

Sprawozdanie z listy 2

Eksploracja danych

Marta Stankiewicz (282244)

Paweł Nowak (282223)

2025-04-30

Spis treści

1 Ocena zdolności separacyjnych zmiennych, dyskretyzacja zmiennych ciągły	2
1.1 Ocena zdolności dyskryminacyjnych zmiennych ciągły	2
1.2 Porównanie różnych metod dyskretyzacji nienadzorowanej.	3
2 Analiza składowych głównych	4
2.1 Porównanie wariancji zmiennych ilościowych.	4
2.2 Badanie korelacji między zmiennymi.	7
2.3 Wyznaczanie składowych głównych.	8
2.4 Wizualizacja danych wielowymiarowych	10
2.5 PCA bez uwzględnienia standaryzacji.	12
2.6 Wnioski końcowe	14
3 Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))	14
3.1 Redukcja wymiaru na bazie MDS	14
3.2 Wizualizacja danych po zastosowaniu skalowania wielowymiarowego	16

Spis rysunków

1 Wykresy skrzypcowo-pudełkowe dla zmiennych ciągły	3
2 Wykresy pudełkowe zmiennych ciągły przed zastosowaniem standaryzacją	5
3 Wykresy pudełkowe zmiennych ciągły po zastosowaniu standaryzacji	6
4 Macierz korelacji dla zmiennych ciągły	7
5 Wykresy pudełkowe dla składowych głównych	8
6 Porównanie udziału wariancji wyjaśnianej przez poszczególne składowe główne	9
7 Skumulowana wariancja wyjaśniana przez kolejne składowe główne	10
8 Porównanie udziału wariancji wyjaśnianej przez poszczególne składowe główne bez standaryzacji	11

9	Wykresy pudełkowe dla składowych głównych bez zastosowanej standaryzacji	12
10	Skumulowana wariancja wyjaśniana przez kolejne składowe główne	13
11	Diagram Sheparda po zastosowaniu MDS dla $k = 2$	14
12	Diagram Sheparda po zastosowaniu MDS dla $k = 2$	15
13	Wykres rozrzutu po zastosowaniu MDS dla $k = 2$	16
14	Wykres rozrzutu po zastosowaniu MDS dla $k = 2$ z grupowaniem według zmiennej 'Survived'	17
15	Wykres rozrzutu po zastosowaniu MDS dla $k = 2$ z grupowaniem według zmiennej 'Pclass'	18
16	Wykres rozrzutu po zastosowaniu MDS dla $k = 2$ z grupowaniem według zmiennej 'Sex'	19

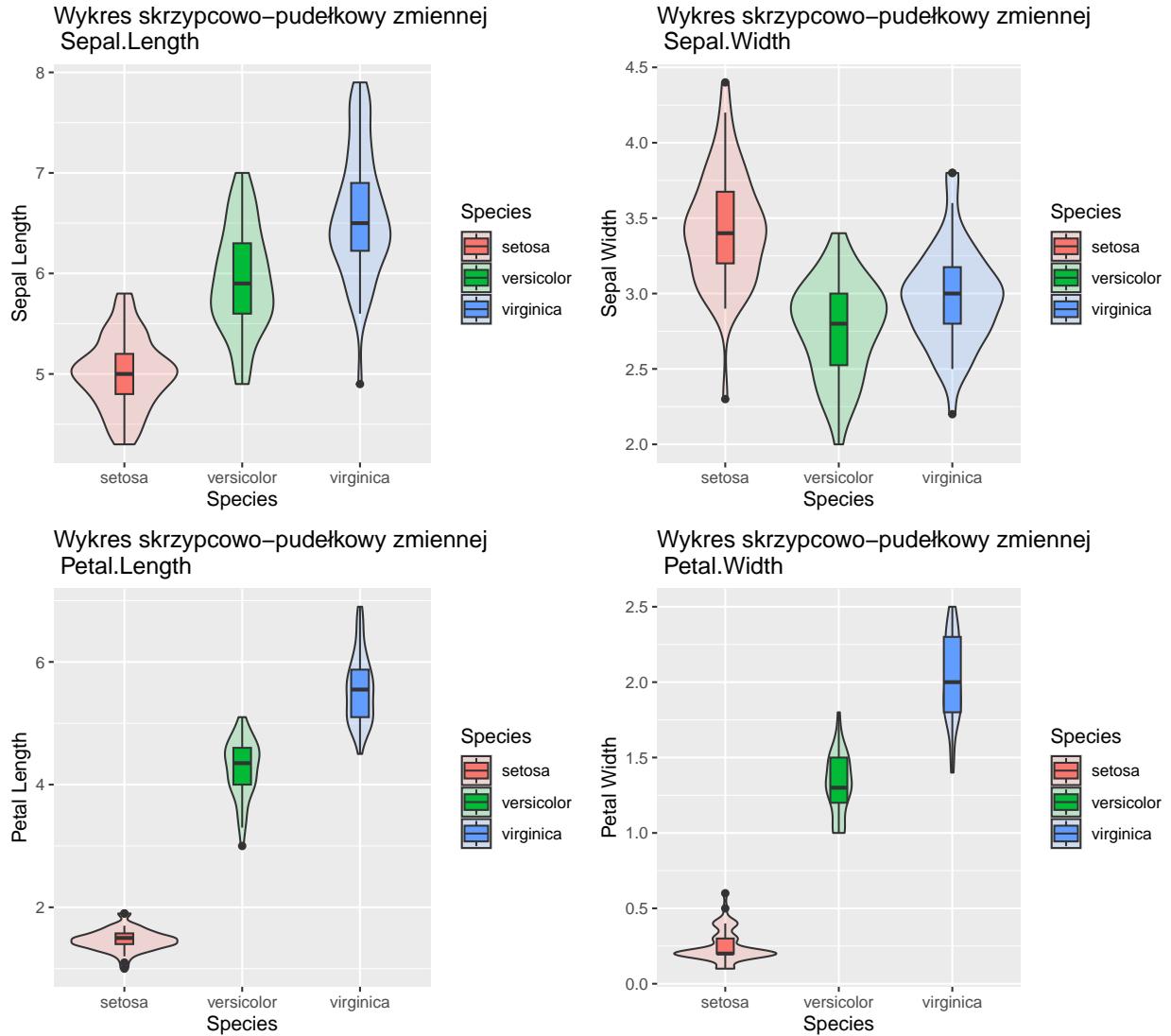
Spis tabel

1	Ocena przedziałowania zmiennej Sepal Width	4
2	Ocena przedziałowania zmiennej Petal Width	4

1 Ocena zdolności separacyjnych zmiennych, dyskretyzacja zmiennych ciągłych

1.1 Ocena zdolności dyskryminacyjnych zmiennych ciągłych.

W celu zbadania zdolności dyskryminacyjnej cech, posłużymy się wykresem skrzypcowo-pudełkowym (tj. wykresem skrzypcowym wraz z wykresem pudełkowym).



Rysunek 1: Wykresy skrzypcowo-pudełkowe dla zmiennych ciągłych

Z wykresów 1 wnioskujemy, że największe zdolności dyskryminacyjne wykazuje zmienna *Petal.Width*. Z kolei najmniejsze zdolności do separacji gatunków obserwujemy u zmiennej *Sepal.Width*.

1.2 Porównanie różnych metod dyskretyzacji nienadzorowanej.

Dla wymienionych wyżej zmiennych (tj. *Petal.Width* oraz *Sepal.Width*) zastosujemy teraz różne techniki przedziałowania (dyskretyzacji) według, odpowiednio, **stałej szerokości przedziału, równej częstości, algorytmu K-średnich, stałych granicach** przedziałów ustalonych przez użytkownika.

