

Dyskretyzacja i redukcja wymiaru na podstawie danych iris, City Quality of Life Dataset, titanic_train

Eksploracja danych

Tomasz Warzecha, album 282261

2025-04-30

Spis treści

1	Dyskretyzacja cech ciągłych	1
1.1	Wczytanie danych	1
1.2	Wybór cech	2
1.3	Porównanie nienadzorowanych metod dyskretyzacji	4
1.4	Podsumowanie	11
2	PCA - analiza składowych głównych	12
2.1	Krótki opis zagadnienia	12
2.2	Przygotowanie danych	12
2.3	Wyznaczanie składowych głównych	13
2.4	Zmienna odpowiadająca poszczególnym składowym	15
2.5	Wizualizacja danych wielowymiarowych	17
2.6	Korelacja zmiennych	18
2.7	Wnioski końcowe	21
3	MSD - skalowanie wielowymiarowe	21
3.1	Wczytanie i przygotowanie danych	22
3.2	Redukcja wymiaru na bazie MDS	22
3.3	Wizualizacja danych	23

1 Dyskretyzacja cech ciągłych

W tym zadaniu pracujemy na danych iris (R-pakiet datasets). Zbiór danych zawiera wyniki pomiarów uzyskanych dla trzech gatunków irysów (tj. setosa, versicolor i virginica) i został udostępniony przez Ronaldą Fishera w roku 1936. Pomiary dotyczą długości oraz szerokości dwóch różnych części kwiatu – działki kielicha (ang. sepal) oraz płatka (ang. petal).

1.1 Wczytanie danych

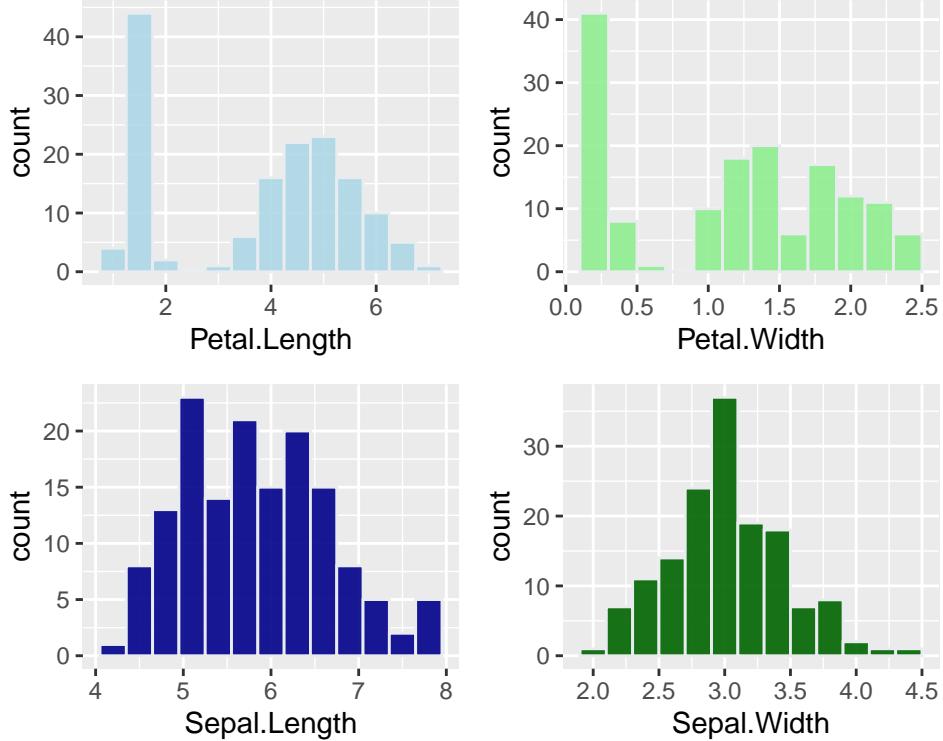
Tabela 1: Struktura danych

x	
Sepal.Length	numeric
Sepal.Width	numeric
Petal.Length	numeric
Petal.Width	numeric
Species	factor

Nasze dane mają 150 przypadków i 5 cech. W powyższej tabeli możemy zobaczyć wszystkie cechy oraz ich typy. Widzimy, że wszystkie zmienne zostały poprawnie rozpoznane. Nasze dane mają 4 cechy numeryczne oraz jedną jakościową. W żadnej z kolumn nie mamy braku wartości.

