

# Raport 2

## Eksploracja danych

Olga Foriasz 277529, Szymon Smoła 282252

2025-04-30

## Spis treści

<b>1</b>	<b>Dyskretyzacja (przedziałowanie) cech ciągłych</b>	<b>1</b>
1.1	Krótki opis zagadnienia . . . . .	1
1.2	Opis eksperymentów/analiz . . . . .	1
1.3	Wyniki . . . . .	2
1.4	Podsumowanie . . . . .	7
<b>2</b>	<b>Analiza składowych głównych - metoda PCA</b>	<b>7</b>
2.1	Krótki opis zagadnienia . . . . .	7
2.2	Przygotowanie danych . . . . .	7
2.3	Wyznaczanie składowych głównych . . . . .	9
2.4	Zmienność odpowiadająca poszczególnym składowym . . . . .	10
2.5	Wizualizacja danych wielowymiarowych . . . . .	12
2.6	Korelacja zmiennych . . . . .	13
2.7	Wnioski końcowe . . . . .	15
<b>3</b>	<b>Skalowanie wielowymiarowe</b>	<b>16</b>
3.1	Krótki opis zagadnienia . . . . .	16
3.2	Wyniki . . . . .	18
3.3	Podsumowanie wniosków . . . . .	22

## 1 Dyskretyzacja (przedziałowanie) cech ciągłych

### 1.1 Krótki opis zagadnienia

#### 1.1.1 Irysy

W pierwszej części raportu przeprowadzimy analizę danych dotyczących irysów, skupiając się na dyskretyzacji cech ciągłych. Dane zawierają wymiary kwiatów trzech gatunków. Główne pytania badawcze to:

1. Jaka cecha wyróżnia się najlepszą i najgorszą zdolnością dyskryminacji
2. Który algorytm dyskretyzacji jest najbardziej skuteczny

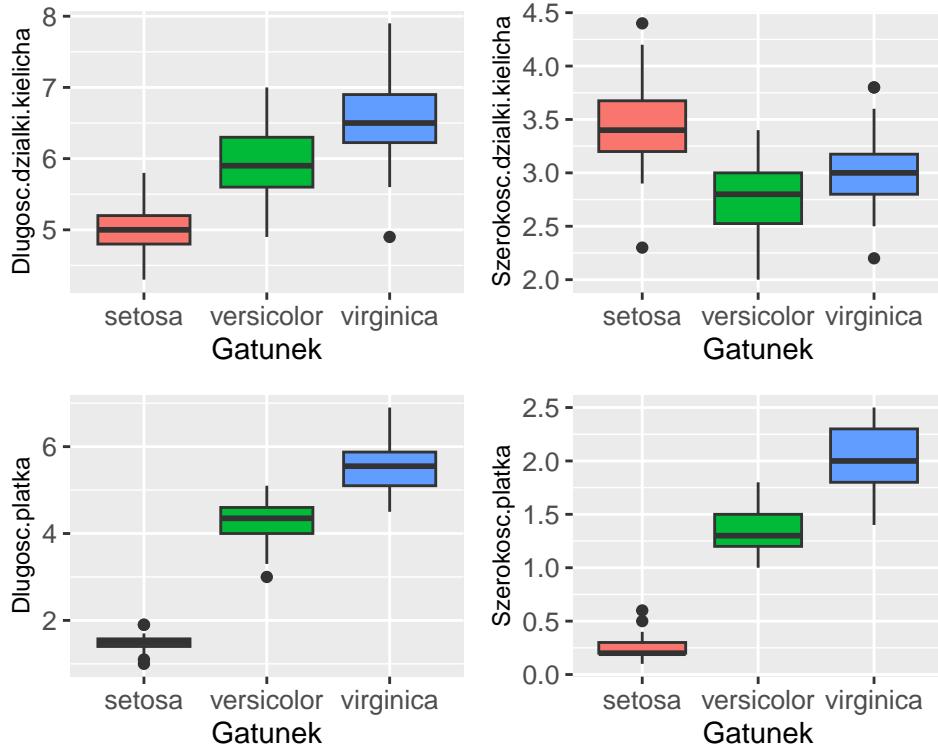
## 1.2 Opis eksperymentów/analiz

### 1.2.1 Wykorzystane narzędzia

- *Metody statystyczne:*
  - Wizualizacja rozkładów
  - Statystyki opisowe
- *Metody dyskretyzacji nienadzorowanej:*
  - Equal frequency
  - Equal width
  - k-means clustering
  - Przedziały zdefinowane przez użytkownika
- *Parametry:*
  - Liczba przedziałów:  $K = 3$

## 1.3 Wyniki

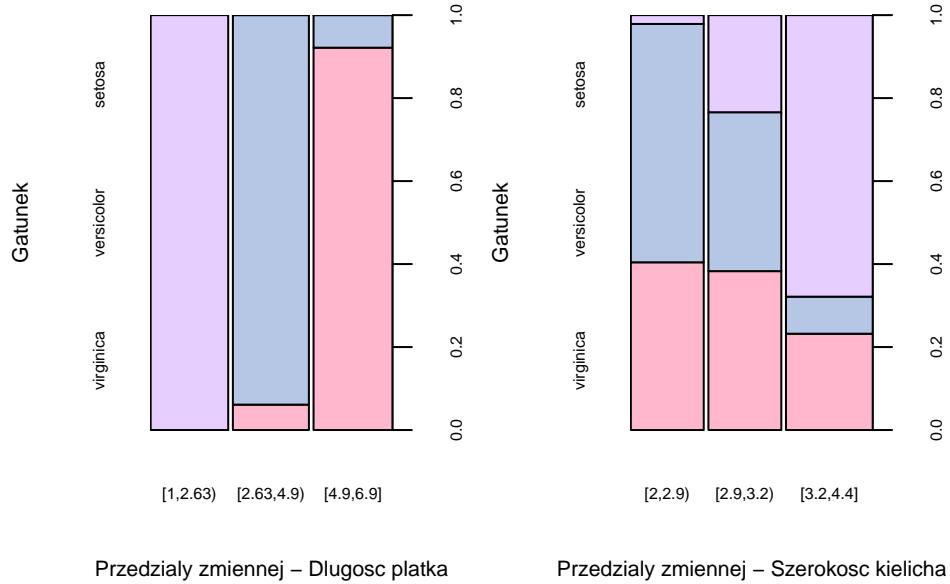
### 1.3.1 Wybór cechy



Rysunek 1: Wykresy pułapkowe dla cech charakteryzujących

Na podstawie Rysunku 1 wynika, że długość płatka najlepiej separuje gatunki, zaś szerokość działki kielicha wykazuje najsłabszą zdolność dyskryminacji. Jesteśmy w stanie to wywnioskować ze względu na niewielkie nałożenie się wartości wykresów przy wykresie długości płatków oraz widoczne nałożenie się wartości wykresów pułapkowych w przypadku szerokości działek kielicha.

### 1.3.2 Porównanie nienadzorowanych metod dyskretyzacji



Rysunek 2: Dyskretyzacja oparta na jednakowej częstotliwości - wykres mozaikowy

Tabela 1: Długość płatka - równa częstotliwość

Przedział	setosa	versicolor	virginica
[1,2.63)	50 (100%)	0 (0%)	0 (0%)
[2.63,4.9)	0 (0%)	46 (93.9%)	3 (6.1%)
[4.9,6.9]	0 (0%)	4 (7.8%)	47 (92.2%)

Tabela 2: Szerokość kielicha - równa częstotliwość

Przedział	setosa	versicolor	virginica
[2,2.9)	27 (57.4%)	1 (2.1%)	19 (40.4%)
[2.9,3.2)	18 (38.3%)	11 (23.4%)	18 (38.3%)
[3.2,4.4]	5 (8.9%)	38 (67.9%)	13 (23.2%)

```
matchClasses(x.tab.equal.freq)
```

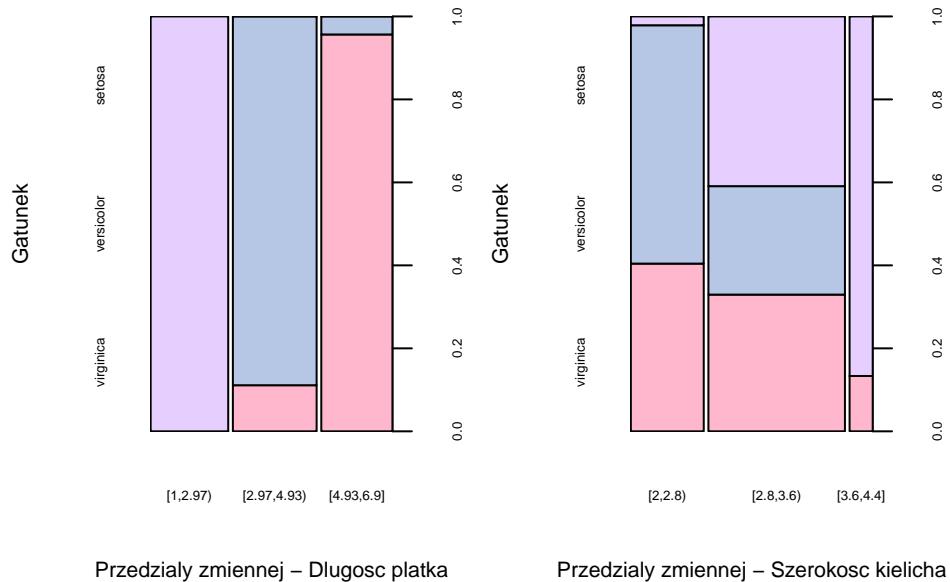
```
## Cases in matched pairs: 95.33 %
##      [1,2.63)    [2.63,4.9)    [4.9,6.9]
## "setosa" "versicolor" "virginica"
```

```
matchClasses(y.tab.equal.freq)
```

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

Z powyższych tabel 1 i 2 oraz z rysunku 2 wynika, że dla długości płatka pierwszy przedział wyznaczony metodą **równych częstotliwości** zawiera wyłącznie rekordy odpowiedzialne za gatunek

setosa, zaś w pozostałych przedziałach widać lekkie wymieszanie gatunków. Obliczony procent zdolności grupowania metodą **równych częstości** dla cechy odpowiedzialnej, za długość płatka wynosi 95.33%. Jednak niestety dla szerokości działki kielicha, metoda ta nie działa perfekcyjnie i uzyskała 55.33% zdolności grupowania.



