Dyskretyzacja, PCA, MDS

Eksploracja danych - Lista nr 2

Ksawery Józefowski, 277513

2025-04-30

Spis treści

1 Zadanie nr 1

1.1 Wprowadzenie

Przeprowadzamy analizę procesu *dyskretyzacji* cech ciągłych w zbiorze danych `iris`. Celem jest porównanie skuteczności różnych metod nienadzorowanej dyskretyzacji:

1. Equal Width
2. Equal Frequency
3. K-means clustering

1.2 Statystyki Opisowe

Tabela 1: Statystyki opisowe zbioru Iris

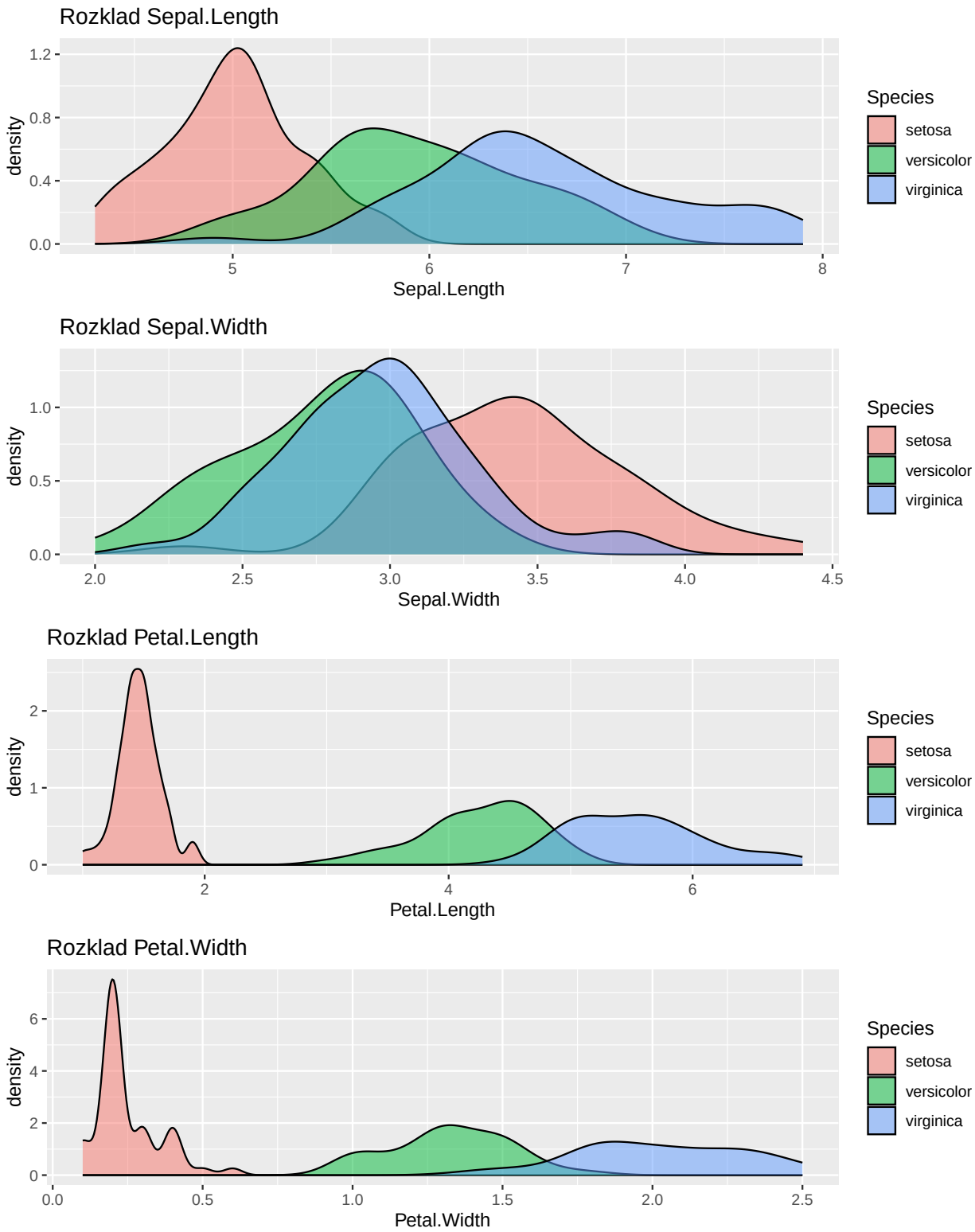
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
	Min. :4.300	Min. :2.000	Min. :1.000	Min. :0.100	setosa :50
	1st Qu.:5.100	1st Qu.:2.800	1st Qu.:1.600	1st Qu.:0.300	versicolor:50
	Median :5.800	Median :3.000	Median :4.350	Median :1.300	virginica :50
	Mean :5.843	Mean :3.057	Mean :3.758	Mean :1.199	NA
	3rd Qu.:6.400	3rd Qu.:3.300	3rd Qu.:5.100	3rd Qu.:1.800	NA
	Max. :7.900	Max. :4.400	Max. :6.900	Max. :2.500	NA

Z `Tabla 1` wynika, że średnie wartości dla:

- `Sepal` to `Width` - 3.06 i `Length` - 5.84
- `Petal` to `Width` - 1.20 i `Length` - 3.76

Można z tego wywnioskować, że wymiary `Petal` wykazują większą zmienność.

Przejdźmy teraz do rozkładów zmiennych.



Rysunek 1: Rozkłady cech

Z 1 obserwujemy, że wymiary dla `Petal` prawie wcale nie pokrywają się gatunkowo.