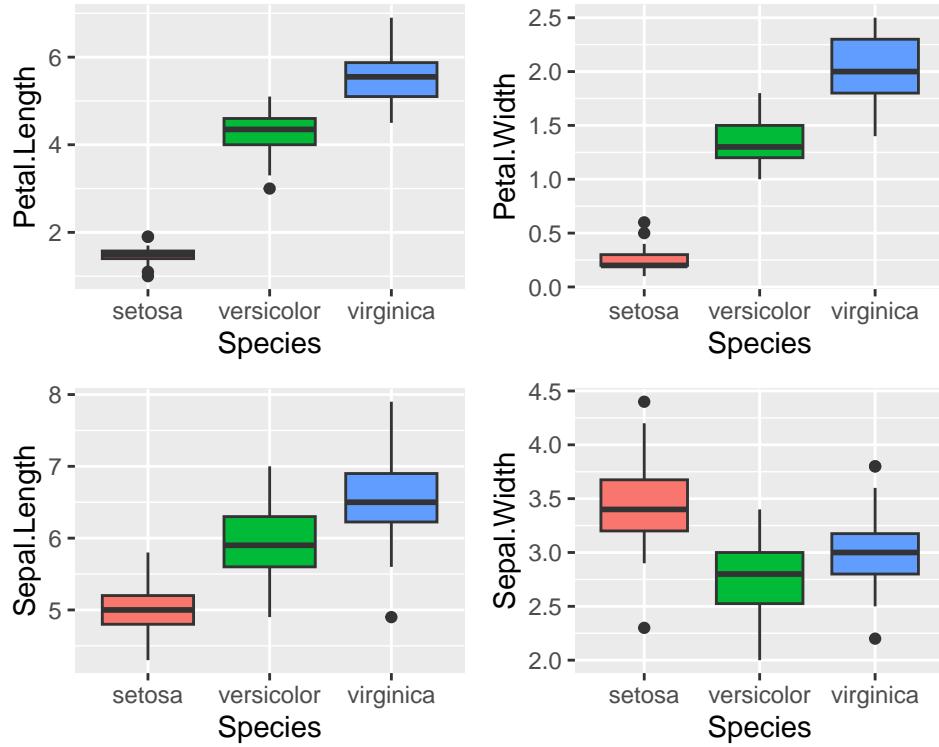
1.2 Wybór cech

Teraz przeanalizujemy nasze dane i wybierzymy zmienną o najgorszej i najlepszej zdolności dyskryminacyjnej.