Rysunek 3: Dyskretyzacja oparta na jednakowej szerokości przedziałów - wykres mozaikowy

Tabela 3: Długość płatka - równe przedziały

Przedział	setosa	versicolor	virginica
[1,2.97)	50 (100%)	0 (0%)	0 (0%)
[2.97,4.93)	0 (0%)	48 (88.9%)	6 (11.1%)
[4.93,6.9]	0 (0%)	2 (4.3%)	44 (95.7%)

Tabela 4: Szerokość działki - równe przedziały

Przedział	setosa	versicolor	virginica
[2,2.8)	1 (2.1%)	27 (57.4%)	19 (40.4%)
[2.8,3.6)	36 (40.9%)	23 (26.1%)	29 (33%)
[3.6,4.4]	13 (86.7%)	0 (0%)	2 (13.3%)

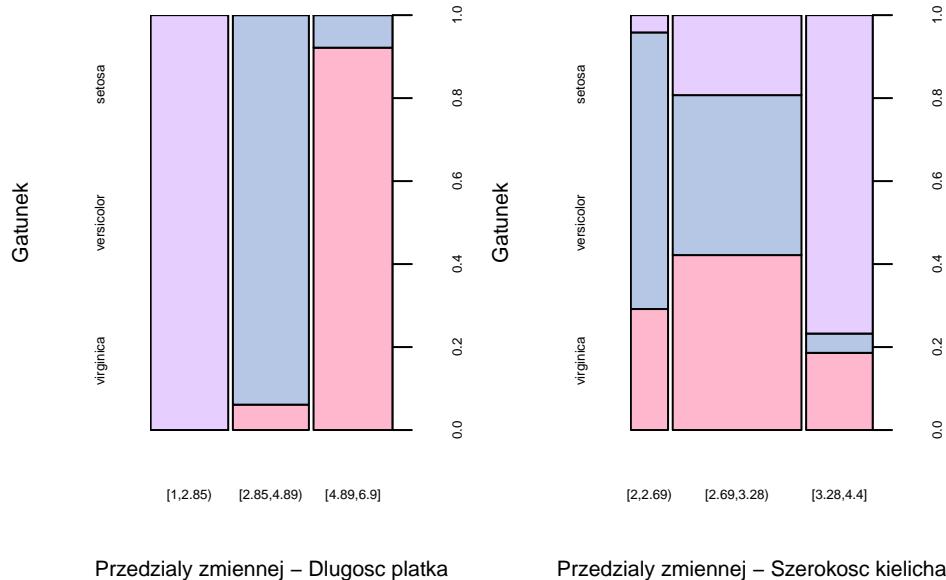
```
matchClasses(x.tab.equal.width)
```

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

```
matchClasses(y.tab.equal.width)
```

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

Analizując wyniki z tabel 3 i 4 oraz z rysunku 3 w analogiczny sposób jak w poprzedniej metodzie. Zauważamy, że metoda **równych szerokości** już nie tak dokładnie grupuje przedziałami nasze dane. Zdolność dyskryminacyjna dla zmiennej odpowiedzialnej za długość płatka wynosi 94.67%, a dla tej odpowiadającej szerokości kielicha wynosi 50.67%. W obu przypadkach jest to gorszy wynik, niż przy metodzie **równych częstotliwości**.



Rysunek 4: Dyskretyzacja oparta na algorytmie k-means - wykres mozaikowy

Tabela 5: Długość płatka - k-means

Przedział	setosa	versicolor	virginica
[1,2.85)	50 (100%)	0 (0%)	0 (0%)
[2.85,4.89)	0 (0%)	46 (93.9%)	3 (6.1%)
[4.89,6.9]	0 (0%)	4 (7.8%)	47 (92.2%)

Tabela 6: Szerokość kielicha - k-means

Przedział	setosa	versicolor	virginica
[2,2.69)	1 (4.2%)	16 (66.7%)	7 (29.2%)
[2.69,3.28)	16 (19.3%)	32 (38.6%)	35 (42.2%)
[3.28,4.4]	33 (76.7%)	2 (4.7%)	8 (18.6%)

```
set.seed(220) # pod stałe wyniki w raporcie

matchClasses(x.tab.disc.k.means)

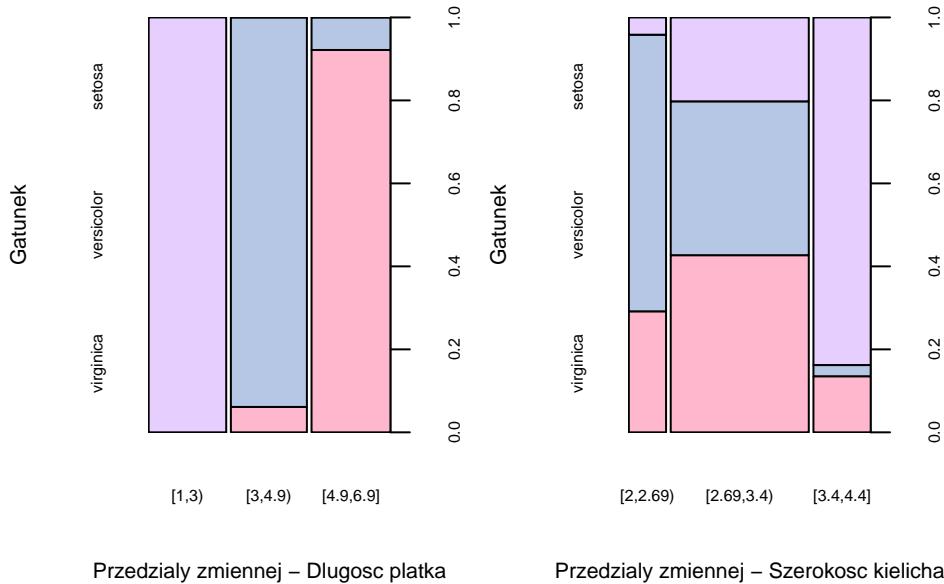
## Cases in matched pairs: 95.33 %

##      [1,2.85)  [2.85,4.89)  [4.89,6.9]
## "setosa" "versicolor" "virginica"
```

```
matchClasses(y.tab.disc.k.means)
```

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

Przyglądając się tabelom 5 i 6 oraz rysunkowi 4 zauważamy, że metoda **k-means clustering** daje podobne rezultaty co metoda **równych częstości**. W wypadku długości płatka rysunek 4 wygląda identycznie jak rysunek 2. Inną sytuację obserwujemy w przypadku wykresu cechy odpowiedzialnej za szerokość działki kielicha. W tym wypadku metoda **k-means clustering** wykazuje się lepszą zdolnością grupującą niż powyżej sprawdzane metody. Jej współczynnik wynosi 56% (dla stałego seeda równego 220)



Rysunek 5: Dyskretyzacja oparta na przedziałach zadanych przez użytkownika - wykres mozaikowy

Tabela 7: Długość płatka - własne przedziały

Przedział	setosa	versicolor	virginica
[1,3)	50 (100%)	0 (0%)	0 (0%)
[3,4.9)	0 (0%)	46 (93.9%)	3 (6.1%)
[4.9,6.9]	0 (0%)	4 (7.8%)	47 (92.2%)

Tabela 8: Szerokość działki - własne przedziały

Przedział	setosa	versicolor	virginica
[2,2.69)	1 (4.2%)	16 (66.7%)	7 (29.2%)
[2.69,3.4)	18 (20.2%)	33 (37.1%)	38 (42.7%)
[3.4,4.4]	31 (83.8%)	1 (2.7%)	5 (13.5%)

```

matchClasses(x.tab.user)

## Cases in matched pairs: 95.33 %
##      [1,3)      [3,4.9)      [4.9,6.9]
## "setosa" "versicolor" "virginica"

matchClasses(y.tab.user)

## Cases in matched pairs: 56.67 %
##      [2,2.69)      [2.69,3.4)      [3.4,4.4]
## "versicolor" "virginica"      "setosa"

```

ykorzystując możliwość ustawienia własnych przedziałów w tabelach 7 i 8 oraz na rysunku 5 udało nam się uzyskać możliwie jak najlepsze zdolności grupowania cech długości płatka i szerokości działki kielicha. Współczynniki zdolności dyskryminacyjnych podanych cech odpowiednio wyniosły 95.33% dla długości płatka oraz 56.67% dla szerokości działki kielicha.

