. Pomimo pewnych zniekształceń odwzorowanych odległości (potwierdzonych przez diagram Sheperda), analiza ujawniła wyraźne klastry oraz wzorce związane z cechami takimi jak płeć, klasa biletu czy miejsce zaokrętowania. Pozycja punktów w przestrzeni oraz ich grupowanie pozwoliły na odczytanie istotnych informacji o pasażerach i zrozumienie struktury danych w nowej przestrzeni.