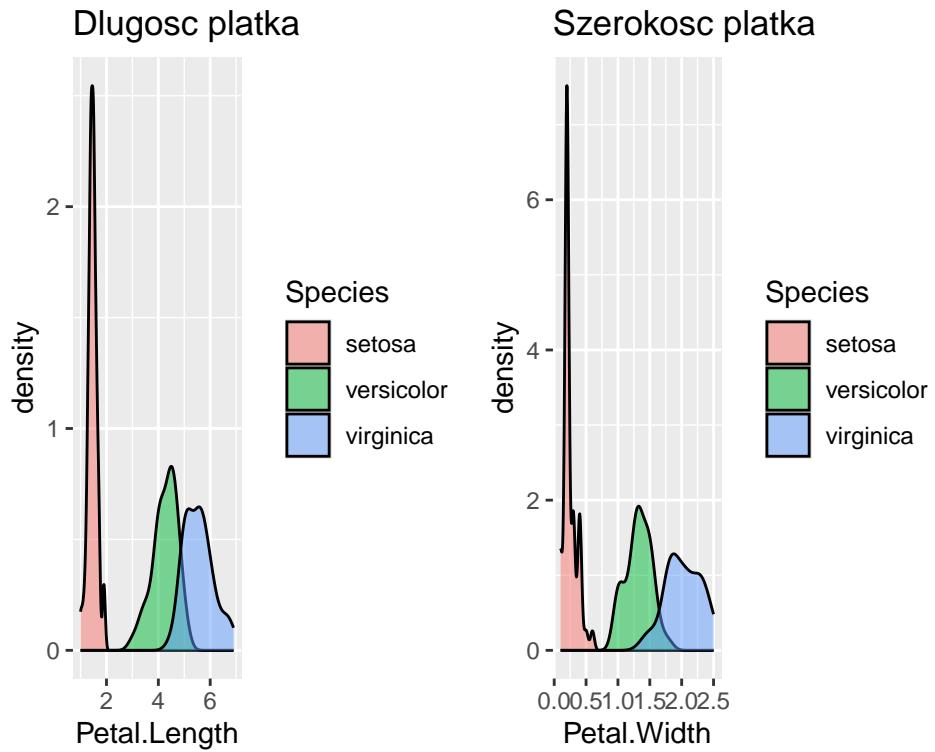


Najbardziej symetryczny wydaje się rozkład zmiennej `Sepal.Width`, oraz mniej `Sepal.Length`. Zmienne opisujące płatek wyraźnie wybijają dla małych wartości, a następnie reszta wartości

ma bardziej symetryczny rozkład. Może to sugerować np. mniejsze płatki kwiatów dla jednego z gatunków co może się nam przydać w dalszej analizie.



Na powyższych wykresach pudełkowych wyraźnie widać, że zmienne dotyczące dzwonka (Sepal), nie pozwala nam dokładnie rozgraniczyć naszych gatunków. Zdecydowanie najgorszą cechą pod tym względem jest **Sepal.Width**, której praktycznie wszystkie trzy pudełka się nakładają. Zmienne dotyczące wymiarów płatka zdecydowanie lepiej pozwolą nam zidentyfikować gatunki, natomiast przyjżyjmy się im lepiej, aby wybrać najlepszą cechę.



Na powyższych wykresach rozkładu widzimy, że gatunek setosa jest dobrze rozróżnialny, natomiast versicolor i virginica delikatnie się nakładają w podobnym stopniu dla długości i szerokości (jednak tu delikatnie mniej). Natomiast przez fakt, że zmienna Petal.Length będzie miała wyższe statystyki, a zatem większe różnice między grupami, wybieramy tą zmienną jako najlepszą do dyskryminacji naszych danych.

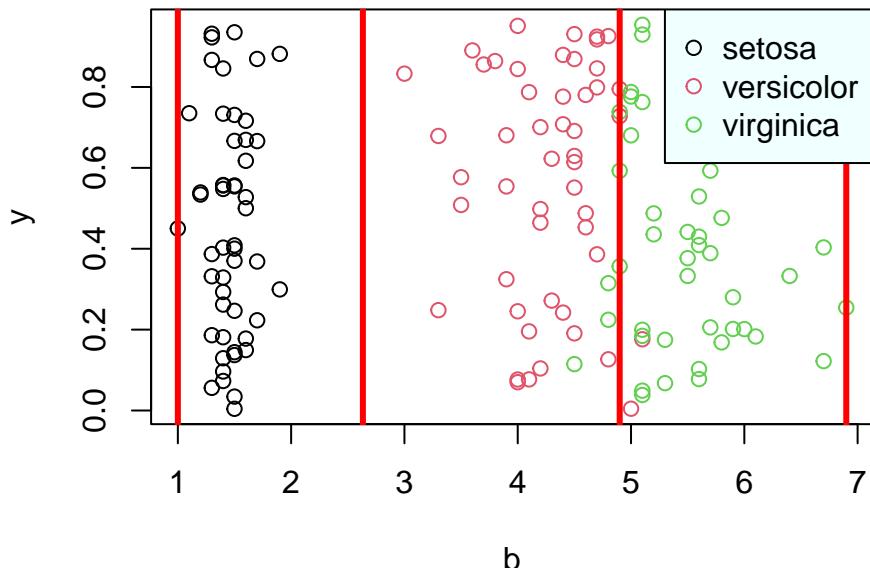
1.3 Porównanie nienadzorowanych metod dyskretyzacji

1.3.1 Metoda: equal frequency discretization

	setosa	versicolor	virginica
[1,2.63)	50	0	0
[2.63,4.9)	0	46	3
[4.9,6.9]	0	4	47

Powyższa tabela przedstawia nam wyniki dyskretyzacji opartej na równych częstościach