## 1.4 Podsumowanie

Najlepszą zdolność dyskryminacyjną w przypadku zmiennej **Długość.Płatka** wykazały się metody **równych częstości i k-means clustering**, które osiągnęły identyczne wyniki (95.33%). W wypadku zmiennej **Szerokość.Działki.Kielicha** najlepsza okazała się być metoda **k-means clustering** osiągając zdolność na poziomie 56%. Można jednak uzyskać jeszcze lepszy wynik (na poziomie 56.67%) ręcznie ustalając przedziały grupowania.

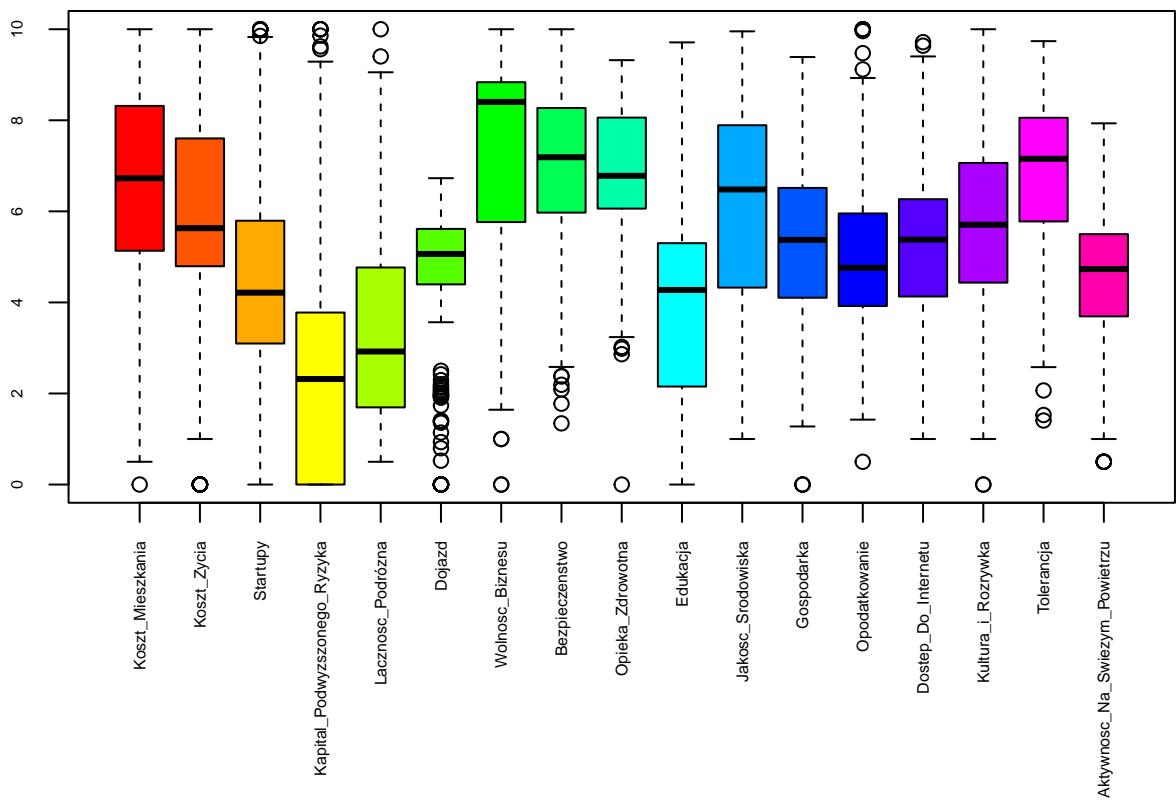
# 2 Analiza składowych głównych - metoda PCA

## 2.1 Krótki opis zagadnienia

Analizowany zbiór danych zawiera wskaźniki jakości życia dla wybranych miast na całym świecie. Dane pochodzą z ze strony Kaggle (źródło: <https://www.kaggle.com/orhankaramanco/de/city-quality-of-life-dataset>) i obejmują różne kategorie, takie jak: bezpieczeństwo, opieka zdrowotna, jakość powietrza, koszty życia, infrastruktura, poziom szczęścia czy dostęp do usług cyfrowych. Celem analizy tego zbioru może być porównanie miast pod względem warunków życia, identyfikacja podobieństw między miastami z różnych kontynentów oraz wizualizacja przestrzenna różnic za pomocą technik takich jak analiza głównych składowych (PCA).

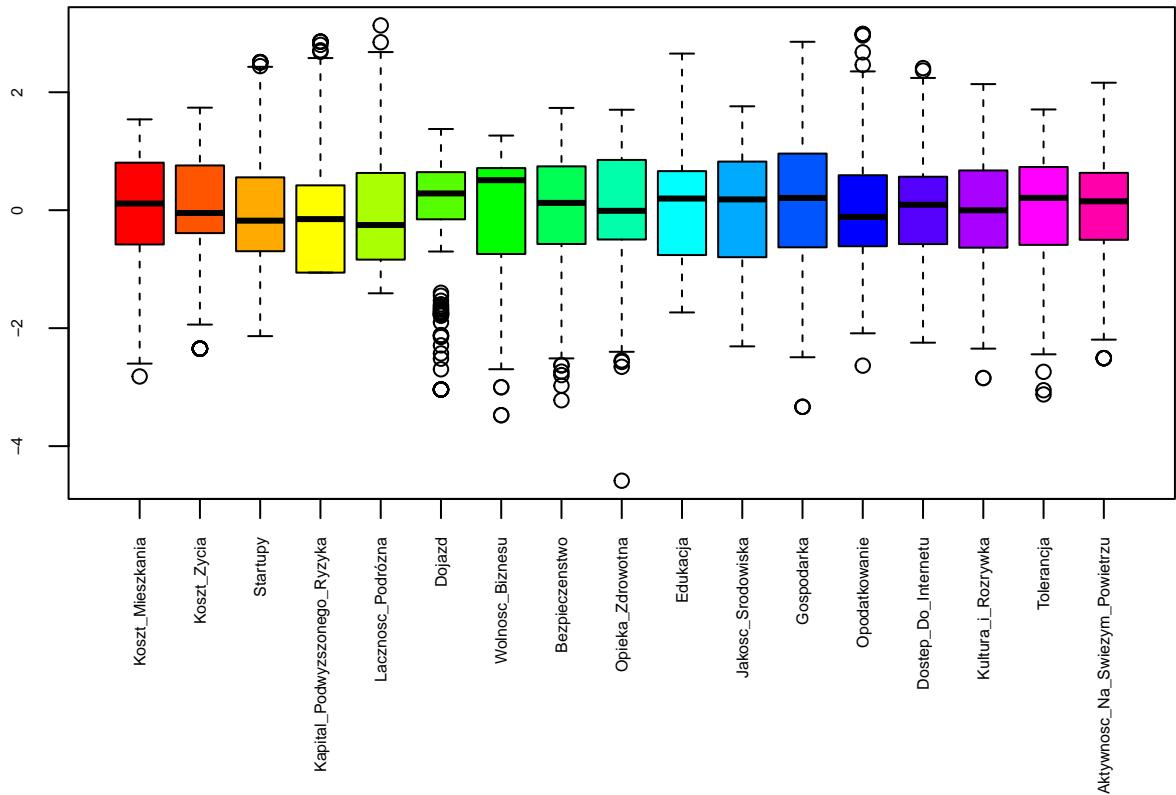
## 2.2 Przygotowanie danych

Analizę zaczniemy od przygotowania danych i sprawdzenia czy wymagana jest standaryzacja.



Rysunek 6: Wykresy pudełkowe

Na Rysunku 6 ukazany jest wykres pudełkowych danych przed standaryzacją. Jak da się zauważyć, niektóre dane mają większą zmienność od innych. Z tego powodu warto zastosować standaryzację, żeby jedne dane nie dominowały nad pozostałymi.



Rysunek 7: Wykresy pudełkowe

Na Rysunku 7 przedstawiony jest wykres zmiennych po standaryzacji.

## 2.3 Wyznaczanie składowych głównych

Tabela 9: Wektory ładunków (PC1, PC2 i PC3)