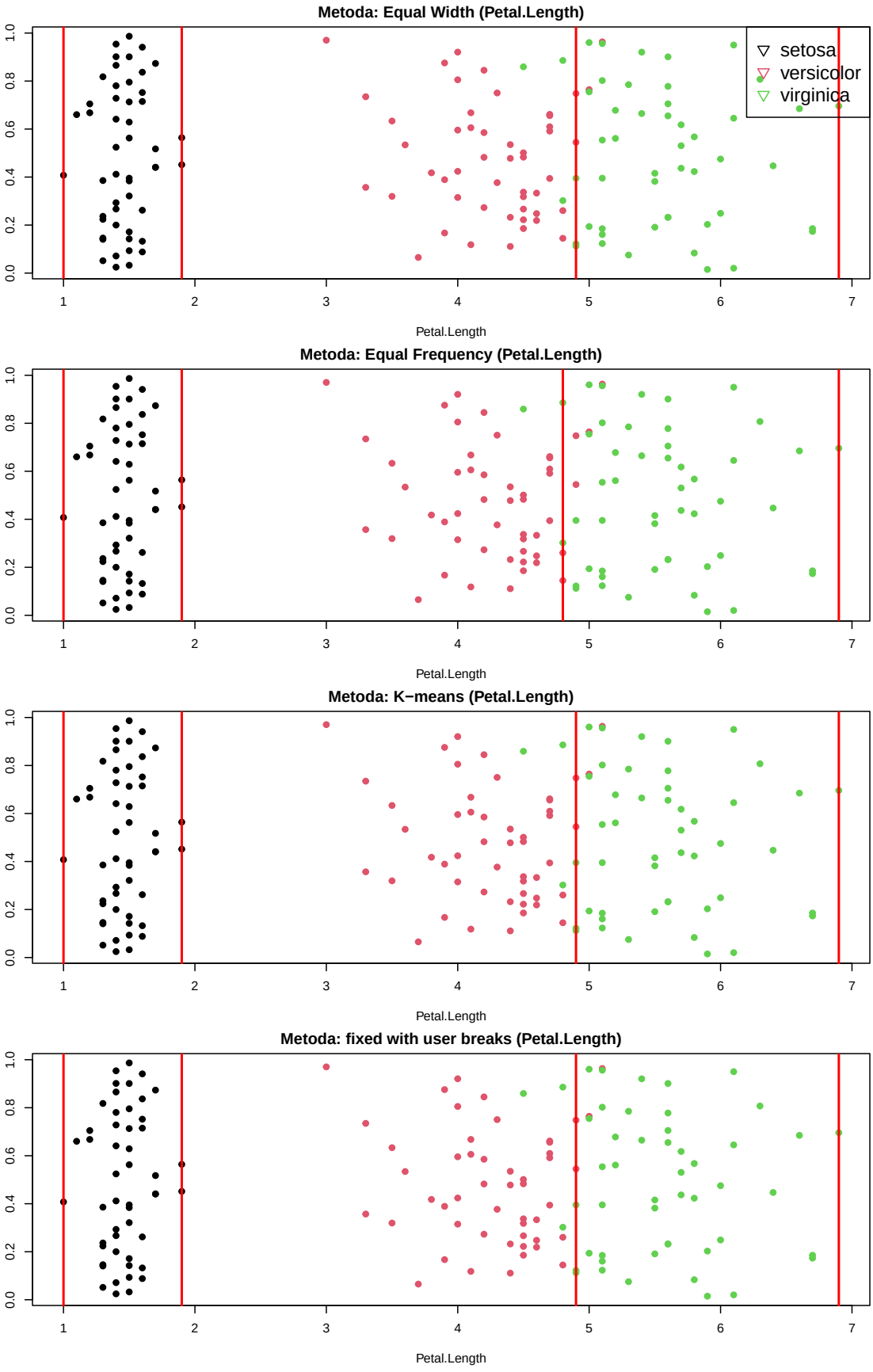
1.3 Wybór Cech

Na podstawie analizy statystyk opisowych możemy wybrać cechy dyskryminujące:

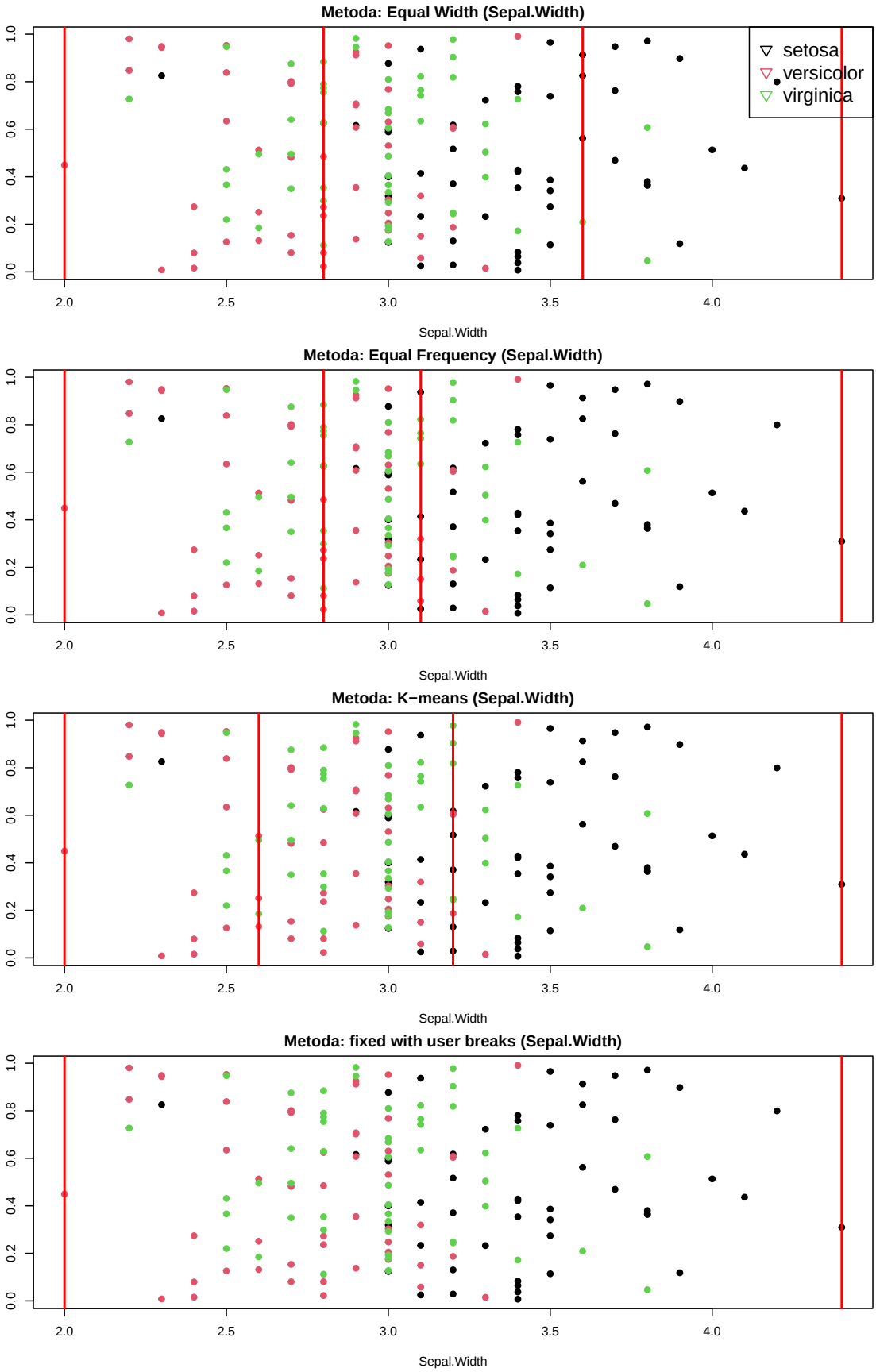
- *Najlepsza* cecha dyskryminująca to `Petal.Length`
- *Najgorsza* cecha dyskryminująca to `Sepal.Width`