Metoda: equal frequency discretization



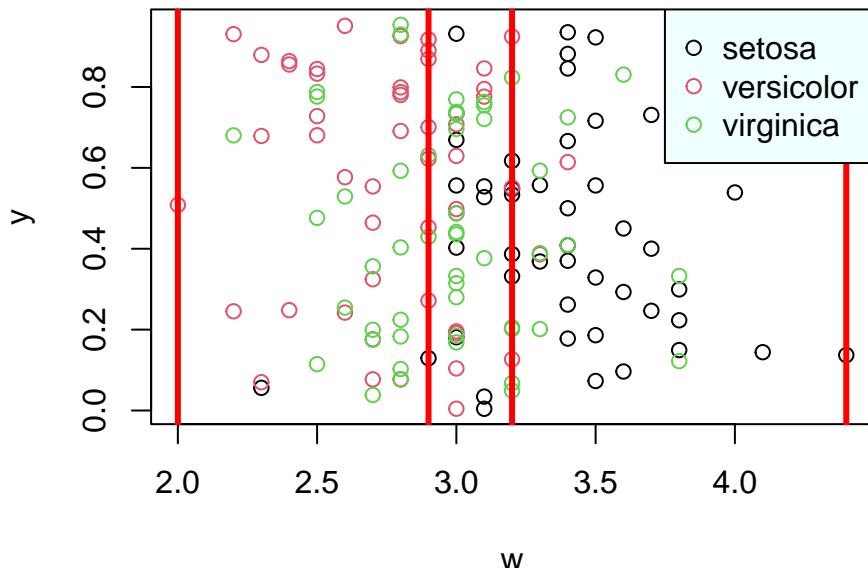
Cases in matched pairs: 95.33 %

x	
[1,2.63)	setosa
[2.63,4.9)	versicolor
[4.9,6.9]	virginica

Widzimy że metoda equal frequency discretization poradziła sobie dostyc dobrze, ze zgodnością ok. 95,3% dla zmiennej Petal.Length. Metoda ta dobrze działa szczególnie gdy dane są równomiernie rozłożone w całym zakresie i nie mają wyraźnych skupisk.

Zobaczmy teraz jak to wygląda dla zmiennej Sepal.Width

Metoda: equal frequency discretization



Cases in matched pairs: 55.33 %

x	
[2,2.9)	versicolor
[2.9,3.2)	versicolor
[3.2,4.4]	setosa

Widzimy, że dla tej zmiennej完全nie zawodzi dyskretyzacja. Dopasowanie jest na poziomie ok. 55,3%.

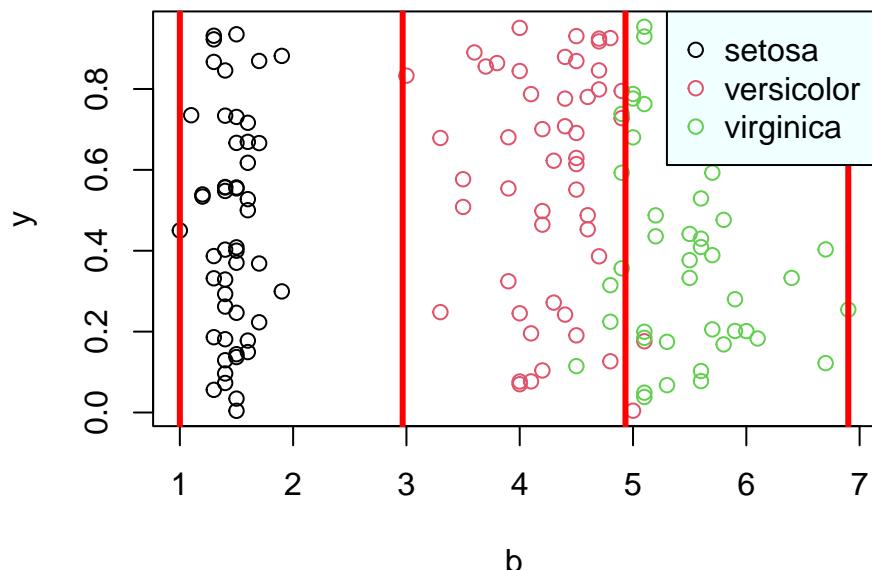
1.3.2 Metoda: equal interval width

Przeprowadźmy teraz analizę dla dyskretyzacji opartej na przedziałach o jednakowej szerokości (ang. equal interval width)

	setosa	versicolor	virginica
[1,2.97)	50	0	0
[2.97,4.93)	0	48	6
[4.93,6.9]	0	2	44

Powyższa tabela przedstawia nam wyniki dyskretyzacji opartej na przedziałach o jednakowej szerokości