1.2.1 Metodologia oceny skuteczności dyskretyzacji

Aby ocenić skuteczność każdej ze wspomnianych metod, przyjęliśmy następującą metodologię. Najpierw dokonaliśmy przedziałowania każdej obserwacji, korzystając ze wszystkich metod, a następnie wybraliśmy tę klasę, która występuje najczęściej (w przypadku tzw. "remisu" wybierana jest dowolna klasa). Następnie sprawdzaliśmy, w ilu przypadkach wynik przedziałowania każdej metody zgadzał się ze zgregowaną klasą. Tę liczbę podzieliliśmy przez liczbę wszystkich przypadków, aby uzyskać procent zgodności danej metody dyskretyzacji. Porównanie różnych metod przedziałowania zostały przedstawione poniżej

Równa częstotliwość	Równa szerokości	K-średnie	Stałe granice
79.330	93.330	72.000	84.670

Tabela 1: Ocena przedziałowania zmiennej Sepal Width

Równa częstotliwość	Równa szerokości	K-średnie	Stałe granice
97.330	100.000	98.670	86.000

Tabela 2: Ocena przedziałowania zmiennej Petal Width

1.2.2 Wnioski dotyczące skuteczności przedziałowania

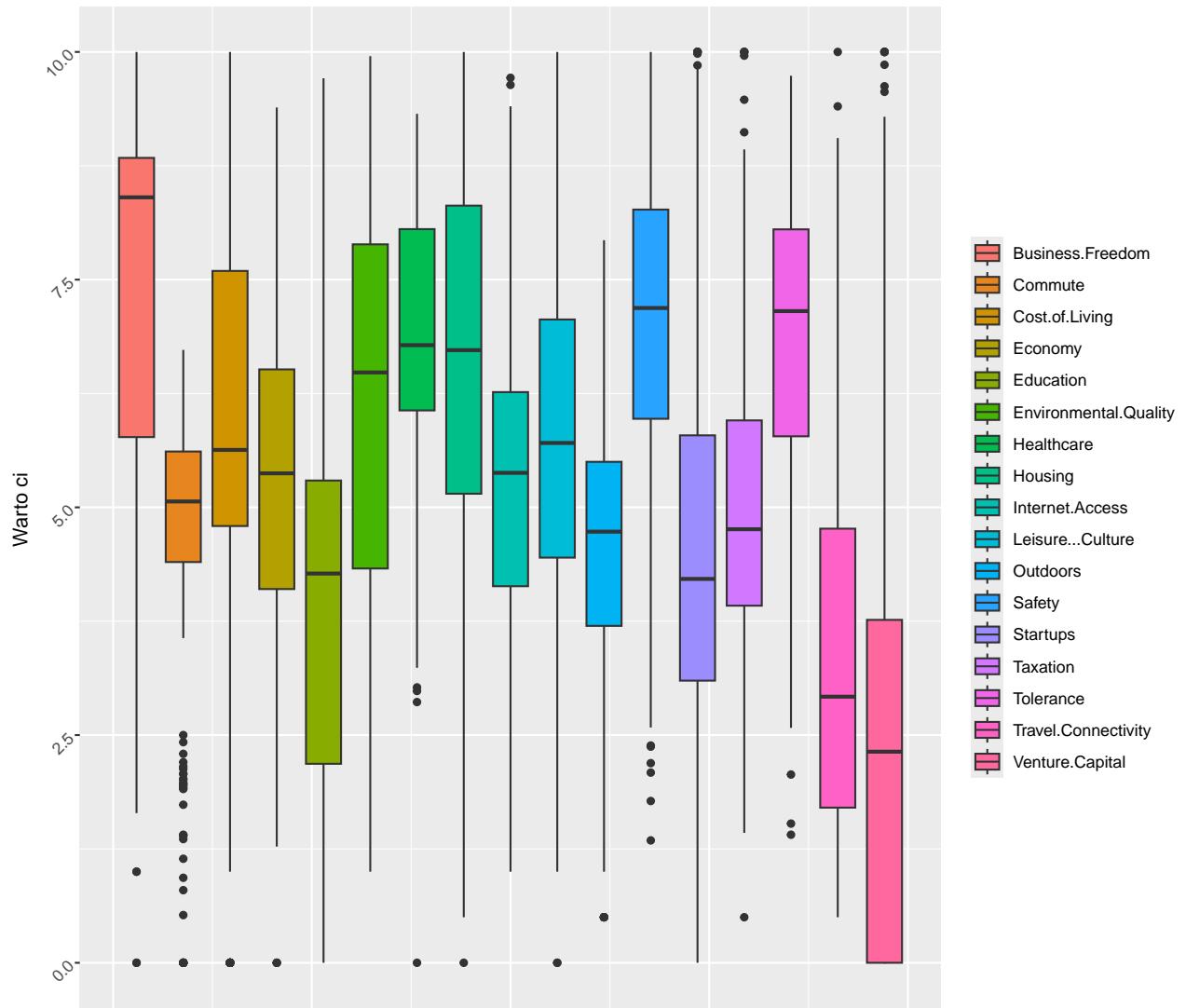
Z tabel 1 oraz 2 możemy wywnioskować, że w obu przypadkach największą skutecznością charakteryzuje się metoda dyskretyzacji oparta na **algorytmie K-średnich**. Z kolei najgorszą skuteczność przedziałowania obserwujemy dla metody opartej na **stałych granicach** przedziału. Wyniki dyskretyzacji zastosowanej dla zmiennej *Petal.Width* znacząco różnią się od wyników przedziałowania zastosowanego dla atrybutu *Sepal.Width*. Jest to zgodne z intuicją — jak wykazaliśmy wcześniej, najgorsze zdolności separacyjne klas wykazuje właśnie zmienna **Sepal.Width**, co znacząco wpływa na niską skuteczność metod przedziałowania. Analogiczna zależność występuje w przypadku cechy **Petal.Width**, która z kolei charakteryzowała się wysokimi zdolnościami dyskryminacyjnymi, co przełożyło się na wysoką dokładność podejść dyskretyzacji.

2 Analiza składowych głównych

2.1 Porównanie wariancji zmiennych ilościowych.

W celu porównania wariancji wszystkich zmiennych ilościowych ze zbioru *uaScoresDataFrame*, posłużymy się wykresami pudełkowymi.

Wykres pudełkowy zmiennych ilościowych

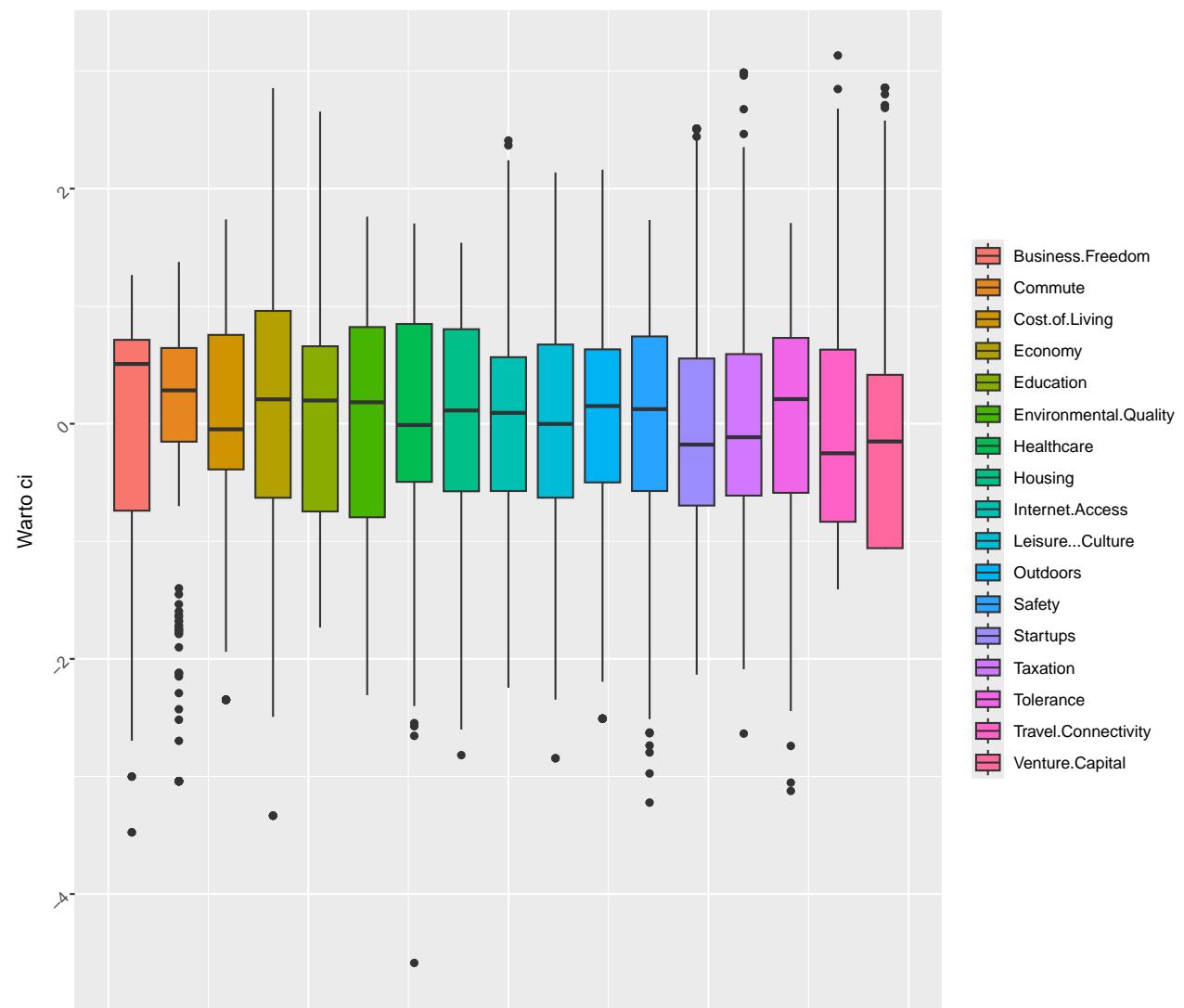


Rysunek 2: Wykresy pudełkowe zmiennych ciągły przed zastosowaniem standaryzacji

Przyjrzyjmy się wykresowi 2. Obserwujemy wysokie zróżnicowanie wariancji. Z jednej strony w badanym zbiorze występują cechy o niskiej dewiacji, która charakteryzuje chociażby zmienną *Commute*. Z drugiej obecność takich zmiennych jak *Environmental.Quality* i *Venture.Capital* pokazują, że nie brakuje atrybutów o wysokiej wariancji. W celu ujednolicenia wariancji zmiennych, konieczne będzie zastosowanie standaryzacji.