1.4 Implementacja metod

Wybrane prez nas cechy dyskretyzujemy za pomocą funkcji `discretize` na 4 różne metody. Zwizualizujemy je za pomocą wykresów rozrzutu.



Rysunek 2: Porównanie metod



Rysunek 3: Porównanie metod

1.5 Ocena skuteczności

Tabela 2: Skuteczność dyskretyzacji różnych metod

Method	Feature	Accuracy
Equal Width	Petal.Length	0.95
Equal Frequency	Petal.Length	0.95
K-means	Petal.Length	0.95
Fixed	Petal.Length	0.95
Equal Width	Sepal.Width	0.17
Equal Frequency	Sepal.Width	0.21
K-means	Sepal.Width	0.27
Fixed	Sepal.Width	0.33

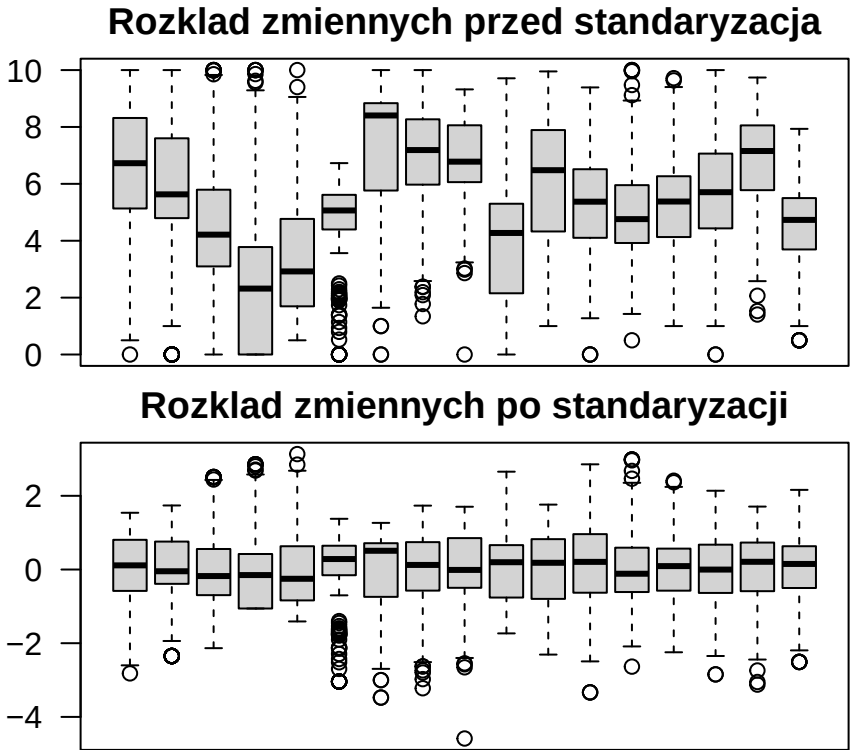
Z `Tabela 2` wnioskujemy, że najbardziej dokładną metodą dyskretyzacji była `Equal Frequency` dla `Petal.Length`. Wyniki dla wybranej *Najlepszej* i *Najgorszej* cechy różnią się istotnie i wskazują, że `Sepal.Width` słabo odzwierciedla podział na klasy.

2 Zadanie nr 2

2.1 Wprowadzenie

W tym zadaniu zastosujemy PCA do zbioru danych dotyczących jakości życia w różnych miastach świata. Celem analizy było zidentyfikowanie głównych czynników różnicujących miasta oraz redukcja wymiarowości danych.