Metoda: equal interval Width Discretization



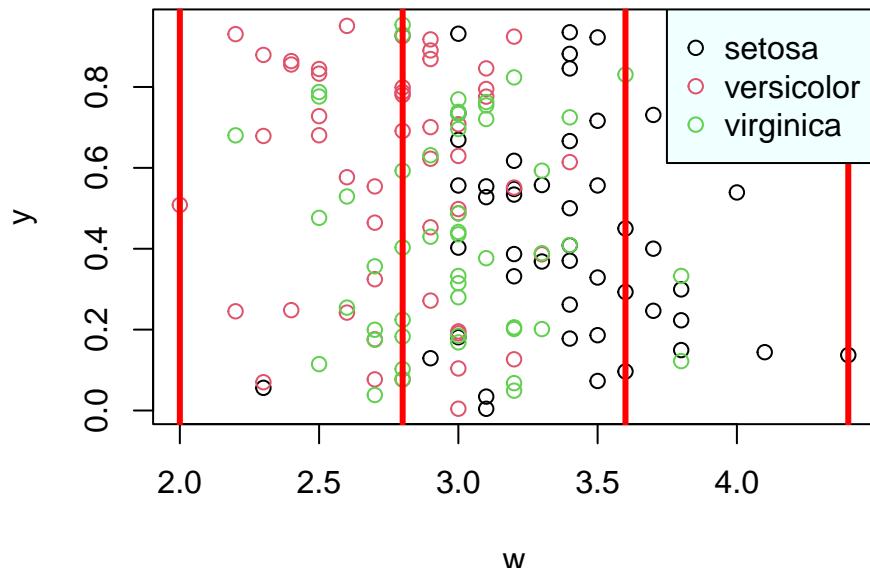
Cases in matched pairs: 94.67 %

x	
[1,2.97)	setosa
[2.97,4.93)	versicolor
[4.93,6.9]	virginica

Widzimy, że ta metoda jest również bardzo dobra, ma zgodność na poziomie ok. 94,7%. Metoda sprawdza się przy danych o nieregularnym rozkładzie – zapewnia równą liczbę obserwacji w przedziałach, co bywa przydatne przy klasyfikacji.

Sprawdźmy teraz jak wygląda ta metoda dla zmiennej `Sepal.Width`:

Metoda: equal interval Width Discretization



Cases in matched pairs: 50.67 %

x	
[2,2.8)	versicolor
[2.8,3.6)	setosa
[3.6,4.4]	setosa

Dla tej zmiennej również zgodność jest słaba, wynosi ok. 50,7%

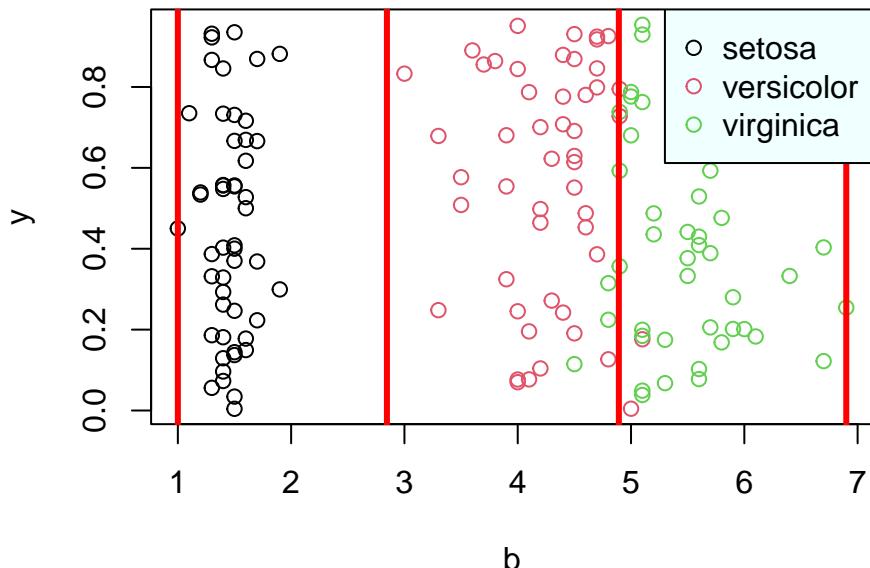
1.3.3 Metoda: k-means discretization

Sprawźmy teraz jak wygląda dyskretyzacja oparta na algorytmie grupowania (ang. k-means discretization)

	setosa	versicolor	virginica
[1,2.85)	50	0	0
[2.85,4.89)	0	46	3
[4.89,6.9]	0	4	47

Powyższa tabela przedstawia nam wyniki dyskretyzacji opartej na algorytmie grupowania.