Na poniższym wykresie 3 pudełkowym widoczne są efekty standaryzacji zastosowane dla zmiennych ze zbioru danych.

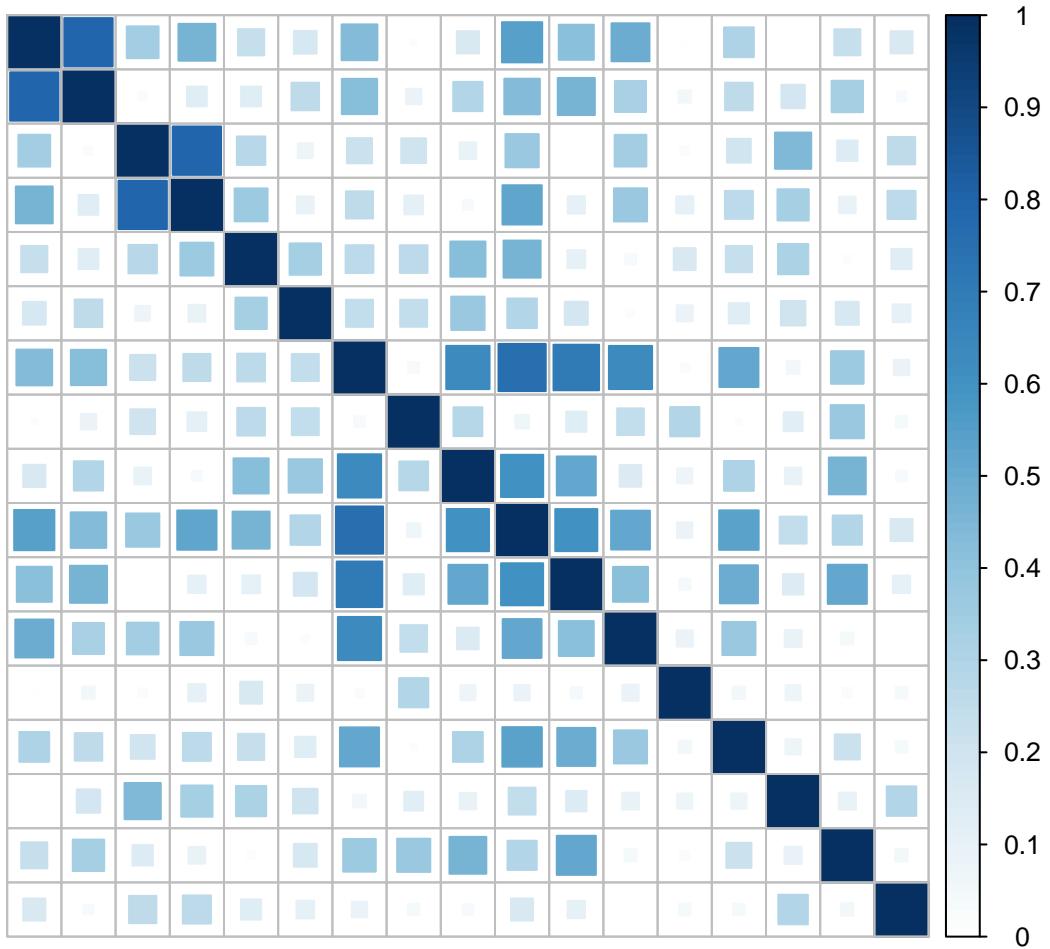
Wykres pułapkowy zmiennych ciągowych



Rysunek 3: Wykresy pułapkowe zmiennych ciągłych po zastosowaniu standaryzacji

2.2 Badanie korelacji między zmiennymi.

Po dokonaniu standaryzacji zmiennych ilościowych, zbadamy jeszcze, jak silne są korelacje między atrybutami w zbiorze danych. Występowanie silnej korelacji świadczy o występowaniu redundantnych zmiennych. Taka redundancja może zostać wyeliminowana za pomocą analizy składowych głównych. Aby poprawić czytelność wykresu, nazwy zmiennych zostały pominięte, a wartość współczynnika została przeskalowana do przedziału [0;1].

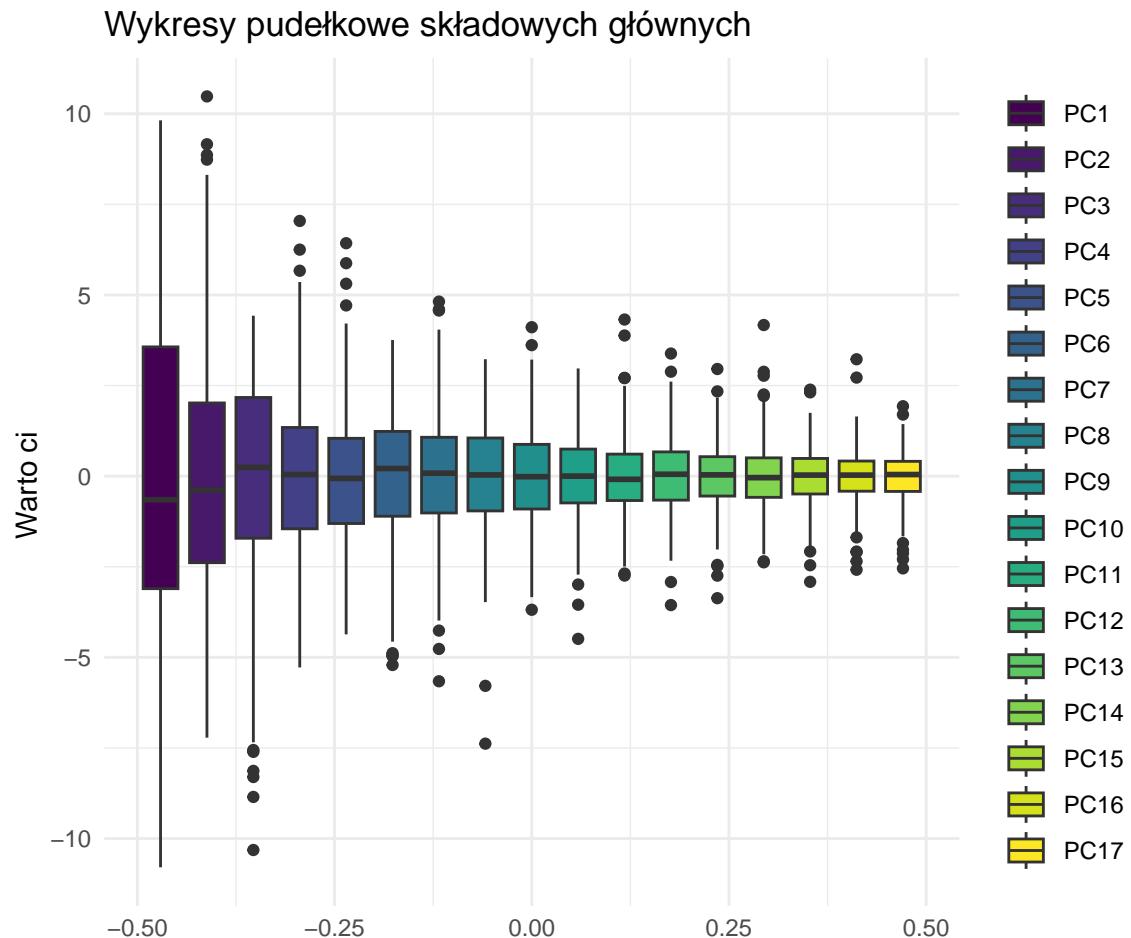


Rysunek 4: Macierz korelacji dla zmiennych ciągłych

Na podstawie rysunku 4 można zauważyć, że w zdecydowanej większości przypadków korelacje pomiędzy zmiennymi są stosunkowo słabe. Niemniej jednak, występują również przypadki skrajne, w których wartości współczynnika korelacji — rozpatrywane w sensie bezwzględnym — zbliżają się do jedności, wskazując na silne liniowe powiązania między wybranymi zmiennymi. W związku z tym należy oczekwać, że redukcja wymiarowości będzie wymagała uwzględnienia relatywnie dużej liczby składowych głównych, aby osiągnąć zakładaną frakcję wyjaśnianej wariancji.