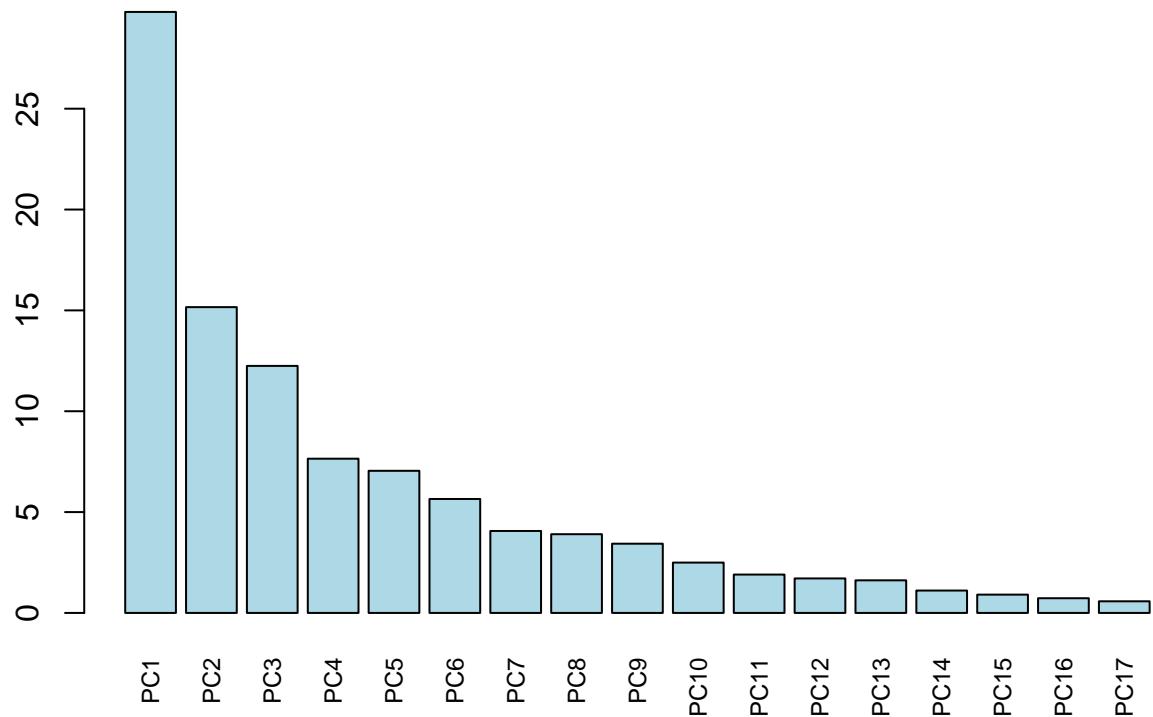
	PC1	PC2	PC3
Koszt_Mieszkania	0.308	0.053	-0.314
Koszt_Zycia	0.260	-0.176	-0.331
Startupy	-0.180	-0.483	0.006
Kapitał_Podwyższonego_Rzyzka	-0.237	-0.427	0.015
Łączność_Podróżna	-0.209	-0.135	-0.340
Dojazd	-0.114	0.026	-0.506
Wolność_Biznesu	-0.377	0.098	0.024
Bezpieczeństwo	-0.039	0.287	-0.333
Opieka_Zdrowotna	-0.280	0.242	-0.281
Edukacja	-0.403	-0.049	-0.074
Jakość_Środowiska	-0.326	0.253	0.054
Gospodarka	-0.273	-0.074	0.309
Opodatkowanie	0.026	0.107	-0.020
Dostęp_Do_Internetu	-0.276	0.023	0.028
Kultura_i_Rozrywka	-0.074	-0.365	-0.305
Tolerancja	-0.190	0.355	-0.103
Aktywność_Na_Świeżym_Powietrzu	-0.092	-0.193	-0.149

Tabela 9 przedstawia wektory ładunków kolejno dla PC1, PC2 oraz PC3 Z powyższych danych możemy zauważać następujące wnioski dotyczące kilku pierwszych wektorów ładunków:

- 1-szy wektor ładunku przypisuje największą wagę zmiennej Edukacja, a na drugim miejscu zmiennej Wolność\_Biznesu. PC1 możemy powiązać z gospodarką danego kraju oraz jej rozwojem.
- 2-gi wektor ładunku największy wkład przypisuje zmiennej Startupy, następnie zmiennej Kapitał\_Podwyższonego\_Rzyzka. Możemy interpretować PC2 jako kontrast nowoczesnego społeczeństwa z takim, w którym większą wagę przykłada się do wartości takich jak np. bezpieczeństwo.
- 3-ci wektor ładunku największą wagę przypisuje zmiennym Dojazd oraz Łączność\_Podróżna. Na trzecim miejscu znajduje się zmienna Koszt\_Zycia. Zatem PC3 wiąże dostępność transportu oraz możliwość dojazdu ze stanem materialnym i kosztami życia. Można to interpretować jako podział względem poziomu życia i codziennej wygody.

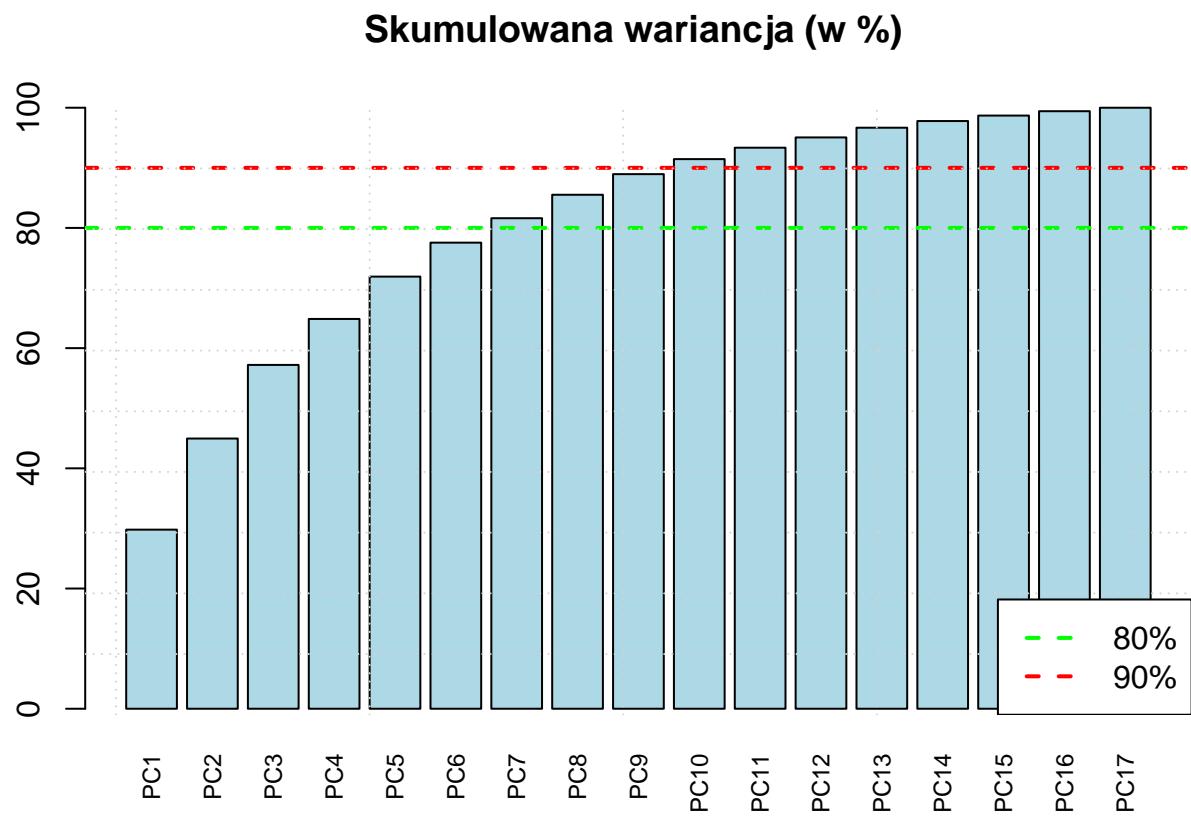
## 2.4 Zmienna odpowiadająca poszczególnym składowym

**Wariancja odpowiadająca poszczególnym składowym (w %)**



Rysunek 8: Wykres wariancji

Na Rysunku 8 przedstawia wariancję dla poszczególnych zmiennych składowych, przedstawioną w procentach. Rysunek ten jest wizualną odpowiedzią na pytanie: jaki procent zmienności odpowiada poszczególnym składowym głównym.

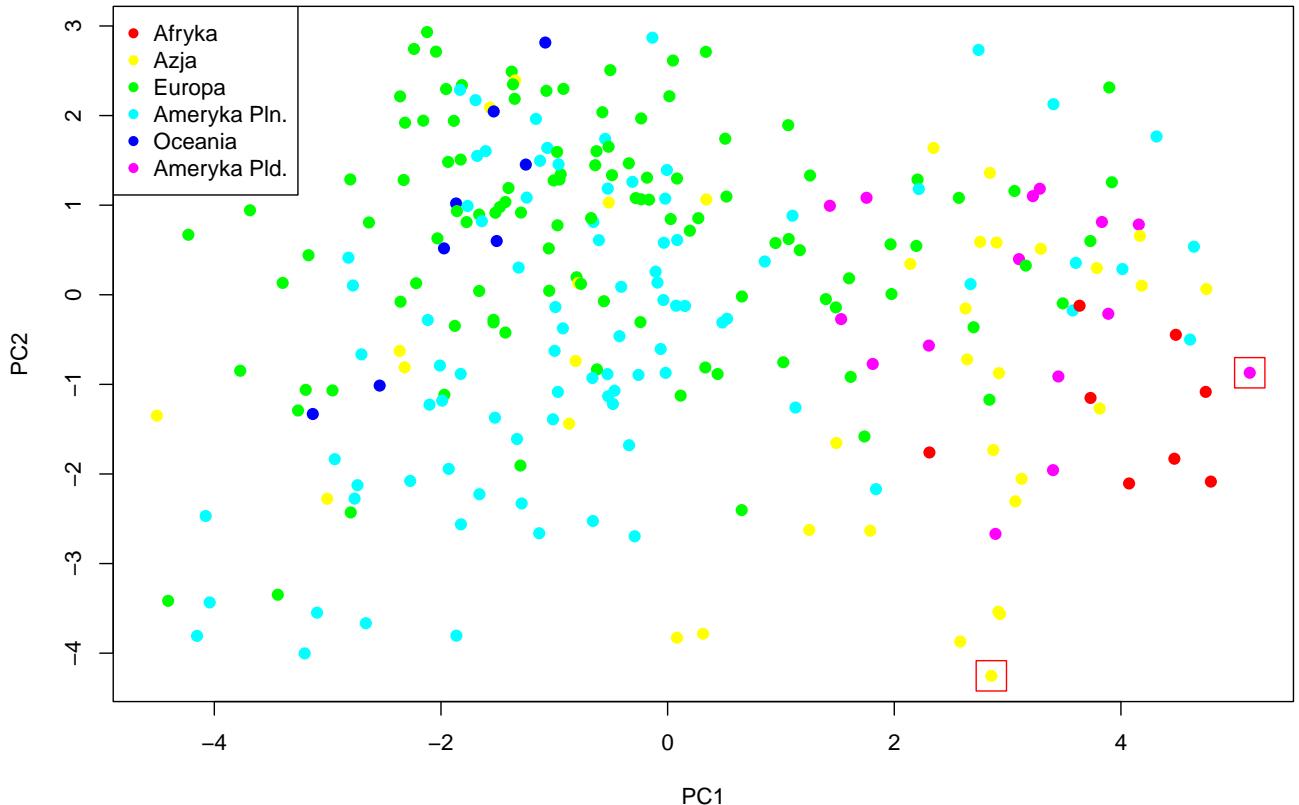


Rysunek 9: Wariancja skumulowana

Wykorzystując Rysunek 9 można zauważyc, że do wyjaśnienia 80% całkowitej zmienności porzebujemy 7 pierwszych składowych (tj.PC1-PC7). Natomiast analogicznie korzystając z wykresu, do wyjaśnienia 90% potrzebne jest 10 pierwszych składowych (tj.PC1-PC10).