Metoda: k-means discretization



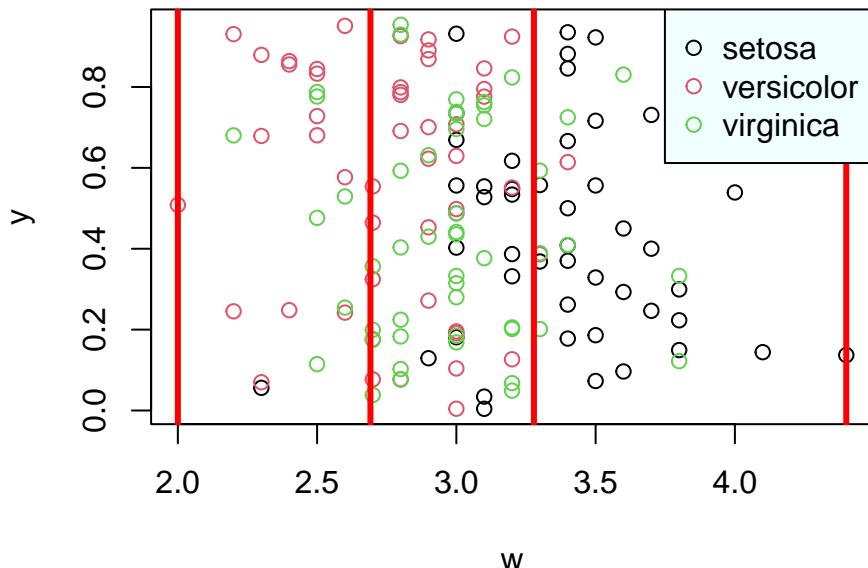
Cases in matched pairs: 95.33 %

x	
[1,2.85)	setosa
[2.85,4.89)	versicolor
[4.89,6.9]	virginica

Metoda klasteryzacji jest również bardzo skuteczna, uzyskała ok.95,3% skuteczności. Najlepsza jest, gdy dane mają naturalne skupiska (klastry) – metoda dopasowuje granice do faktycznej struktury danych.

Sprawdzamy teraz tę metodę dla zmiennej Sepal.Width:

Metoda: k-means discretization



Cases in matched pairs: 56 %

x	
[2,2.69)	versicolor
[2.69,3.28)	virginica
[3.28,4.4]	setosa

W tym przypadku metoda klasteryzacji poradziła sobie trochę lepiej, jednak w dalszym ciągu dopasowanie jest na poziomie 56%, co jest słabym wynikiem.

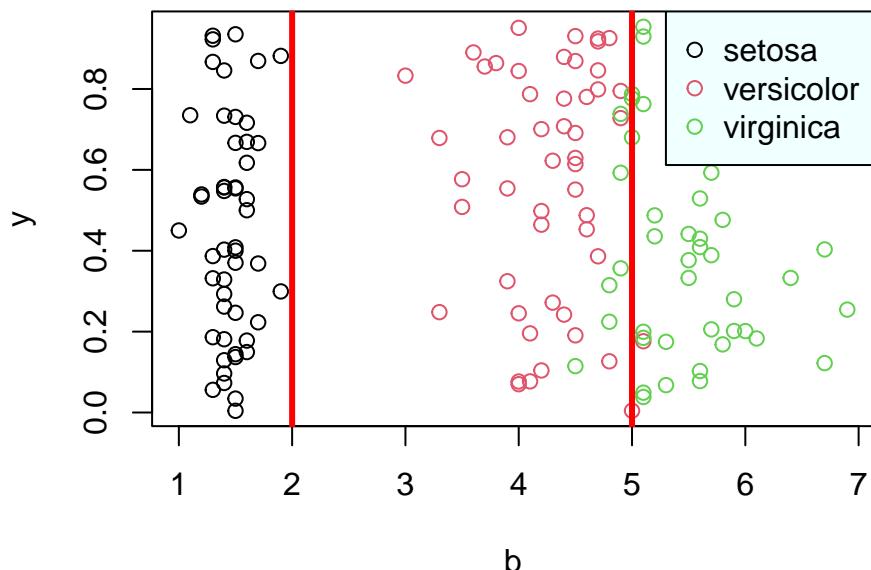
1.3.4 Metoda: fixed (user provided breaks)

Teraz przejdźmy do dyskretyzacji z przedziałami zadanymi przez użytkownika.

	setosa	versicolor	virginica
small	50	0	0
medium	0	48	6
large	0	2	44

Powyższa tabela przedstawia nam wyniki dyskretyzacji opartej na przedziałach zadanych przez użytkownika.