2.3 Wyznaczanie składowych głównych.

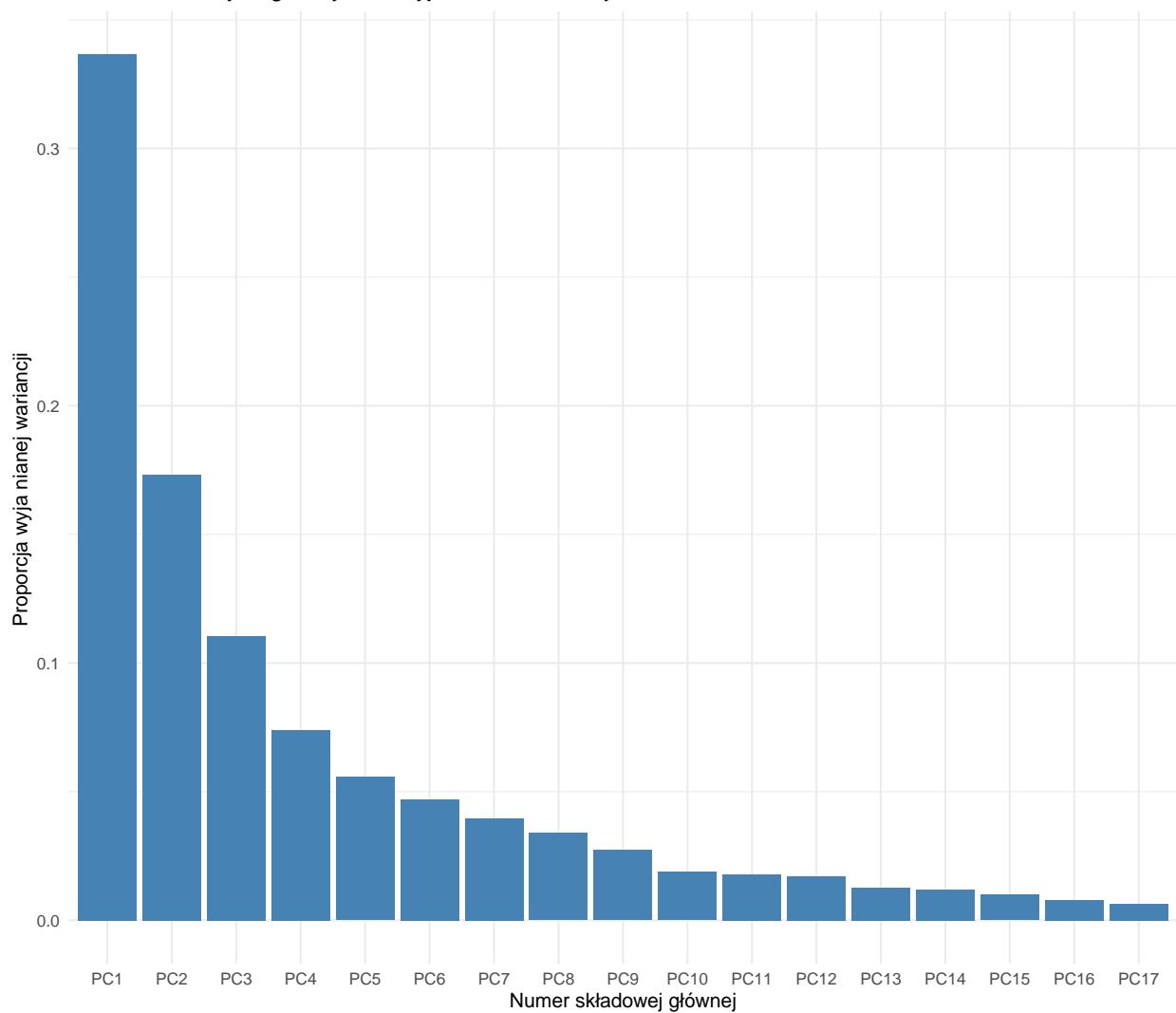
Dla analizowanego zbioru zmiennych ciągłyzych zostanie przeprowadzona analiza głównych składowych. W jej ramach porównany zostanie rozrzut składowych oraz stopień, w jakim wyjaśniają one całkowitą wariancję danych. Na zakończenie, na podstawie skumulowanej wariancji wyjaśnianej przez kolejne składowe, wyznaczona zostanie minimalna liczba komponentów niezbędnych do osiągnięcia poziomu co najmniej 80% lub 90% całkowitej wariancji.



Rysunek 5: Wykresy pudełkowe dla składowych głównych

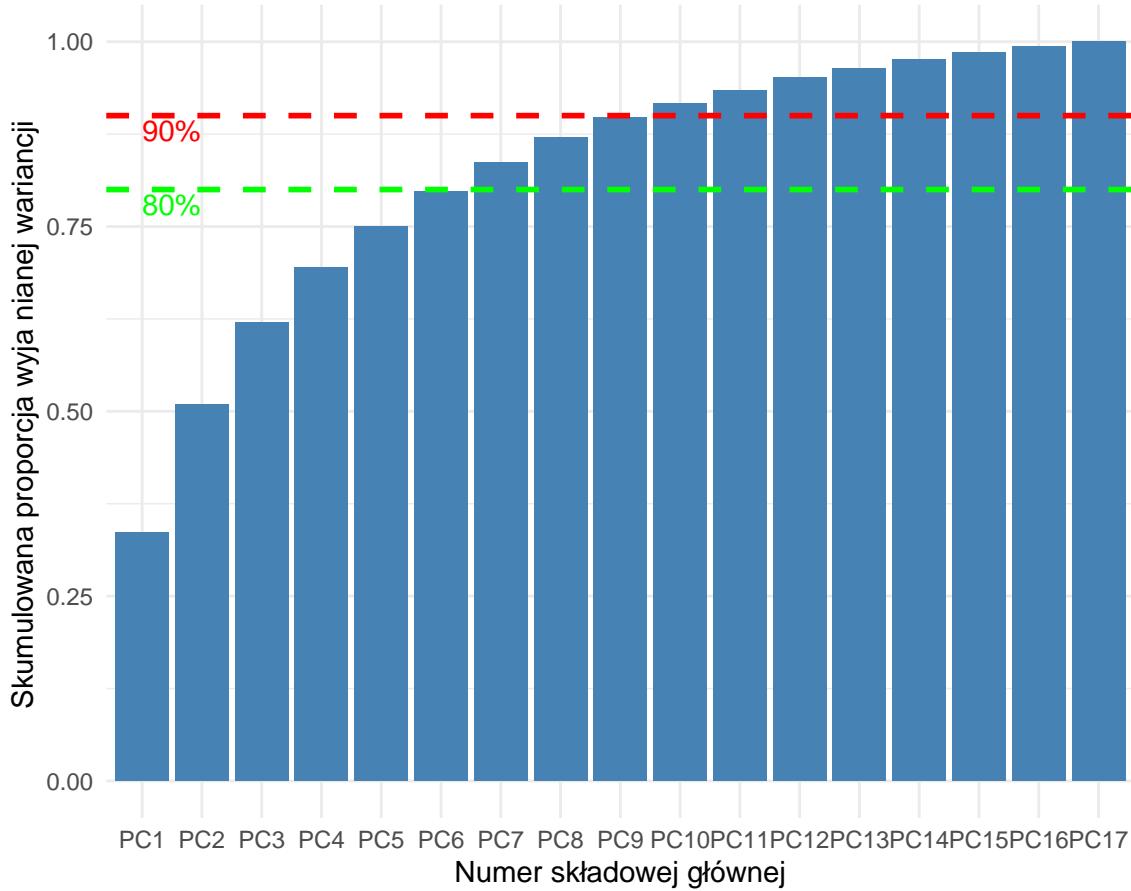
Mając już wyznaczone składowe główne, możemy odpowiedzieć na pytanie o minimalną liczbę komponentów niezbędnych do osiągnięcia założonej frakcji wyjaśnianej wariancji. W celu ilustracji udziału poszczególnych składowych w ogólnej wariancji danych, odwołajmy się do wykresu 6. Największe przyrosty wariancji obserwowane są dla pierwszych czterech składowych głównych, po czym tempo wzrostu wyraźnie spowalnia. W związku z tym, wstępna analiza sugeruje, iż uwzględnienie jedynie 4–6 pierwszych składowych głównych (spośród wszystkich 17) może być wystarczające do uzyskania satysfakcyjującego poziomu odwzorowania struktury danych.