## 2.5 Wizualizacja danych wielowymiarowych

Dane – wykres rozrzutu 2D



Rysunek 10: Wykres rozrzutu 2D

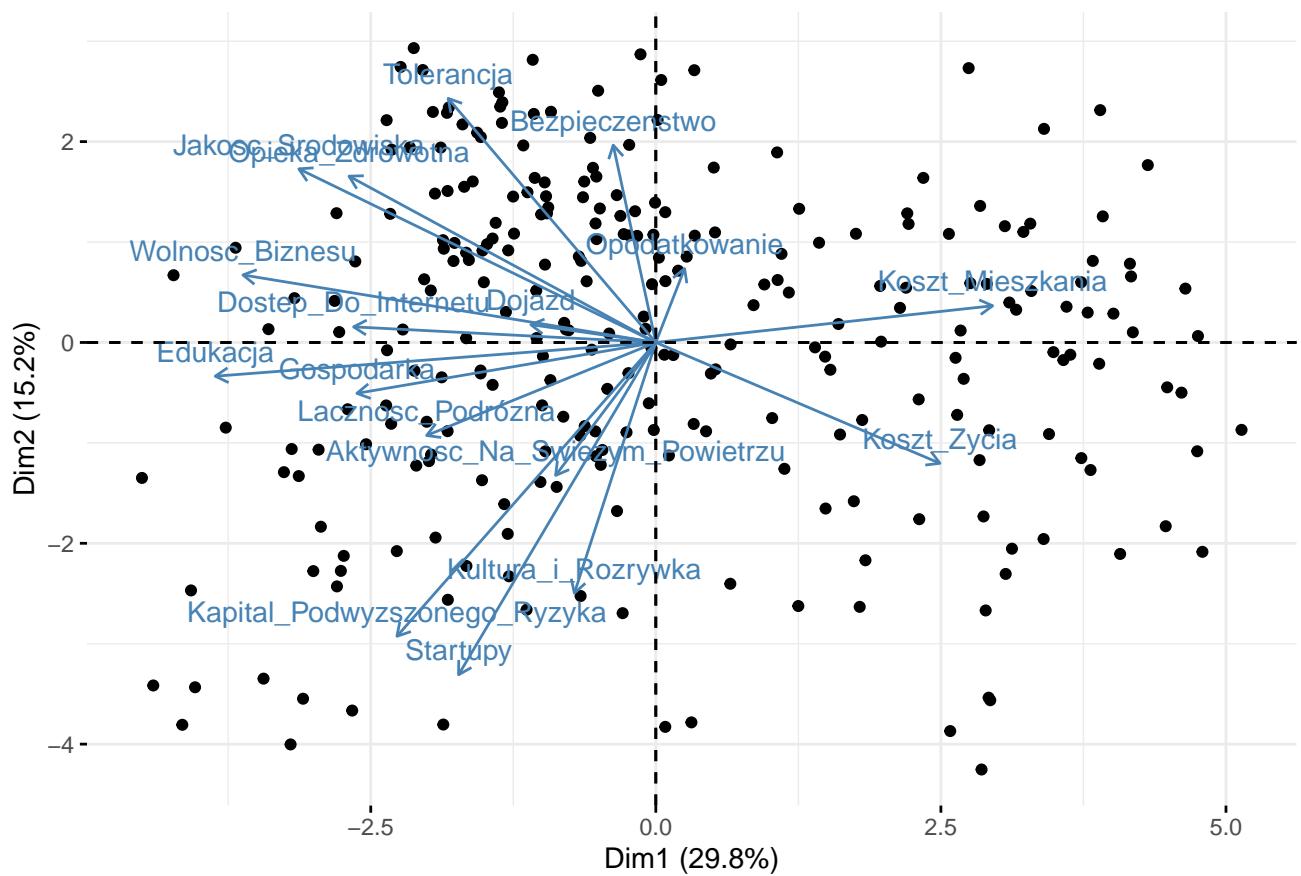
Na podstawie Rysunku 10 można zaobserwować, że dla niektórych kontynentów dane skumulowane są na wykresie w pewnych mniejszych obszarach. Świadczy to o małym zróżnicowaniu wartości wektorów składowych głównych, podczas gdy dla pozostałych dane są bardziej rozproszone. Może to świadczyć o skrajnych różnicach na terenie poszczególnych kontynentów. Pomimo tego, można jednak w większości przypadków znaleźć dla każdego kontynentu obszar, w którym znajdują się wartości wektorów składowych.

Na wykresie można dostrzec naturalne grupowanie miast. Dla przykładu, miasta w Europie, Oceanii oraz Ameryce Północnej mają podobne wartości, co może świadczyć o porównywalnym poziomie rozwoju. Analogicznie, kontynenty Ameryka Południowa, Afryka i Azja są na podobnym poziomie według wektorów składowych głównych, jednak znacznie odstają od wcześniej wspomnianych trzech kontynentów.

Największą wartość PC1 przypisuje się miastu-państwu Singapur, natomiast najniższą wartość PC2 – Delhi, stolicy Indii. Singapur jest dobrze rozwiniętym miastem z silną gospodarką. Jest także jednym z najbogatszych państw, gdzie komfort życia jest na wysokim poziomie. W Delhi, mimo postępującego rozwoju, mniejszy nacisk kładziony jest na nowoczesne rozwiązania czy innowacyjne pomysły. Stąd niższy wskaźnik dotyczący startupów czy kapitału podwyższzonego ryzyka.

## 2.6 Korelacja zmiennych

PCA – Biplot



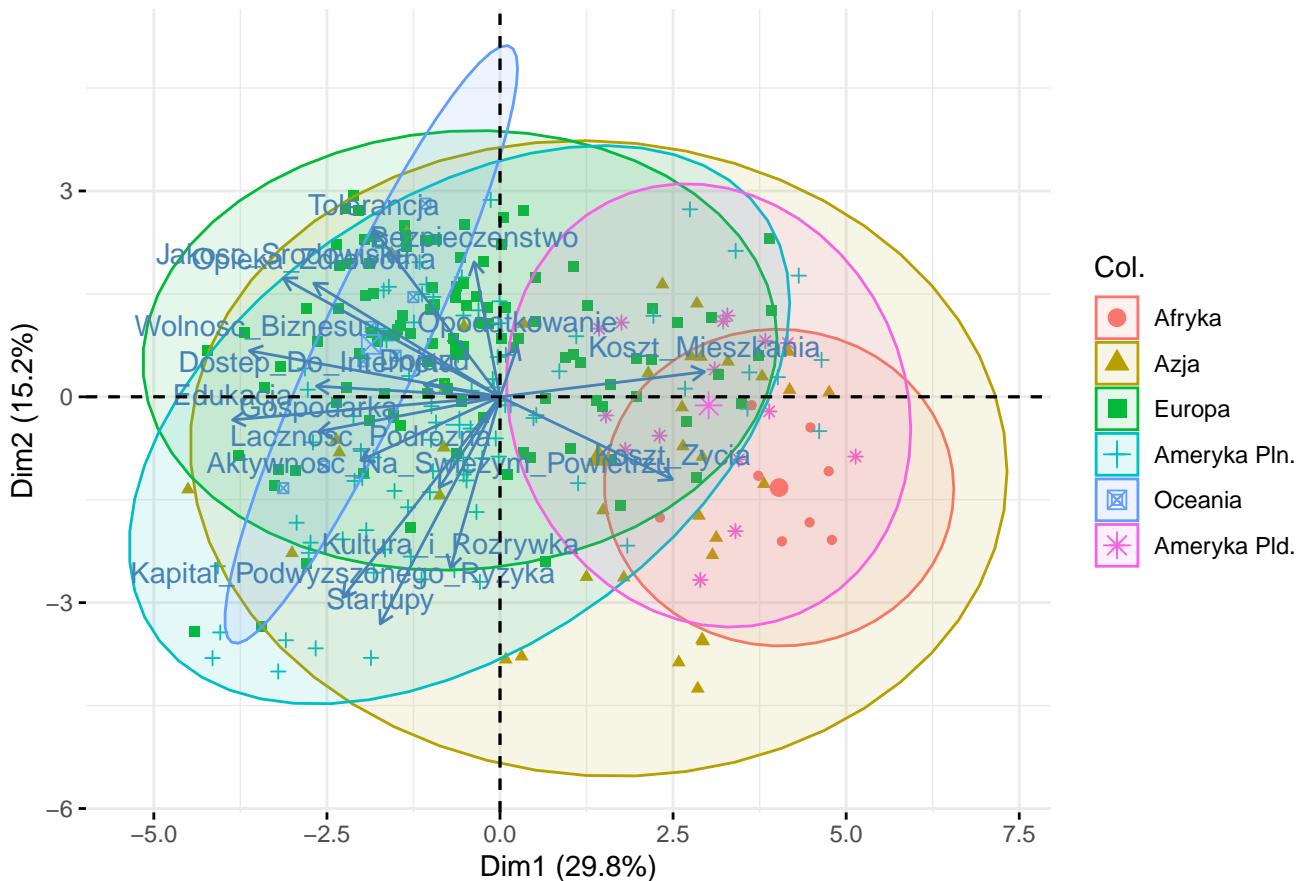
Rysunek 11: Dwuwykres

Na dwuwykresie przedstawionym na Rysunku 11 można odczytać, które zmienne są skorelowane dodatnio, które ujemnie, a które nie są wcale skorelowane. Do jednych z najbardziej dodatnio skorelowanych zmiennych należą np. Startupy, Kultura\_i\_Rozrywka oraz Kapitał\_Podwyższzonego\_Ryzyka, a także Gospodarka i Edukacja. Można to interpretować jako zjawisko, w którym te zmienne wspólnie się rozwijają.