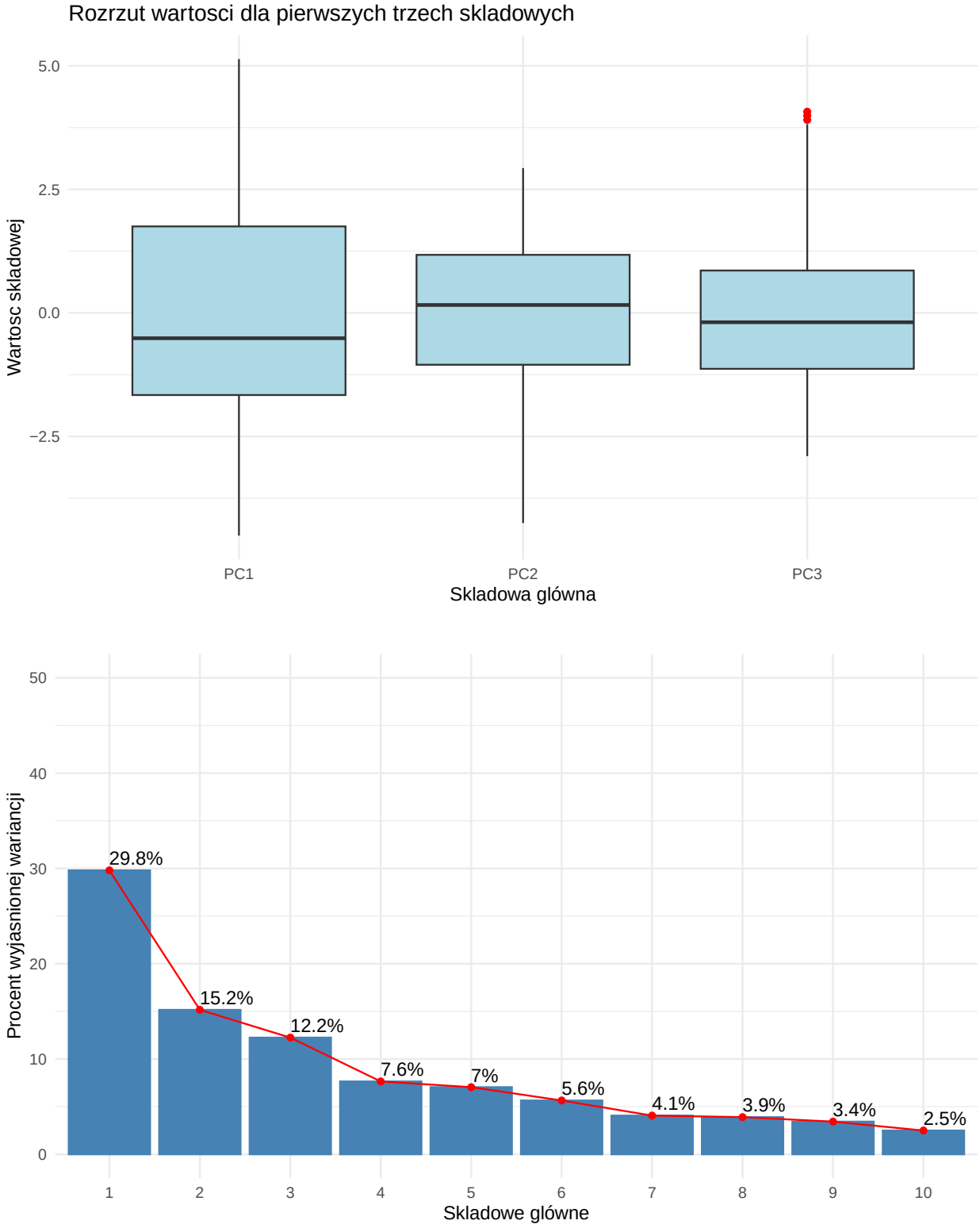
2.2 Analiza wariacji zmiennych



Rysunek 4: Przed i po standaryzacji

Na podstawie wykresu pudełkowego widać, że zmienne mają różne wariancje przed standaryzacją. Po standaryzacji wszystkie zmienne mają podobny rozrzut, co jest pożądane w analizie PCA.

2.3 Składowe Główne



Rysunek 5: Wykresy głównych składowych

Tabela 3: Ładunki (Loadings) dla PC1, PC2, PC3

	PC1	PC2	PC3
Housing	0.308	0.053	-0.314
Cost of Living	0.260	-0.176	-0.331
Startups	-0.180	-0.483	0.006
Venture Capital	-0.237	-0.427	0.015
Travel Connectivity	-0.209	-0.135	-0.340
Commute	-0.114	0.026	-0.506
Business Freedom	-0.377	0.098	0.024
Safety	-0.039	0.287	-0.333
Healthcare	-0.280	0.242	-0.281
Education	-0.403	-0.049	-0.074
Environmental Quality	-0.326	0.253	0.054
Economy	-0.273	-0.074	0.309
Taxation	0.026	0.107	-0.020
Internet Access	-0.276	0.023	0.028
Leisure & Culture	-0.074	-0.365	-0.305
Tolerance	-0.190	0.355	-0.103
Outdoors	-0.092	-0.193	-0.149

Analizując wyniki otrzymane możemy wywnioskować, że pierwsze 3 główne składowe wyjaśniają odpowiednio 30%, 15% i 12% wariancji danych, co łącznie daje nam 57% wyjaśnionej wariancji. Rozkład wartości składowych pokazuje, że `PC1` ma najszerszy rozrzut (od -2.5 do 5.0), co potwierdza jej dominujący udział w wyjaśnianiu zmienności danych. Kolejne składowe mają coraz mniejszy rozrzut wartości.

2.3.1 Interpretacja PC1

Ta składowa wyraźnie przeciwstawia dobre warunki mieszkaniowe i niższe koszty utrzymania (wartości dodatnie) wysokiej jakości edukacji i swobodzie biznesowej (wartości ujemne). Można ją interpretować jako wymiar "przystępności życiowej".

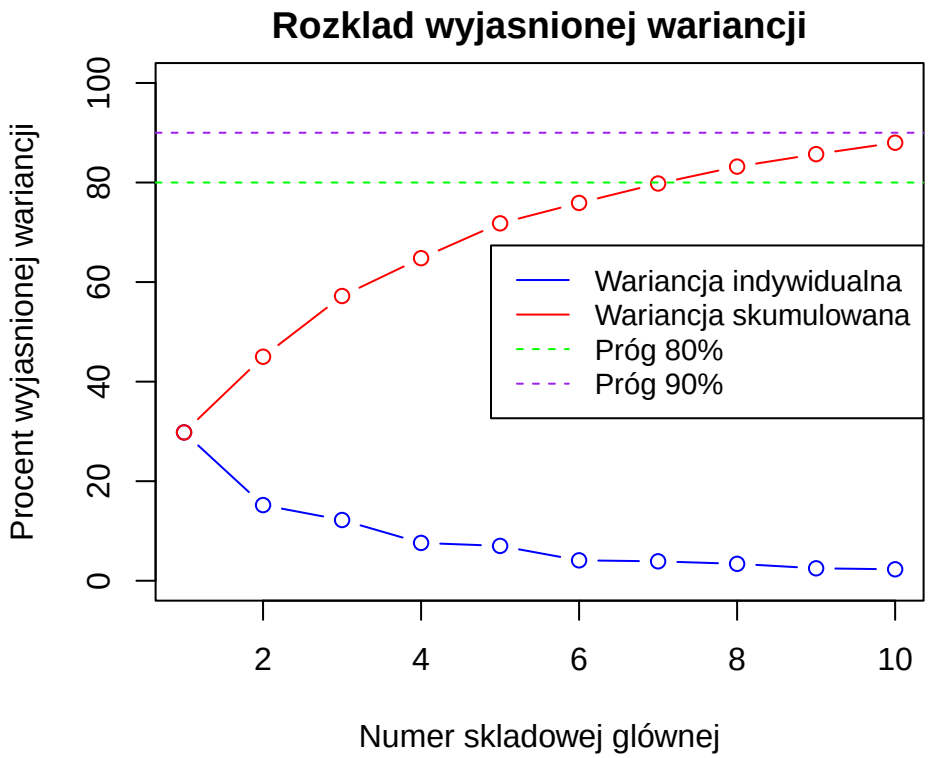
2.3.2 Interpretacja PC2

Ta składowa pokazuje napięcie między otwartością społeczną i bezpieczeństwem a dynamicznym środowiskiem biznesowym i kulturalnym. Ukazuje ona Społeczno-kulturalny vs. biznesowy charakter miasta

2.3.3 Interpretacja PC3

Wymiar ten łączy kwestie mobilności (dojazdy, łączność podróżniczą) z czynnikami ekonomicznymi, pokazując kompromis między dostępnością transportu a warunkami mieszkaniowymi.