Udział składowych głównych w wyjaśnianiu wariancji



Rysunek 6: Porównanie udziału wariancji wyjaśnianej przez poszczególne składowe główne

Skumulowana wariancja wyjaśniana przez składowe główne



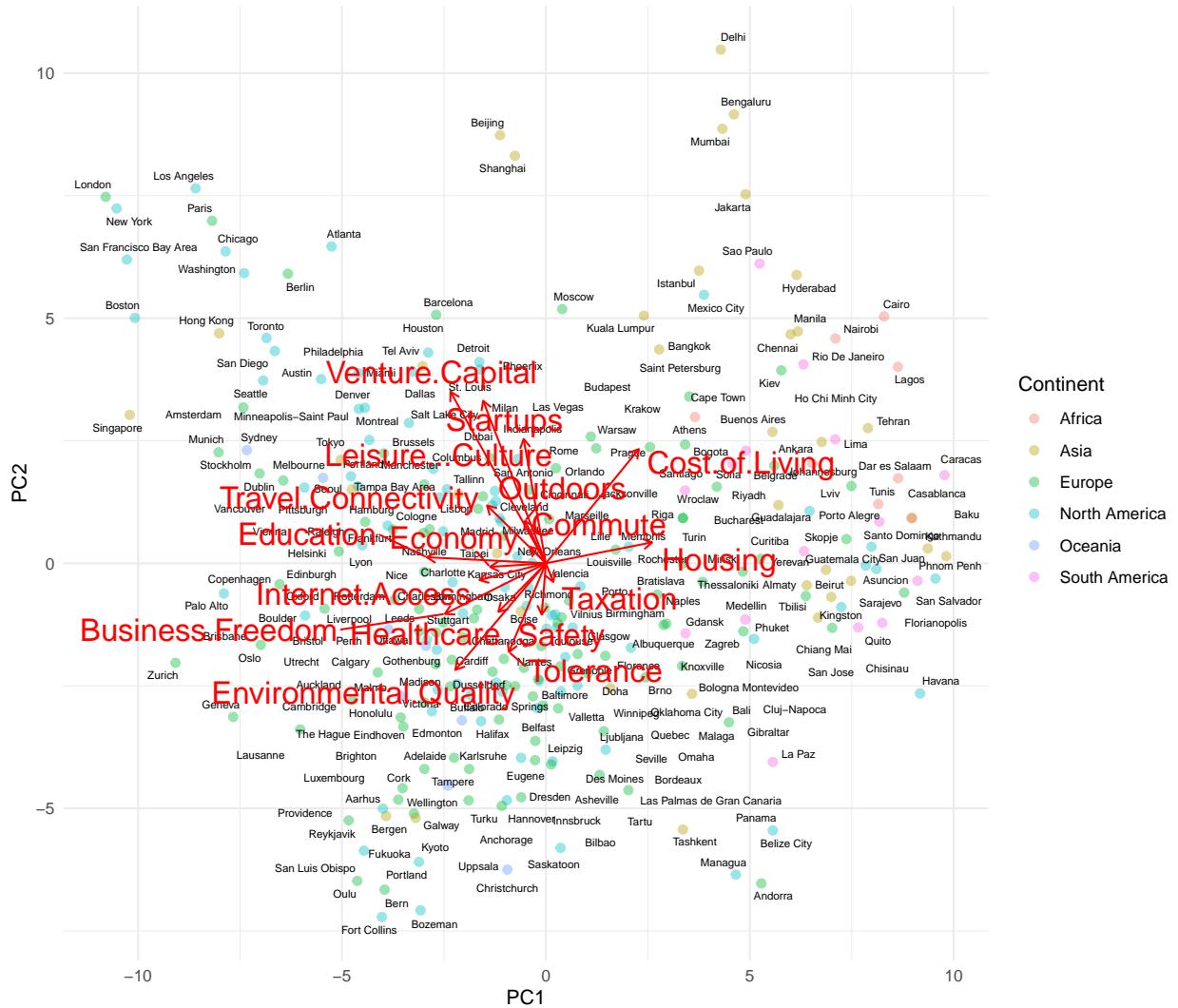
Rysunek 7: Skumulowana wariancja wyjaśniana przez kolejne składowe główne

Na podstawie wykresu 7 możemy określić liczbę składowych głównych niezbędnych do osiągnięcia założonego poziomu wyjaśnianej wariancji. W celu odtworzenia 80% całkowitej wariancji zmiennych oryginalnych wystarczające okazuje się uwzględnienie pierwszych sześciu składowych głównych. Natomiast osiągnięcie progu 90% wymaga rozszerzenia tego zbioru do dziewięciu składowych.

2.4 Wizualizacja danych wielowymiarowych

Po przeprowadzeniu redukcji wymiarowości za pomocą analizy głównych składowych (PCA), dokonano wizualizacji danych w przestrzeni wyznaczonej przez dwie pierwsze składowe, które kumulatywnie wyjaśniają największą część zmienności w zbiorze. Celem tej analizy jest identyfikacja potencjalnych struktur skupiskowych wśród obserwacji (miast) oraz wykrycie ewentualnych obserwacji odstających, które znaczco odbiegają od pozostałych pod względem analizowanych cech.

Wizualizacja składowych głównych



Rysunek 8: Porównanie udziału wariancji wyjaśnianej przez poszczególne składowe główne bez standaryzacji

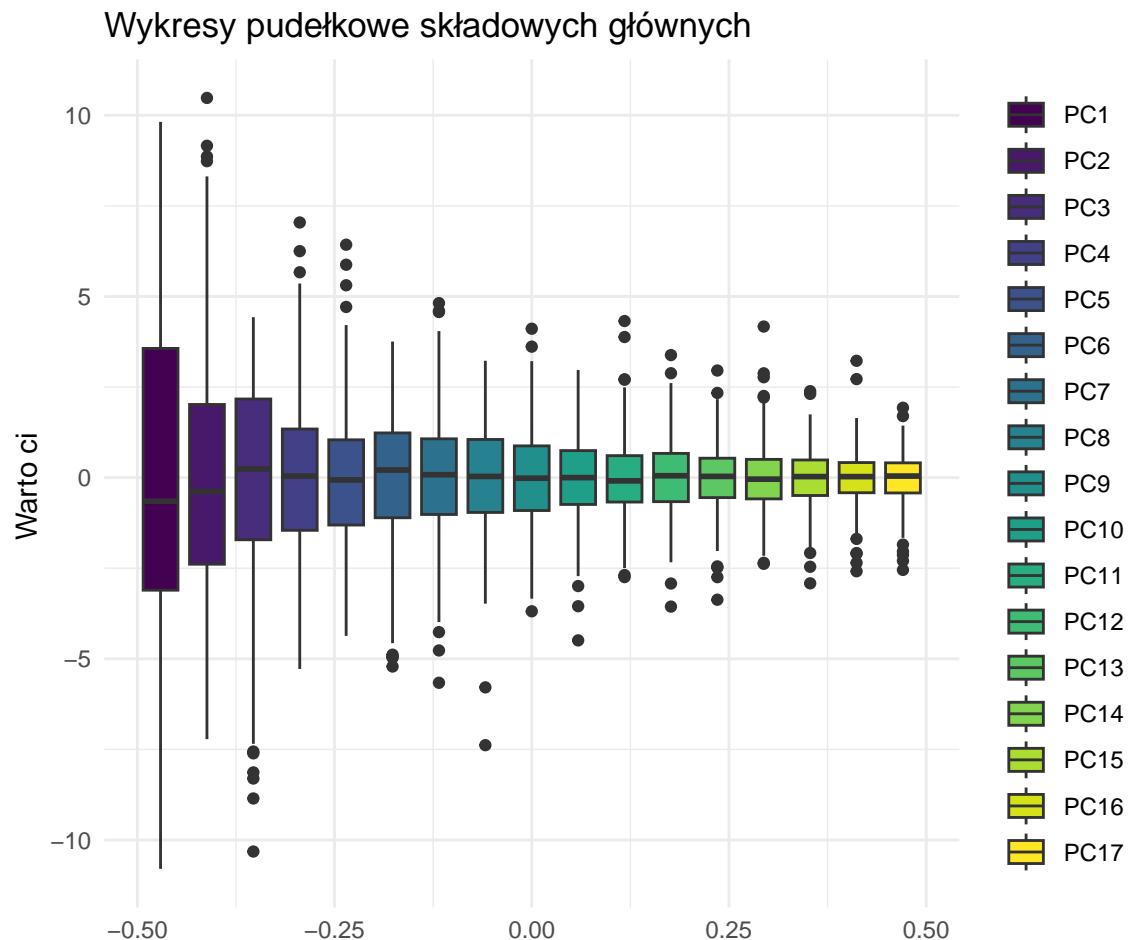
Większość miast koncentruje się w centralnej części wykresu, co sugeruje znaczny stopień podobieństwa między nimi pod względem analizowanych cech. Mimo to, zauważalne są wyraźnie wyodrębnione grupy miast, które odbiegają od głównego skupiska. W lewym górnym obszarze wykresu wyróżniają się m.in. Londyn, Nowy Jork, Los Angeles czy Chicago – miasta o szczególnym znaczeniu globalnym. Są to zarówno stolice silnych gospodarek (np. Berlin, Londyn), jak i wiodące centra biznesowe i kulturowe (np. Nowy Jork, Los Angeles). Charakterystyczny dla tej grupy jest wysoki poziom rozwoju infrastruktury i aktywności związanych z kulturą czasu wolnego (Leisure Culture) oraz intensywna obecność kapitału wysokiego ryzyka (Venture Capital).

Drugą wyraźnie wyodrębnioną grupę stanowią miasta zlokalizowane w prawym górnym obszarze wykresu, takie jak Delhi, Pekin, Meksyk czy Bengaluru. Są to dynamicznie rozwijające

się metropole o rosnącym znaczeniu finansowym, technologicznym oraz kulturowym. W odróżnieniu od wcześniej wspomnianych globalnych centrów, charakteryzują się one relatywnie niskimi kosztami życia, co stanowi istotny czynnik przyciągający zarówno mieszkańców, jak i inwestorów oraz przedsiębiorców poszukujących korzystnych warunków do rozwoju działalności.