Zmienne skorelowane ujemnie to np. Koszt\_Mieszkania i Edukacja, czy Koszt\_Zycia i Opieka\_Zdrowotna. Oznacza to, że przy wzroście jednej zmiennej, druga spada – czyli np. wyższe koszty życia niekoniecznie idą w parze z lepszą opieką zdrowotną.

Zmienne nieskorelowane, czyli takie, których rozwój lub zahamowanie nie ma przełożenia na inne zmienne, to np. Aktywnosc\_Na\_Swiezym\_Powietrzu i Jakośc\_Środowiska, Opodatkowanie i Jakość\_Środowiska, czy też Tolerancja..

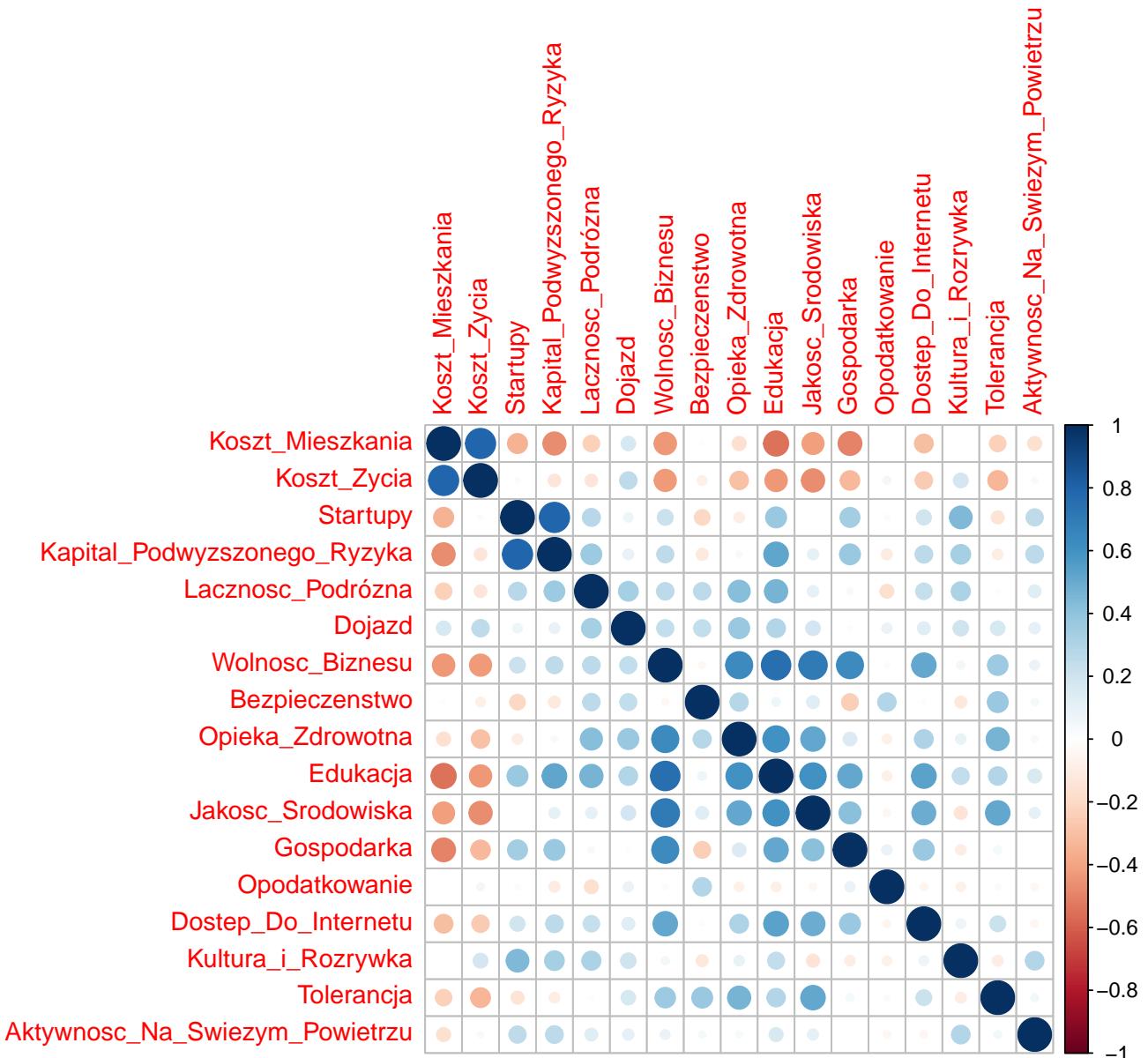
## PCA – Biplot



Rysunek 12: Dwuwykres z podziałem na kontynenty

Rysunek 12 również przedstawia dwuwykres zmiennych, jednak z dodatkowym podziałem na kontynenty. Można zauważać, że Oceania, Europa i Ameryka Północna znajdują się w jednym obszarze wykresu, co może wskazywać na wysoki poziom rozwoju gospodarczego i ekonomicznego. Z kolei Afryka i Ameryka Południowa są umiejscowione po lewej stronie wykresu. Może to sugerować, że koszty życia i mieszkania są tam niższe, jednak równocześnie wiele zmiennych gospodarczych, kulturowych i ekonomicznych – takich jak opieka zdrowotna, kultura i rozrywka czy dostęp do internetu – osiąga znacznie niższe wartości niż na wcześniej wspomnianych kontynentach.

Na wykresie widoczne jest również, że miasta azjatyckie są rozproszone po całym obszarze, co może świadczyć o dużym zróżnicowaniu poziomu rozwoju między nimi.



Rysunek 13: Macierz korelacji

Korzystając z Rysunku 13, możemy wysnuć analogiczne wnioski jak na podstawie dwuwykresu. Zależności zauważone wcześniej potwierdzają się w danych odczytanych z macierzy korelacji. Jednak dzięki jej bardziej przejrzystej formie, można łatwiej dostrzec dodatkowe zależności, np. silną dodatnią korelację między zmiennymi: Wolność\_Biznesu i Gospodarka, Opodatkowanie i Edukacja, Startupy i Kapitał\_Podwyzszonego\_Ryzyka, Koszt\_Zycia i Koszt\_Mieszkania.

Ujemna korelacja występuje np. między zmiennymi: Edukacja a Koszt\_Mieszkania i Koszt\_Zycia, a także Jakość\_Srodowiska z Kosztem\_Zycia i Startupami. Brak korelacji obserwujemy np. między Opodatkowaniem a Tolerancją.

Można więc zauważyć, że wnioski są spójne, jednak analiza macierzy korelacji zmniejsza ryzyko błędnej interpretacji dzięki większej czytelności.

## 2.7 Wnioski końcowe

Wykonując analizę składowych głównych, powiązaliśmy wektor PC1 z rozwojem gospodarczym, PC2 z nowoczesnością państwa, a PC3 z poziomem i wygodą życia. Na podstawie wykresów do-

strzegliśmy także, że do wyjaśnienia 80% i 90% całkowitej zmienności potrzebujemy odpowiednio 7 i 10 pierwszych składowych głównych.

Następnie, korzystając z mapy rozproszenia, mogliśmy zaobserwować, że kontynenty takie jak Europa, Oceania i Ameryka Północna znajdują się na podobnym poziomie rozwoju oraz gospodarczym. Analogicznie, w jednym obszarze wykresu znalazły się Azja, Afryka oraz Ameryka Południowa. Choć oddalone od wcześniej wspomnianych kontynentów, te trzy regiony również znajdują się na porównywalnym poziomie według analizowanych wskaźników.

Ciekawą obserwacją było także przypisanie największej wartości wektora PC1 Singapurowi, co może świadczyć o silnej gospodarce tego miasta-państwa, oraz najwyższej wartości PC2 Delhi, co może wskazywać na mniejsze znaczenie innowacyjności i nowoczesnych rozwiązań.

Na początku została wykonana standaryzacja danych, dzięki której późniejsza analiza była bardziej adekwatna i czytelna, a ważne informacje nie zostały zagubione. Bez standaryzacji jedna zmienna mogłaby zdominować dane, co prowadziłoby do zniekształconych wyników. Ponadto, dzięki standaryzacji, wykresy rozrzutu i dwuwykresy były czytelne i przejrzyste, a dane nie były skupione w jednym miejscu.