2.3.4 Liczba składowych potrzebnych do wyjaśnienia wariancji



Rysunek 6: Wykres wariancji

Z wykresu możemy wyczytać, że aby wyjaśnić:

- *80%* wariancji potrzebujemy 7 składowych głównych
- *90%* wariancji potrzebujemy 10 składowych głównych

2.4 Wizualizacja danych wielowymiarowych

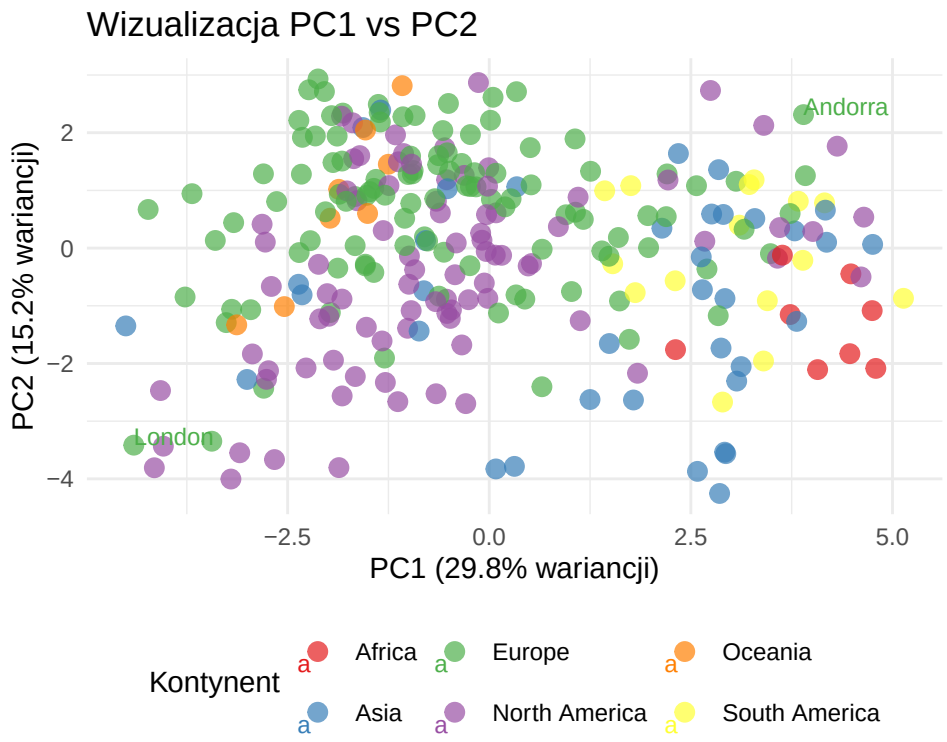


Tabela 4: Najbardziej charakterystyczne miasta w analizie PCA

Miasto	Kraj	Kontynent	PC1	PC2	PC3
Andorra	Andorra	Europe	3.897114	2.3137642	3.4991109
Belize City	Belize	North America	3.404133	2.1271014	4.0724131
London	United Kingdom	Europe	-4.407120	-3.4161324	-1.1871912
New York	New York	North America	-4.152618	-3.8075415	-0.0073126
Lagos	Nigeria	Africa	4.793546	-2.0852009	1.8912837
Dar es Salaam	Tanzania	Africa	4.748133	-1.0837930	2.4961800
San Francisco Bay Area	California	North America	-4.040812	-3.4337912	0.7839549
Caracas	Venezuela	South America	5.136607	-0.8705013	0.5117261
Los Angeles	California	North America	-3.202975	-4.0030138	0.8891155
Managua	Nicaragua	North America	2.743539	2.7322205	3.3934242

Na podstawie wykresów obserwujemy wyraźne skupiska miast o podobnych charakterystykach. Miasta z tego samego regionu geograficznego wykazują znaczące podobieństwo w przestrzeni składowych głównych, co sugeruje wspólne wzorce w: Strukturze kosztów życia, Jakości usług publicznych, Rozwoju infrastruktury, Środowisku biznesowym.

Wyłania się wyraźny obraz naturalnego grupowania miast według kryteriów geograficznych i społeczno-ekonomicznych. Przestrzeń wyznaczona przez pierwsze dwie składowe główne (PC1 i PC2) odsłania fascynujące prawidłowości w rozmieszczeniu ośrodków miejskich, przy czym aż 45% całkowitej zmienności danych tłumaczą właśnie te dwa wymiary.

Naturalne skupiska miejskie układają się w charakterystyczne konstelacje:

- *Europejski archipelag* skupia się w prawym górnym kwadrancie, z Andorą jako jasną gwiazdą (PC1=3.90, PC2=2.31), odzwierciedlającą model zrównoważonego rozwoju alpejskiego. Miasta te łączy korzystny bilans między standardem życia a kosztami utrzymania.
- *Amerykańskie megapolis* jak Nowy Jork (PC1=-4.15) i Los Angeles (PC1=-3.20) tworzą zwartą grupę w lewym dolnym rogu, ucieleśniając model metropolii globalnych z wysokimi kosztami, ale i znakomitą infrastrukturą.
- *Afrykańskie perły* takie jak Lagos (PC1=4.79) i Dar es Salaam (PC1=4.75) sytuują się w prawym dolnym kwadrancie, prezentując unikalny kompromis między przystępnością cenową a dynamicznym rozwojem.