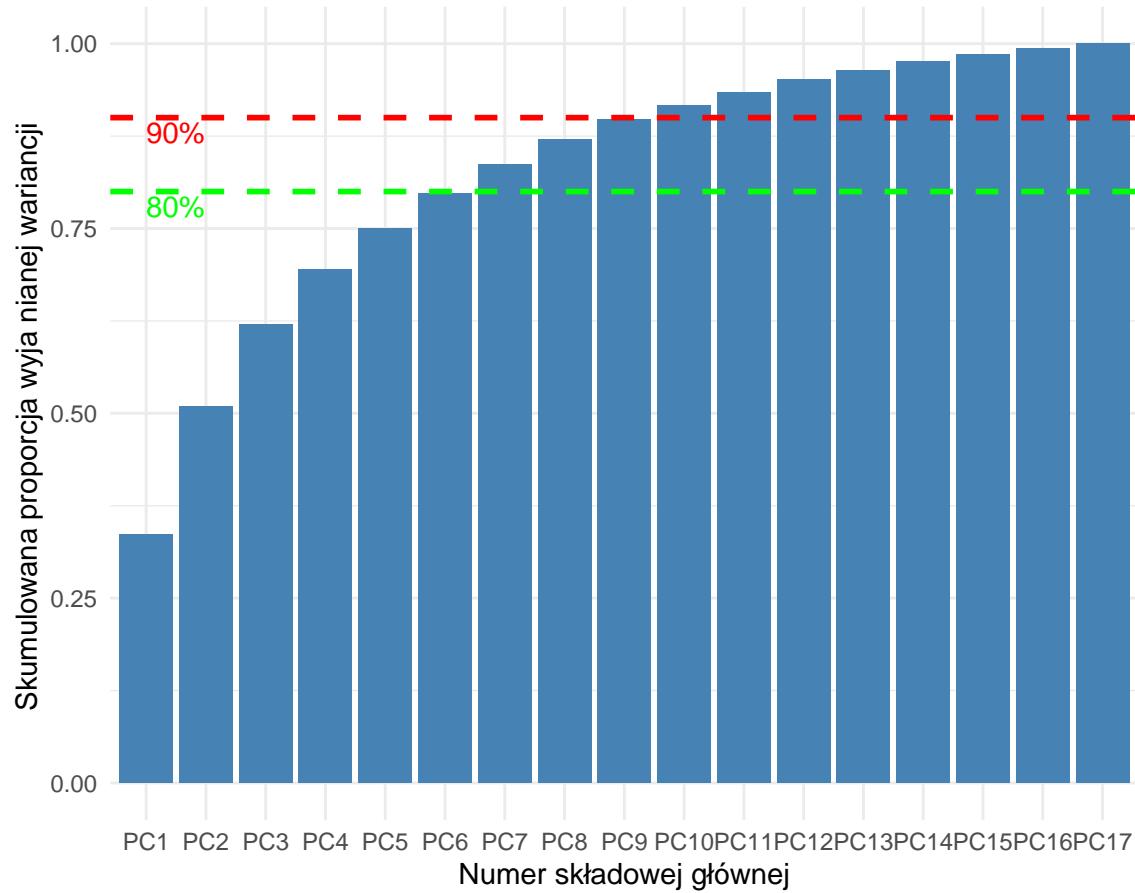
2.5 PCA bez uwzględnienia standaryzacji.

W dalszej części analizy porównana zostanie efektywność składowych głównych wyznaczonych na podstawie niestandardyzowanych zmiennych. Celem jest ocena, czy liczba komponentów niezbędnych do osiągnięcia z góry określonego progu wyjaśnionej wariancji istotnie różni się od tej uzyskanej przy zastosowaniu analizy głównych składowych z uprzednią standaryzacją danych.



Rysunek 9: Wykresy pudełkowe dla składowych głównych bez zastosowanej standaryzacji

Skumulowana wariancja wyjaśniana przez składowe główne



Rysunek 10: Skumulowana wariancja wyjaśniana przez kolejne składowe główne

Wyniki przedstawione na wykresach 9 oraz 10 wskazują, że brak standaryzacji danych nie wpływa istotnie na liczbę głównych składowych niezbędnych do osiągnięcia zadowalającego poziomu wyjaśnionej wariancji. Łączna wariancja kolejnych składowych nie odbiega w sposób znaczący od łącznej wariancji obserwowanej na wykresie 7, gdzie uprzednio zastosowano standaryzację danych.

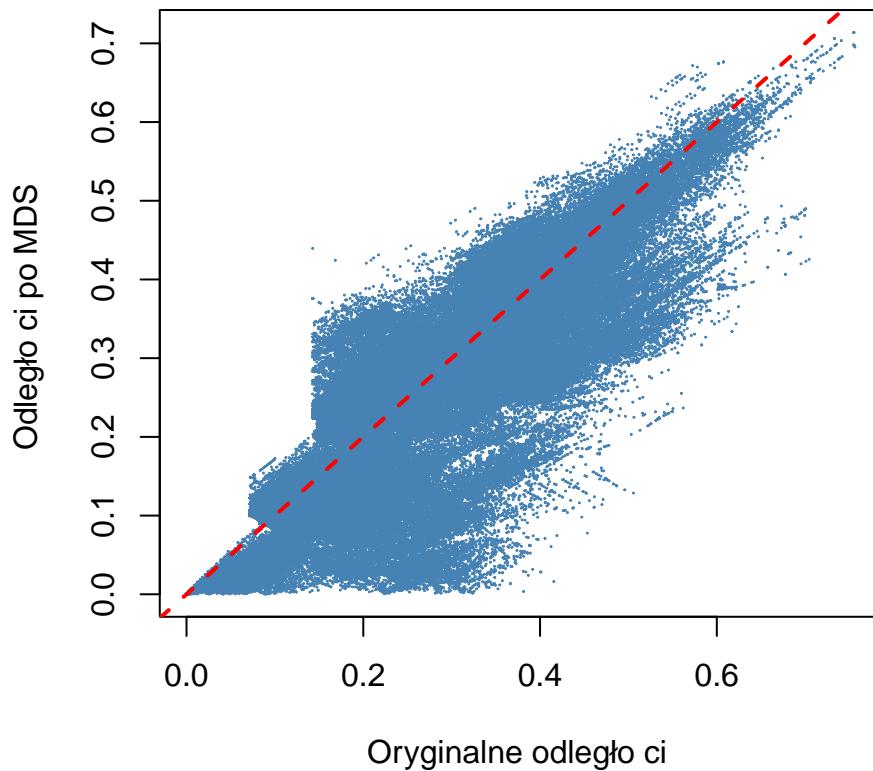
2.6 Wnioski końcowe

Na podstawie przeprowadzonej analizy PCA stwierdzono, że standaryzacja zmiennych ciągłych nie wpłynęła znacząco na strukturę głównych składowych. Może to sugerować, że dane te były już wcześniej porównywalne pod względem skali. W wyniku analizy ustalono, że do zachowania 90% całkowitej wariancji wystarcza 9 spośród 17 składowych, co oznacza, że część cech może być redundantna. Uzyskane wyniki wskazują, że dane zawierają pewien poziom powtarzalności informacji, co może być wykorzystane w dalszej analizie, na przykład przez ograniczenie liczby zmiennych wejściowych w modelach predykcyjnych. Redukcja wymiarowości może przyczynić się do uproszczenia modeli, skrócenia czasu obliczeń oraz poprawy ich interpretowalności.