## 3 Skalowanie wielowymiarowe

### 3.1 Krótki opis zagadnienia

Celem raportu jest analiza danych dotyczących pasażerów Titanica przy użyciu skalowania wielowymiarowego (MDS). Dane zawierają informacje o charakterystykach pasażerów, takich jak wiek, płeć, klasa pasażerska oraz informację o przeżyciu katastrofy. Główne pytania badawcze to:

- Czy możliwe jest zredukowanie wymiaru danych przy zachowaniu istotnych informacji?
- Czy istnieje widoczny podział pasażerów na grupy związane z przeżyciem katastrofy, płecią lub klasą pasażerską?
- Czy w danych występują obserwacje nietypowe?

#### 3.1.1 Przeprowadzone analizy

W projekcie wykonano redukcję wymiaru przy użyciu skalowania wielowymiarowego (MDS) na danych dotyczących pasażerów Titanica. Główne etapy obejmują:

- *Przygotowanie danych:*
  - Usunięcie zbędnych zmiennych (ID, nazwisk, numerów biletów itp.)
  - Konwersję typów zmiennych (Płeć, Klasa, Port)
  - Obsługę brakujących wartości (usunięcie rekordów z brakującymi danymi dla kluczowych zmiennych)
- *Redukcję wymiaru:*
  - Obliczenie macierzy odmiенноścia z użyciem odległości Gowera (dla danych mieszanych: ilościowych i jakościowych)
  - Wykonanie niemetrycznego MDS (NMDS) z 2 wymiarami docelowymi ( $k = 2$ )
  - Ocena jakości odwzorowania za pomocą współczynnika stresu i diagramu Sheparda

- *Wizualizacje i interpretacje:*
  - Wykresy rozrzutu z podziałem na:
    - \* Przeżycie (Survived)
    - \* Płeć (Sex)
    - \* Klasę (Pclass)
  - Dodanie elips grupowych (80% przedziału ufności) i obserwacji odstających
  - Analiza skupisk i wzorców przestrzennych

### 3.1.2 Wykorzystane narzędzia

**Metody statystyczne:**

- Skalowanie wielowymiarowe (MDS) – wariant niemetryczny
- Analiza skupień oparta na odległościach
- Obliczanie odległości dla identyfikacji outlierów (wyników odstających)

**Wizualizacje:**

- Wykresy rozrzutu 2D z ggplot2 (geom\_point, stat\_ellipse)
- Diagram Sheparda (Shepard z pakietu MASS)
- Kontury gęstości (stat\_density\_2d)
- Etykiety outlierów (geom\_label\_repel z ggrepel)

**Testy jakościowe:**

- Ocena stresu MDS
- Wizualna analiza separacji grup na wykresach

### 3.1.3 Wykorzystane parametry

**Dane wejściowe:**

- *Próbka:* n = 712 pasażerów (po usunięciu braków danych)
- *Zmienne:*
  - Ilościowe: Wiek, Liczba rodzeństwa/małżonków, Liczba rodziców/dzieci, Opłata za bilet
  - Jakościowe: Płeć, Klasa, Port
  - Jakościowa: Przeżycie – używana tylko do interpretacji

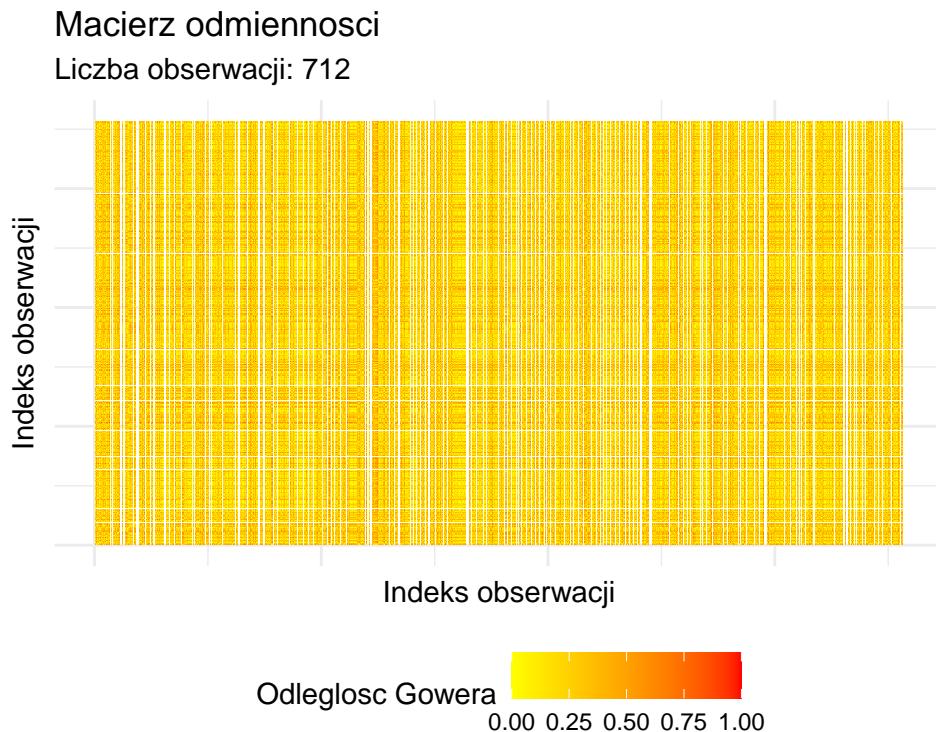
**Parametry MDS:**

- *Metryka:* odległość Gowera (dla danych mieszanych)
- *Algorytm:* niemetryczny MDS (isoMDS z pakietu MASS)
- *Wymiar docelowy:* d = 2
- *Współczynnik stresu:* 17.622%, wynik sugeruje, że:

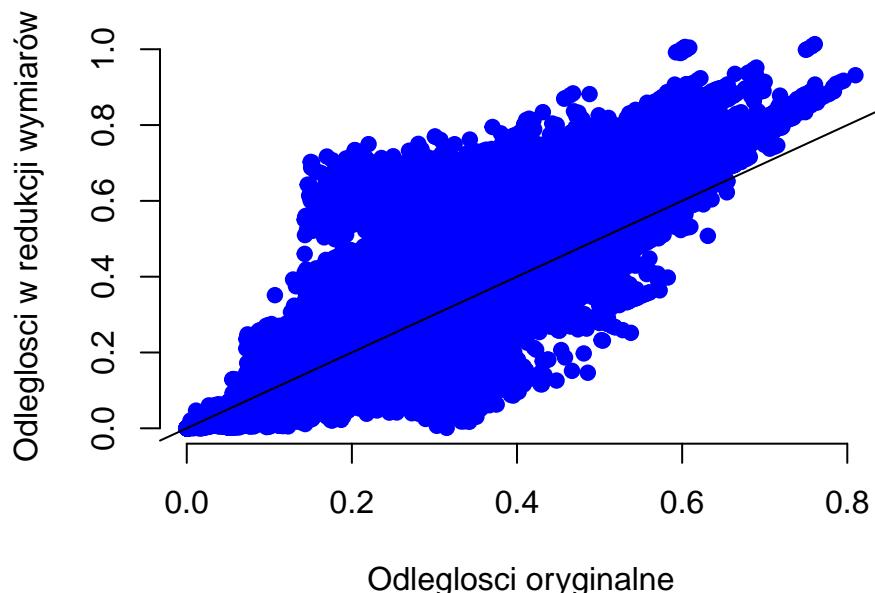
- Najsilniejsze trendy (np. podział na klasy/płeć) są prawdopodobnie realnie zachowane
- Detale i słabsze wzorce mogą być zniekształcone
- Odległości między punktami na wykresie są przybliżone (nie dokładne)

## 3.2 Wyniki

### 3.2.1 Wykresy i tabele



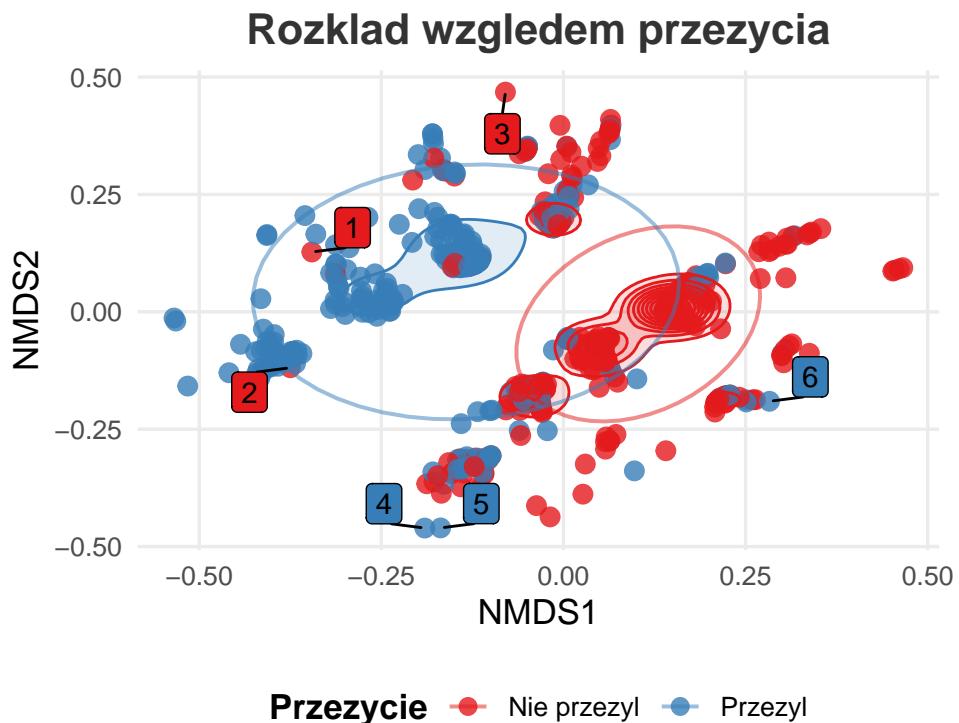
Rysunek 14: Wizualizacja macierzy odmiенноścii