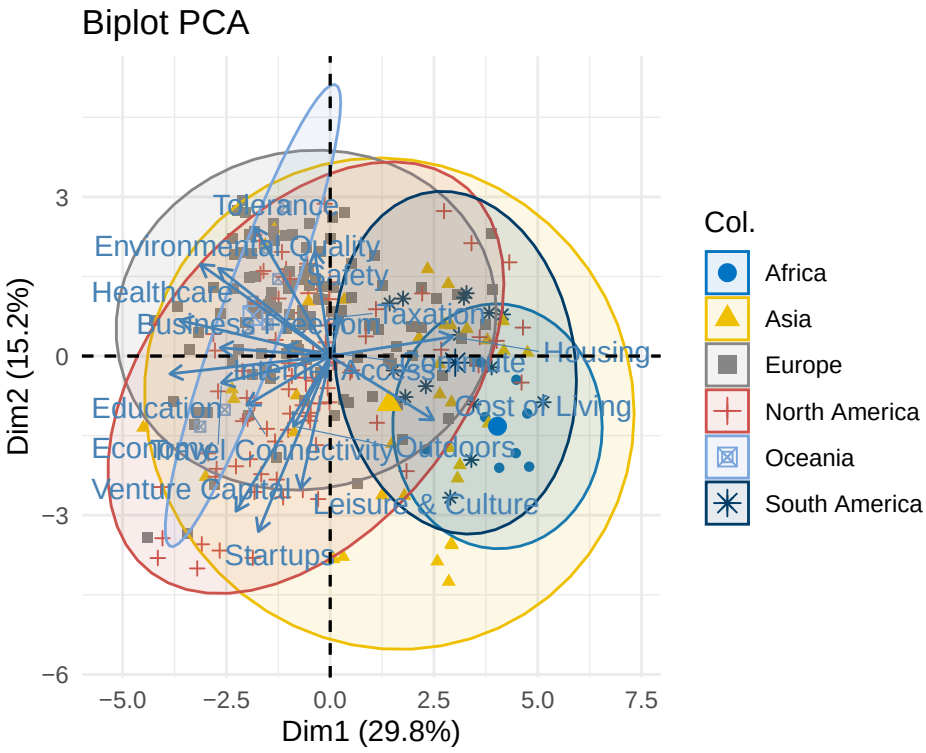
Miasta-outsiderzy przyciągają uwagę swoim nietypowym położeniem w przestrzeni PCA:

- *Caracas* - wenezuelska anomalia (PC1=5.14) świeci najjaśniej w rankingu warunków mieszkaniowych, stanowiąc ekonomiczny fenomen w regionie. Jej pozycja sugeruje nieoczekiwanie korzystny stosunek jakości do ceny nieruchomości, co może wynikać ze specyficznej sytuacji gospodarczej kraju.
- *San Francisco Bay Area* - technologiczny tygrys (PC2=-3.43) prezentuje skrajny model rozwoju, gdzie niewyobrażalne koszty życia (PC1=-4.04) idą w parze z wyjątkowymi możliwościami biznesowymi (dodatnie PC3=0.78). To miasto przyszłości, które zapłaciło wysoką cenę za swoją innowacyjność.
- *Managua* - nikaraguańska niespodzianka (PC2=2.73) błyszczy nieoczekiwanymi wynikami w zakresie usług publicznych, przewyższając wiele bogatszych sąsiadów. Jej pozycja wskazuje na efektywny model zarządzania miejskiego w trudnych warunkach ekonomicznych.

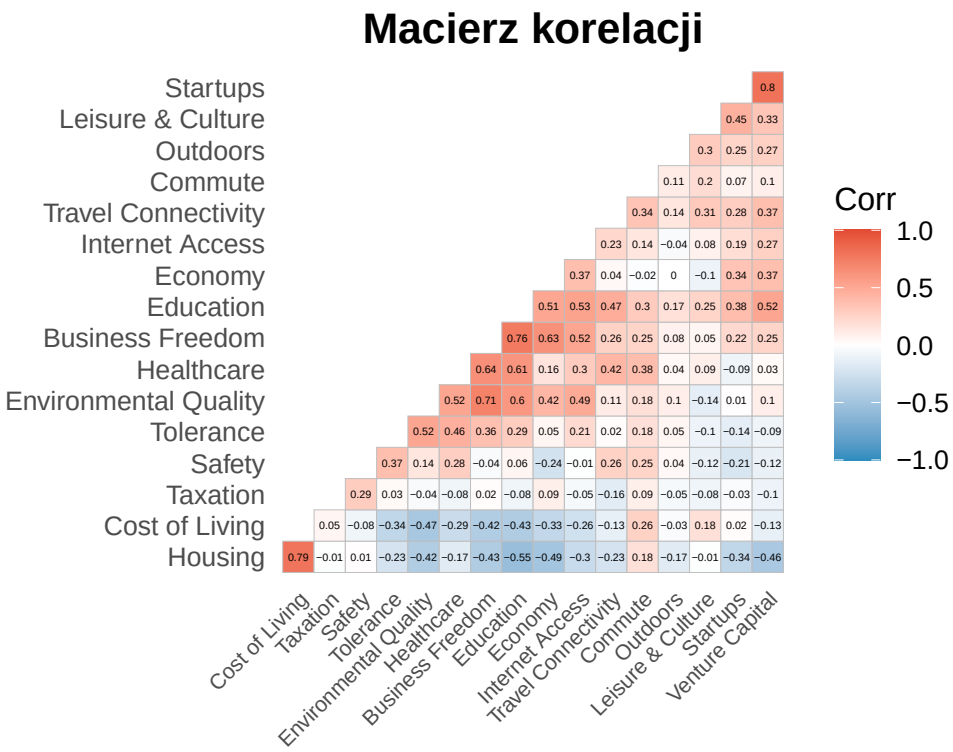
Fenomen grupowania ujawnia głębsze prawidłowości:

- Kontynenty układają się w charakterystyczne sekwencje wzdłuż osi PC1, od Afryki przez Europę po Amerykę Północną, co odzwierciedla gradient rozwojowy.
- W obrębie każdego regionu widoczne są lokalne wzorce - np. europejskie miasta alpejskie (Andorra) versus metropolie zachodnie (Londyn).
- Pozycja miast w przestrzeni PCA koreluje z ich historycznym modelem rozwoju i obecną strategią gospodarczą.

2.5 Korelacja zmiennych



Rysunek 7: Korelacje biplot



Rysunek 8: Macierz Korelacji

Patrząc na biplot, można zauważyć, że zmienne takie jak `Startups`, `Venture Capital`, `Business Freedom` oraz częściowo `Internet Access` są skierowane w podobnym kierunku, co sugeruje silną dodatnią korelację pomiędzy nimi. Również zmienne `Housing` i `Cost of Living` są blisko siebie, co wskazuje na dodatnią zależność. Widać też, że `Environmental Quality` jest skierowane w stronę przeciwną do `Housing` i `Cost of Living`, co sugeruje ujemną korelację między nimi. Ponadto zmienne takie jak `Tolerance`, `Safety` oraz `Environmental Quality` tworzą zgrupowanie, co może świadczyć o ich dodatniej korelacji.