3 Skalowanie wielowymiarowe (Multidimensional Scaling (MDS))

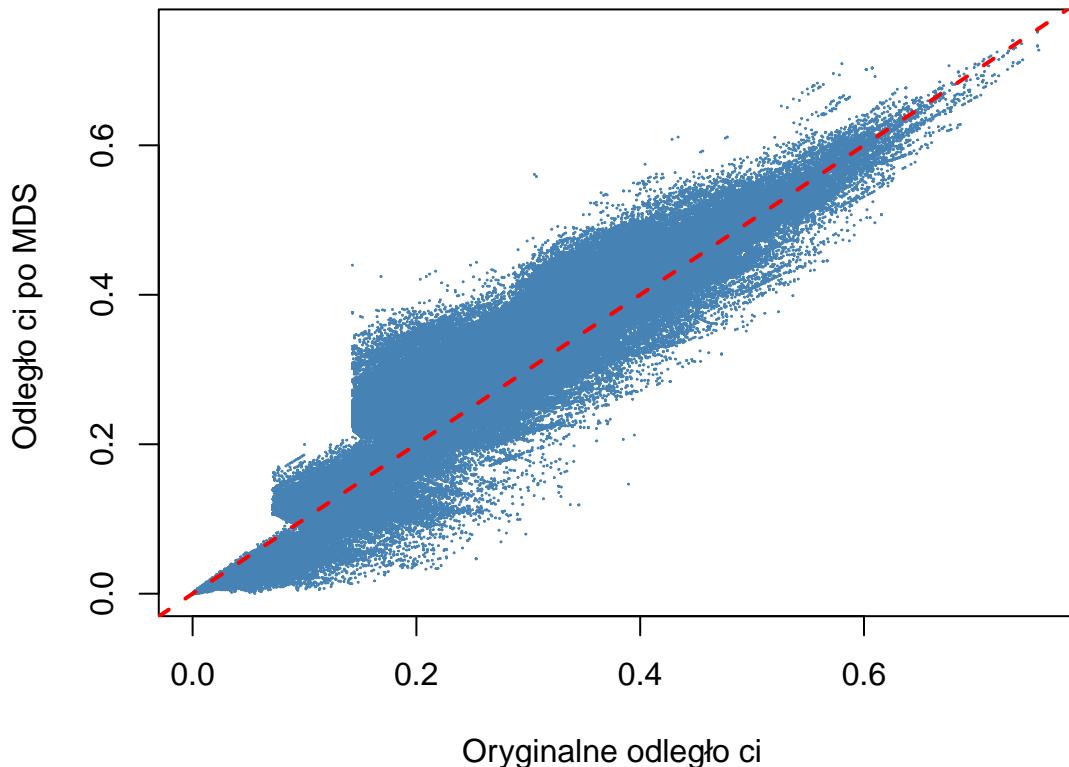
3.1 Redukcja wymiaru na bazie MDS

Diagram Sheparda (dla wymiaru 2)



Rysunek 11: Diagram Sheparda po zastosowaniu MDS dla k = 2

Diagram Sheparda (dla wymiaru 3)

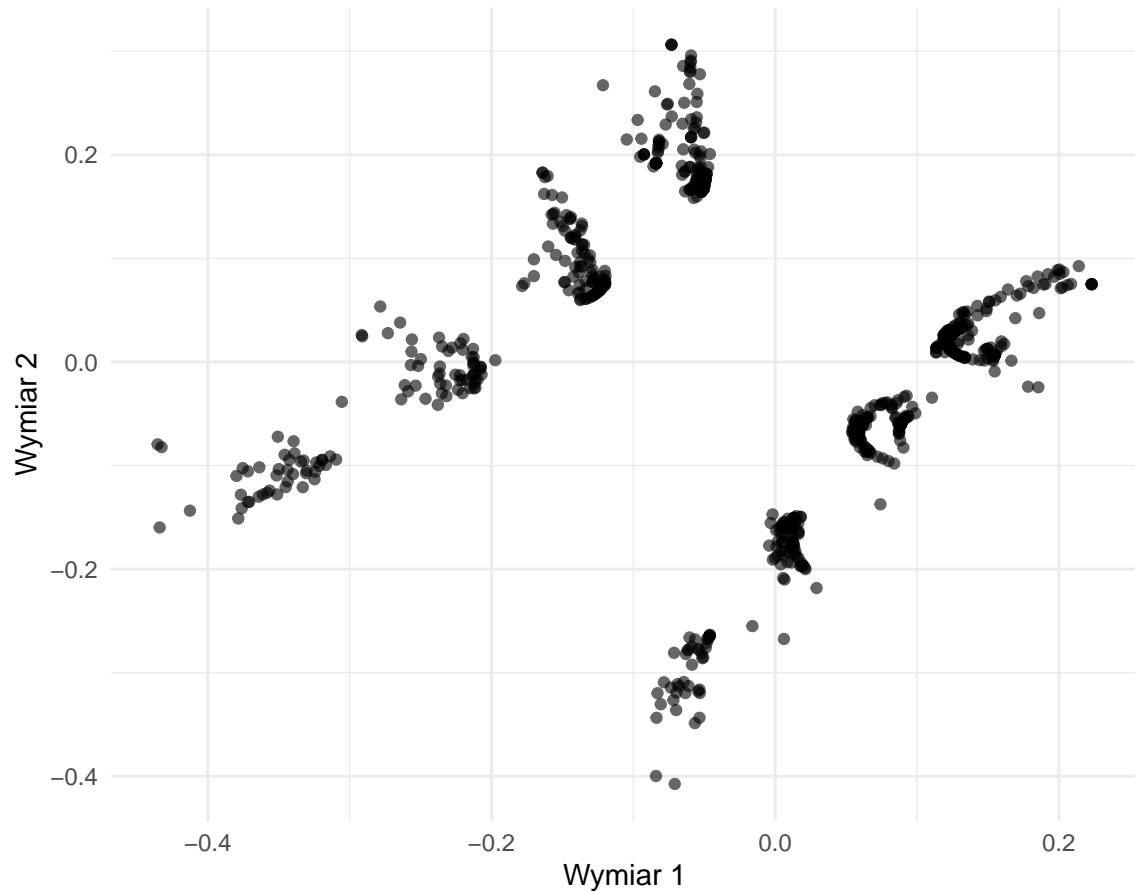


Rysunek 12: Diagram Sheparda po zastosowaniu MDS dla $k = 2$

Analiza diagramu Sheparda wskazuje, że skalowanie klasyczne w przestrzeni dwuwymiarowej charakteryzuje się niską jakością odwzorowania – obserwacje wykazują istotne odchylenia od linii idealnej, co świadczy o niewystarczającej dokładności odwzorowania struktur odległościowych. Zwiększenie liczby wymiarów do trzech prowadzi do zauważalnej poprawy – punkty na wykresie są znacznie bliższe linii odniesienia, co potwierdza wyższą trafność rekonstrukcji oryginalnych relacji odległościowych w przestrzeni trójwymiarowej.

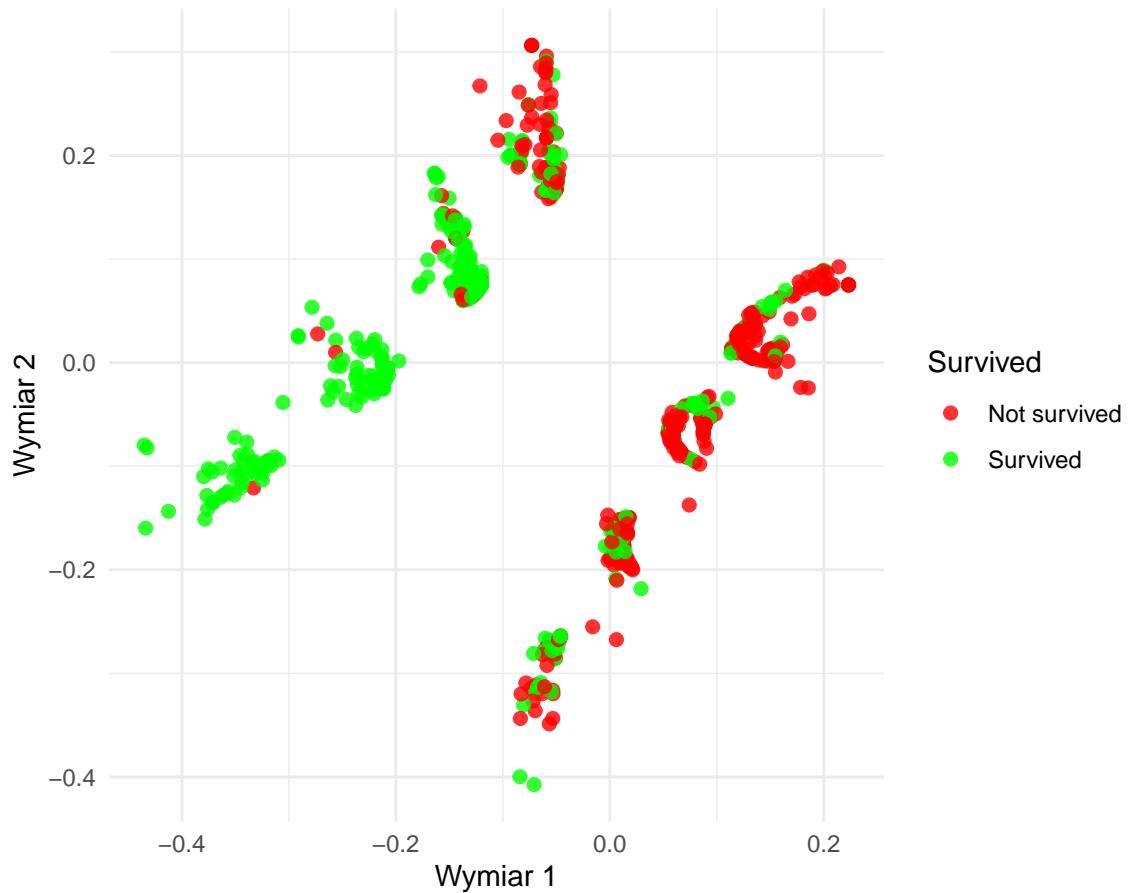
3.2 Wizualizacja danych po zastosowaniu skalowania wielowymiarowego