Rysunek 15: Diagram Sheparda

**Diagram Sheparda** (Rysunek 15)

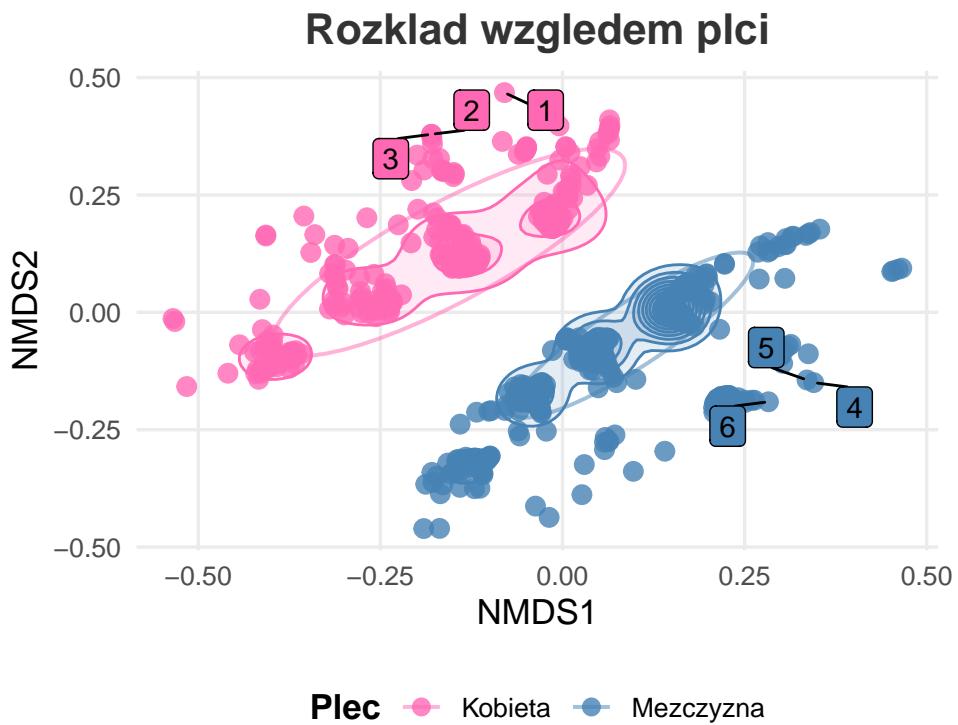
- *Jakość odwzorowania:*
  - Diagram Sheparda pokazuje zależność między oryginalnymi odległościami (Gowera) a odległościami w przestrzeni MDS. Jeśli punkty leżą blisko linii 1:1, odwzorowanie jest dobre. W tym przypadku widoczne jest pewne rozproszenie, co sugeruje, że NMDS przybliża odległości, ale nie idealnie.
- *Wnioski:*
  - Współczynnik stresu (17.622 %) wskazuje na umiarkowaną jakość odwzorowania. Silne trendy (np. podział na klasy/płeć) są widoczne, ale detale mogą być zniekształcone.



Rysunek 16: Wykres rozkładu względem przeżycia

#### Wykres rozkładu względem przeżycia (Rysunek 16)

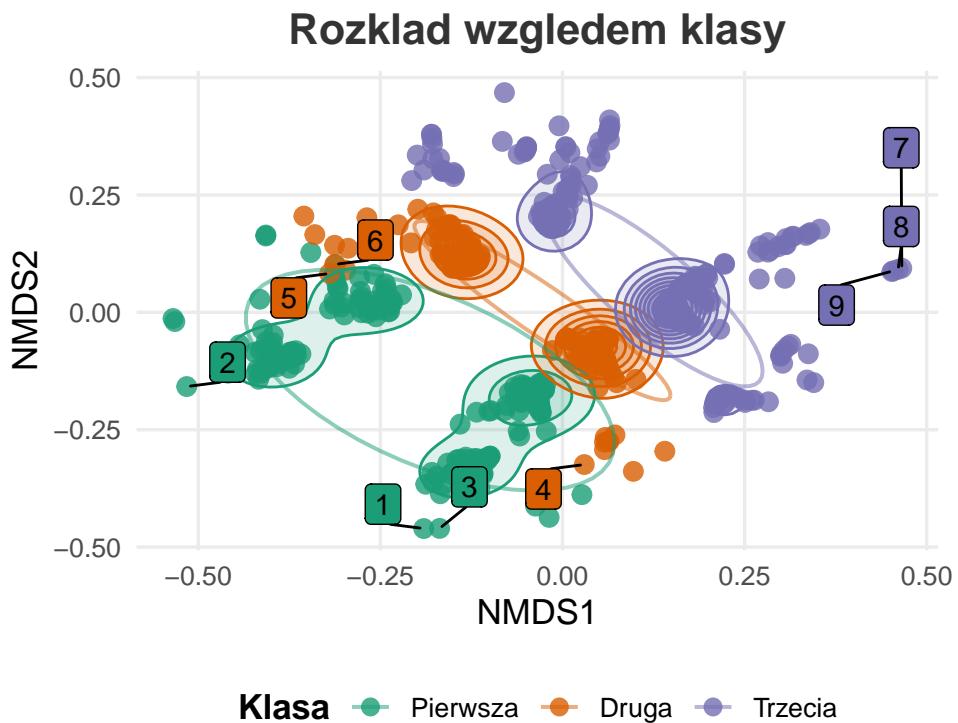
- *Podział na grupy:*
  - Widoczny jest częściowy podział na grupy “przeżył” vs “nie przeżył”, ale z dużym nakładaniem się elips (80% przedział ufności). Sugeruje to, że przeżycie zależy od innych czynników (np. klasy, płci), a nie tylko od pozycji w MDS.
- *Obserwacje odstające:*
  - Zidentyfikowano kilka outlierów, które mogą odpowiadać nietypowym przypadkom (np. pasażerom o skrajnych wartościach wieku lub opłaty).



Rysunek 17: Wykres rozkładu względem płci

#### Wykres rozkładu względem płci (Rysunek 17)

- *Podział na grupy:*
  - Silny podział ze względu na płeć – kobiety i mężczyźni tworzą wyraźne skupiska z minimalnym nakładaniem się. Odzwierciedla to historyczne fakty (priorytet dla kobiet podczas ewakuacji).
- *Obserwacje odstające:*
  - Outliery mogą dotyczyć np. osób z ekstremalnymi wartościami, możliwe związanymi z opłatą lub wiekiem.



Rysunek 18: Wykres rozkładu względem klasy

### Wykres rozkładu względem klasy (Rysunek 18)

- *Podział na grupy:*
  - Klasa pasażerska silnie wpływa na pozycję w MDS. Pasażerowie klasy 1 są wyraźnie oddzieleni od klasy 3, co może odzwierciedlać różnice w opłatach, lokalizacji kajut lub priorytetach ewakuacyjnych.
- *Obserwacje odstające:*
  - Nietypowe rekordy mogą dotyczyć pasażerów o niestandardowych cechach związanych z wiekiem, opłatą lub inną cechą ilościową z oryginalnej ramki danych.

Tabela 10: Przeżywalność pasażerów Titanica według klasy i płci

Klasa	Płeć	Liczba osób			% przeżycia
		Przeżył	Nie przeżył	Suma	
Pierwsza	Kobieta	80	3	83	96.4
Pierwsza	Mężczyzna	40	61	101	39.6
Druga	Kobieta	68	6	74	91.9
Druga	Mężczyzna	15	84	99	15.2
Trzecia	Kobieta	47	55	102	46.1
Trzecia	Mężczyzna	38	215	253	15.0

Źródło: Dane pochodzą ze zbioru titanic\_train z pakietu titanic

### Główne wnioski tabeli przeżywalności:

- *Kobiety:* Miały znacznie wyższą przeżywalność niż mężczyźni.
- *Klasa 1:* Najwyższa przeżywalność (96.8% dla kobiet, 39.6% dla mężczyzn), co potwierdza hipotezę o priorytecie dla bogatszych pasażerów.

- *Klasa 3:* Najniższa przeżywalność (46.1% kobiet, 15% mężczyzn), co może wynikać z gorszej lokalizacji kajut.

### 3.3 Podsumowanie wniosków

**Redukcja wymiaru:** - NMDS skutecznie uwidocznił główne trendy (płeć, klasa), ale stres (około 17.622%) wskazuje na przybliżony charakter odwzorowania.

**Podziały grupowe:** - Najsilniejsze dla płci i klasy – co zgadza się z historycznymi danymi.

**Outliery:** - Zidentyfikowano nietypowe przypadki, wymagające dalszej analizy (np. ze względu na ekstremalne wartości wieku/opłaty).

**Tabela przeżywalności:** - Potwierdza, że płeć i klasa były kluczowymi czynnikami przeżycia, co tłumaczy częściowo nakładanie się grup na wykresach MDS.