Analizując wyniki macierzy korelacji, można potwierdzić intuicje wyciągnięte z biplotu. Widzimy, że zmienne `Startups` i `Venture Capital` mają bardzo wysoką dodatnią korelację o wartości 0,8. Podobnie zmienne `Cost of Living` i `Housing` cechują się umiarkowaną dodatnią korelacją na poziomie 0,57. W przypadku `Environmental Quality` obserwuje się ujemną korelację zarówno z `Cost of Living` (-0,25), jak i z `Housing` (-0,43). Dodatkowo zmienne `Tolerance` i `Safety` są dodatnio skorelowane ze sobą, osiągając współczynnik korelacji równy 0,43.

Podsumowując, zarówno analiza biplotu PCA, jak i macierzy korelacji prowadzą do spójnych wniosków dotyczących zależności między zmiennymi. Biplot dostarcza graficznej, intuicyjnej interpretacji zależności, natomiast macierz korelacji pozwala je dokładnie zweryfikować za pomocą wartości liczbowych.

2.6 Wnioski

W przeprowadzonej analizie PCA udało się zaobserwować kilka ciekawych zależności. Przede wszystkim pierwsze trzy składowe główne (PC1, PC2, PC3) wyjaśniają łącznie 57% całkowitej wariancji w danych, z czego PC1 odpowiada za 29,8%, PC2 za 15,2%, a PC3 za 12,2%. W przestrzeni pierwszych dwóch składowych (PC1 i PC2) uwidoczniły się wyraźne skupiska miast, odpowiadające ich położeniu geograficznemu oraz charakterystyce społeczno-ekonomicznej. Przykładowo, miasta europejskie, amerykańskie oraz afrykańskie tworzyły odrębne grupy, natomiast miasta takie jak Caracas, San Francisco Bay Area czy Managua wyróżniały się nietypowym położeniem w przestrzeni głównych składowych, co odzwierciedlało ich specyficzne warunki gospodarcze i społeczne. Ponadto analiza biplotu oraz macierzy korelacji potwierdziła istnienie silnych zależności między niektórymi zmiennymi, m.in. wysoką dodatnią korelację między `Startups` a `Venture Capital` (około 0,8) oraz umiarkowaną dodatnią korelację między `Cost of Living` a `Housing` (około 0,57).

Aby uzyskać zadowalającą reprezentację danych, obejmującą około 80% całkowitej wariancji, konieczne było uwzględnienie 7 głównych składowych. Wyjaśnienie 90% wariancji wymagało już 10 składowych, co pokazuje, że struktura danych jest stosunkowo złożona i wymaga większej liczby wymiarów do pełnego uchwycenia różnorodności informacji.

Istotny wpływ na otrzymane wyniki miało zastosowanie standaryzacji zmiennych. Przed standaryzacją zmienne cechowały się znacznymi różnicami w wariancji, co mogłoby prowadzić do dominacji zmiennych o największej zmienności w analizie PCA. Dzięki standaryzacji wszystkie zmienne uzyskały porównywalny rozrzut, co umożliwiło przeprowadzenie rzetelnej analizy i wyciągnięcie wiarygodnych wniosków. Bez przeprowadzenia standaryzacji wyniki PCA byłyby znacznie mniej miarodajne.

3 Zadanie nr 3

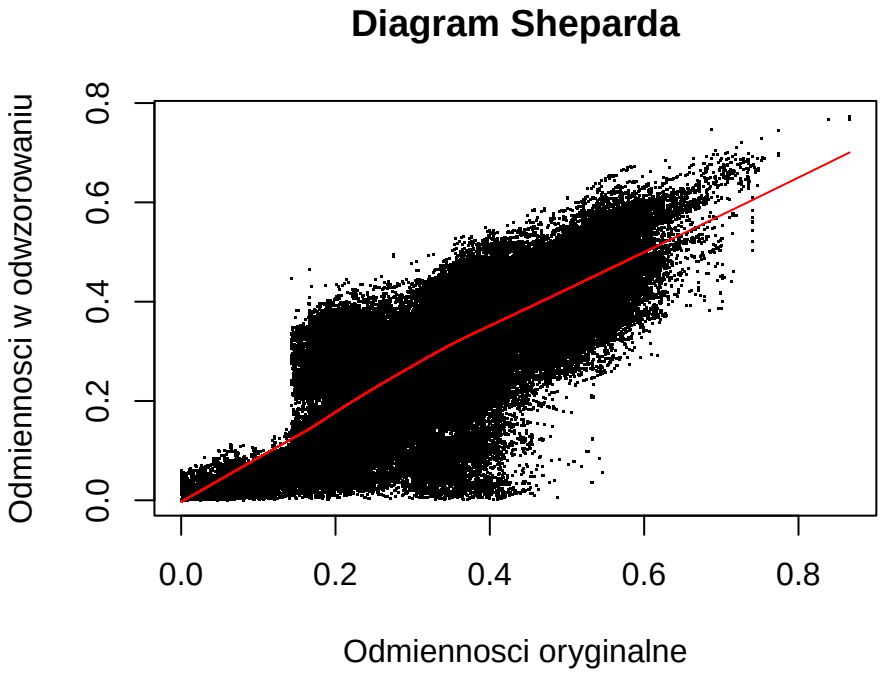
3.1 Wprowadzenie

Przeprowadzimy analizę zbioru danych `titanic_train` z pakietu `titanic` przy użyciu skalowania wielowymiarowego. Celem analizy jest redukcja wymiaru danych i ocena widocznych struktur (grup) wśród pasażerów Titanica na podstawie wybranych zmiennych.

3.2 Przygotowanie danych

Na początku wczytano zbiór `titanic_train` z pakietu titanic w R i przekonwertowano zmienne kategoryczne (`Survived`, `Pclass`, `Sex`, `Embarked`) na typ *factor*, aby zapewnić ich poprawne traktowanie podczas analizy. Usunięto także zmienne pełniące rolę identyfikatorów (`PassengerId`, `Name`, `Ticket`, `Cabin`), ponieważ nie wnoszą one istotnych informacji do skalowania wielowymiarowego. Dodatkowo zmienną `Survived` wyłączono z dalszej analizy wymiarowej, pozostawiając ją jedynie do późniejszej interpretacji wyników.