MDS ($k = 2$): Wykres punktowy bez grupowania



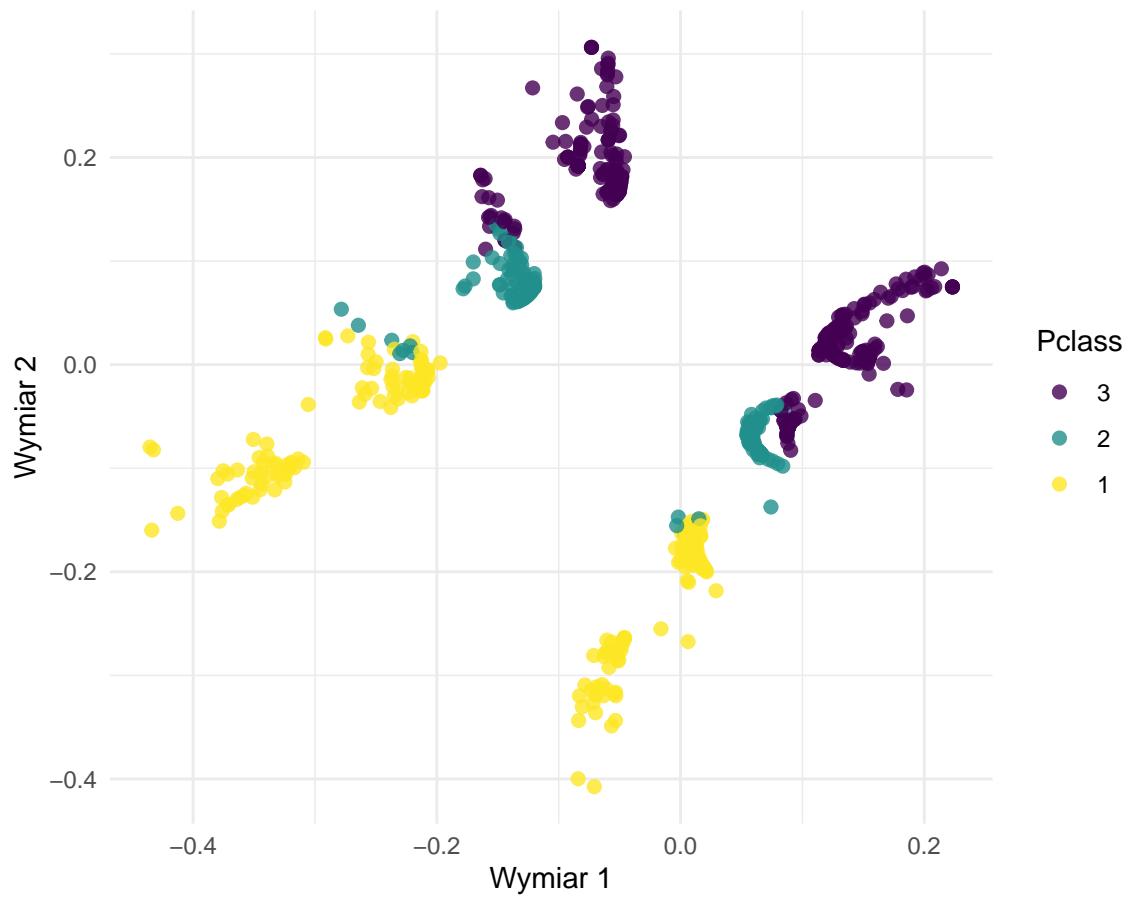
Rysunek 13: Wykres rozrzutu po zastosowaniu MDS dla $k = 2$

MDS ($k = 2$): Wykres rozrzutu wg zmiennej Survived



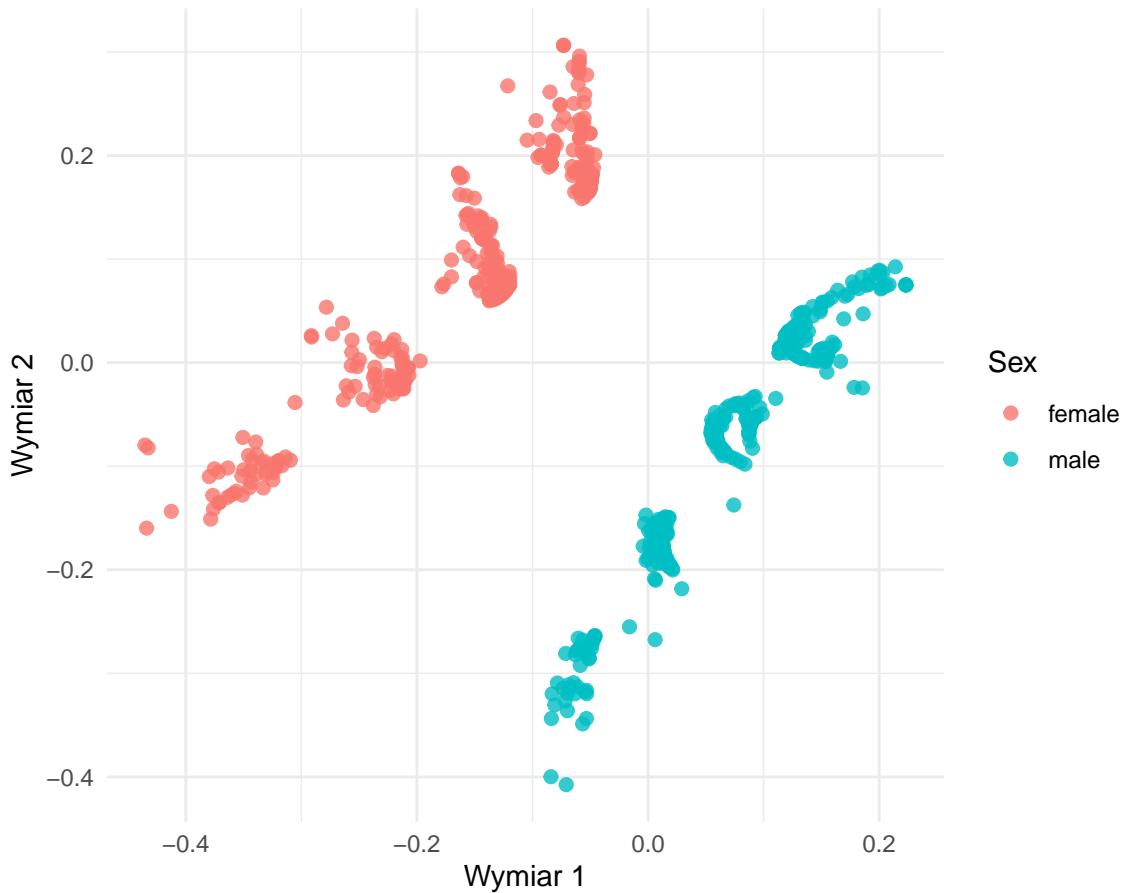
Rysunek 14: Wykres rozrzutu po zastosowaniu MDS dla $k = 2$ z grupowaniem według zmiennej 'Survived'

MDS ($k = 2$): Wykres rozrzutu wg zmiennej Survived



Rysunek 15: Wykres rozrzutu po zastosowaniu MDS dla $k = 2$ z grupowaniem według zmiennej 'Pclass'

MDS ($k = 2$): Wykres rozrzutu wg zmiennej Survived



Rysunek 16: Wykres rozrzutu po zastosowaniu MDS dla $k = 2$ z grupowaniem według zmiennej 'Sex'

Na wykresie 14 widoczny jest wyraźny podział danych na dwa skupiska. Po oznaczeniu przynależności poszczególnych obserwacji do grup odpowiadających wartościom zmiennej **Survived**, można zauważać, że w górnym skupisku dominują osoby, które przeżyły (**Survived == 1**), natomiast w dolnym – osoby, które nie przeżyły (**Survived == 0**). Na wykresie można dostrzec kilka punktów oddalonych od głównych skupisk, np. pojedyncze punkty znajdujące się po lewej stronie wykresu lub w dolnej części, z dala od skoncentrowanych grup. Są to obserwacje odstające, które mogą reprezentować osoby o nietypowym zestawie cech (np. bardzo młody wiek, nietypowa kombinacja klasy i płci) lub wskazywać na błędy danych lub wyjątkowe przypadki. Cechy użyte do analizy MDS w pewnym stopniu odzwierciedlają różnice związane z przeżyciem, choć nie jest to rozdział idealny – kolory są częściowo wymieszane w niektórych skupiskach. Po uwzględnieniu podziału według płci 16 widać, że wśród osób ocalałych największą grupę stanowią kobiety. Dalsza analiza z uwzględnieniem klasy podróży 15 (zmienna **Pclass**) pokazuje, że wśród kobiet, które przeżyły, zdecydowaną większość stanowią pasażerki podróżujące pierwszą klasą. W przypadku mężczyzn natomiast klasa podróży nie miała istotnego wpływu na przeżywalność – liczba ocalałych była zbliżona niezależnie od klasy.