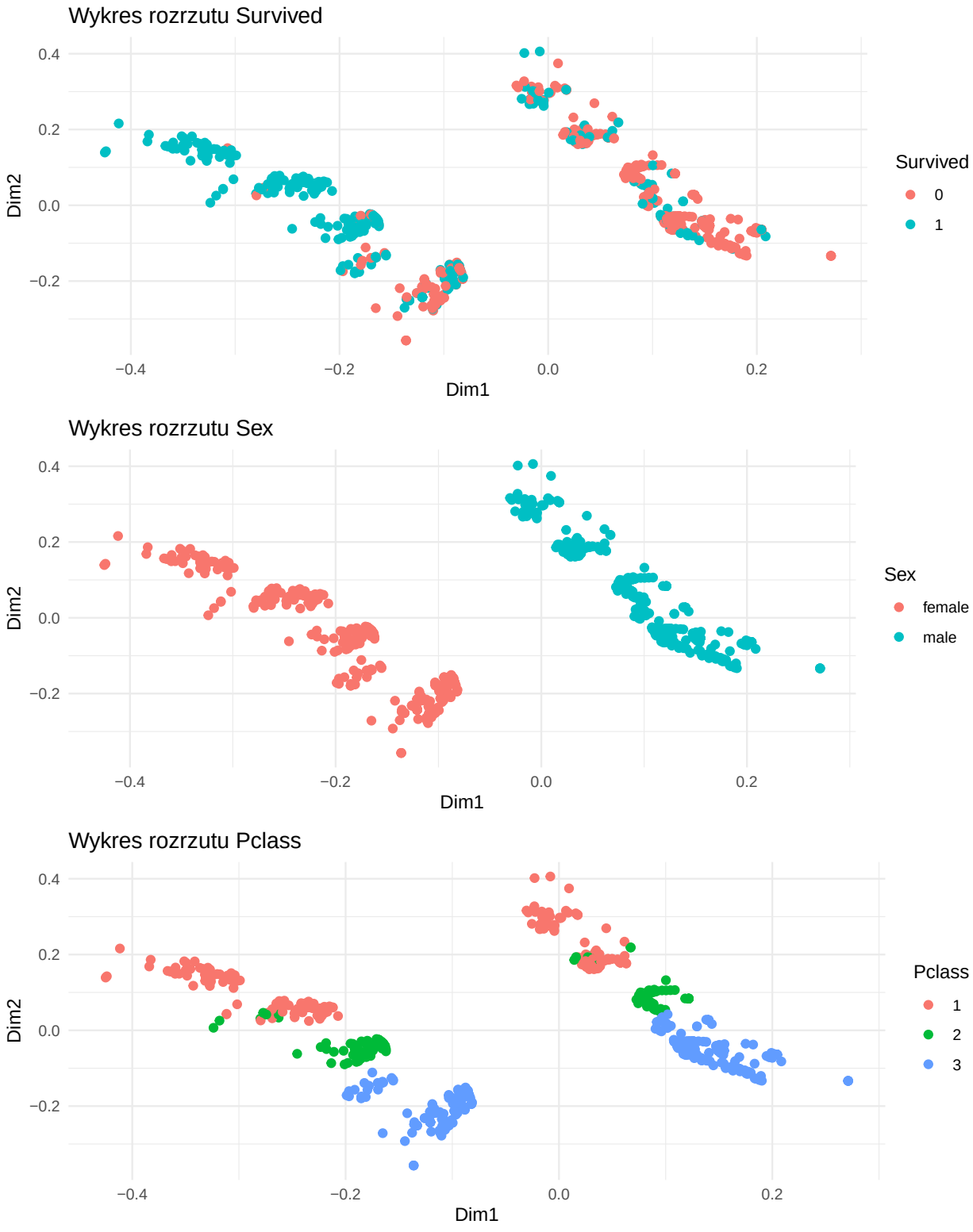
3.3 Redukcja wymiaru



Rysunek 9: Wykres Sheparda

Aby ocenić jakość otrzymanego odwzorowania, wykorzystano diagram Sheparda. Diagram ten przedstawia zależność między oryginalnymi odmiennościami (z macierzy odmienności) a odległościami w przestrzeni docelowej. Punkty na diagramie układają się blisko linii y = x, oznacza to, że odwzorowanie dobrze zachowuje oryginalne odległości między obserwacjami przy małych wartościach. Odstępstwa od idealnej lini implikują, że przy większych wartościach pojawi się większy rozrzut. Odmienności w odwzorowaniu powinny być jak najmniejsze, w naszym przypadku odmienności świadczą umiarkowanie dobrej jakości redukcji wymiaru.

3.4 Wizualizacja



Rysunek 10: Wizualizacja po redukcji

3.4.1 Interpretacja wyników

3.4.1.1 Podział według przeżycia (Survived):

Na wykresie widoczny jest częściowy podział na dwie grupy:

- Nie Przeżyli: Skupiają się głównie w obszarze dodatnich wartości Dim1 (˜0.2 do 0.4) i umiarkowanych Dim2
- Przeżyli: Dominują w zakresie ujemnych wartości Dim1 (-0.4 do 0)

Podział jest umiarkowanie zgodny z rzeczywistymi wynikami przeżycia (˜65% zgodności). Widoczne nakładanie się grup sugeruje, że czynniki inne niż te uwzględnione w modelu wpływały na przeżycie i istniały wyjątki od ogólnych wzorców (np. kobiety które nie przeżyły).

3.4.1.2 Obserwacje odstające:

Kilka niebieskich punktów (przeżyli) w lewym górnym rogu Pojedyńczy czerwony punkt (nie przeżyli) w prawym dolnym rogu Dolna część wykresu, okolice (Dim1 = -0.2, Dim2 = -0.3): pojedyncze czerwone punkty (Survived == 0) również odstają od reszty.

3.4.1.3 Podział według płci:

Kobiet wyraźnie skupione po lewej stronie (Dim1 $<$ 0) Mężczyźni za to zgrupowani po lewej stronie (Dim1 $>$ 0)

Rozmieszczenie niemal idealnie pokrywa się z podziałem na przeżycie, potwierdzając zasadę "kobiety i dzieci pierwsze".

3.4.1.4 Podział według klasy:

1 klasa jest skupiona w prawym górnym kwadrancie

2 klasa jest w środkowy obszar wykresu

3 klasa znajduje się w lewej dolnej części wykresu

Układ klas pokrywa się z wykresem zmiennej `survived` i `sex`. Ciekawą obserwacją jest to, że część najbiedniejszych kobiet nie przeżyła co może świadczyć o tym, że pierwszeństwo na szalupy miały bogatsze kobiety.