Metoda: fixed (user provided breaks)



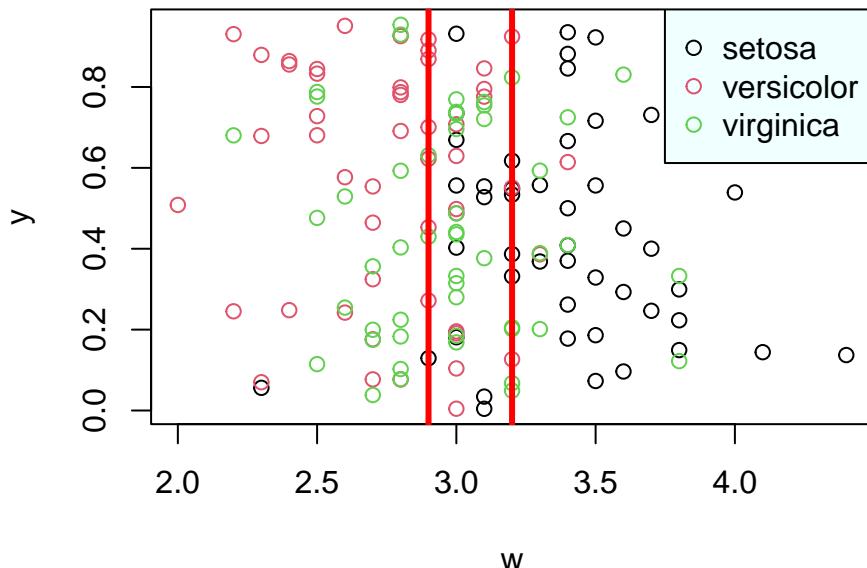
Cases in matched pairs: 94.67 %

x	
small	setosa
medium	versicolor
large	virginica

W tej metodzie uzyskujemy zgodność na poziomie ok.94,7%. Największym minusem tej metody jest konieczność wprowadzania granic przez użytkownika, co w większości przypadków jest trudne i męczące. Jednak w niektórych typach danych, gdzie są one oddalone od siebie i mają specyficzne wartości taka metoda również może się przydać

Sprawdźmy teraz co dostaniemy dla zmiennej Sepal.Width

Metoda: fixed (user provided breaks)



Cases in matched pairs: 55.33 %

x	
small	versicolor
medium	versicolor
large	setosa

W tej metodzie dla zmiennej opisującej szerokość działki dostajemy niezadowalający wynik.

1.4 Podsumowanie

Na podstawie przeprowadzonych analiz widać wyraźnie, że skuteczność dyskretyzacji mocno zależy od tego, jak dobrze dana cecha rozróżnia gatunki. Dla zmiennej Petal.Length, którą wybraliśmy jako najlepszą (bo najłatwiej było na jej podstawie rozróżnić gatunki), praktycznie każda metoda dała bardzo dobre wyniki – zgodność sięgała nawet 95%. Zarówno metody równej szerokości, równej liczności, k-means, jak i ta z ręcznymi programami, działały podobnie dobrze. To pokazuje, że jak mamy dobrą cechę, to nawet prosta metoda może nam dać dobre rezultaty.

Z kolei dla zmiennej Sepal.Width, która słabo odróżnia gatunki (pułecka się na siebie nakładały, nie było wyraźnych różnic), wszystkie metody wypadały raczej słabo – zgodność była na poziomie około 50–56%, niezależnie od wybranej metody. W skrócie: jeśli cecha jest zła, to żadna metoda dyskretyzacji jej nie „naprawi”.

Podsumowując – najlepsze wyniki osiągamy wtedy, gdy dobra cecha zostaje sparowana z odpowiednią metodą dyskretyzacji – choć w praktyce wybór metody ma mniejsze znaczenie niż wybór właściwej cechy.

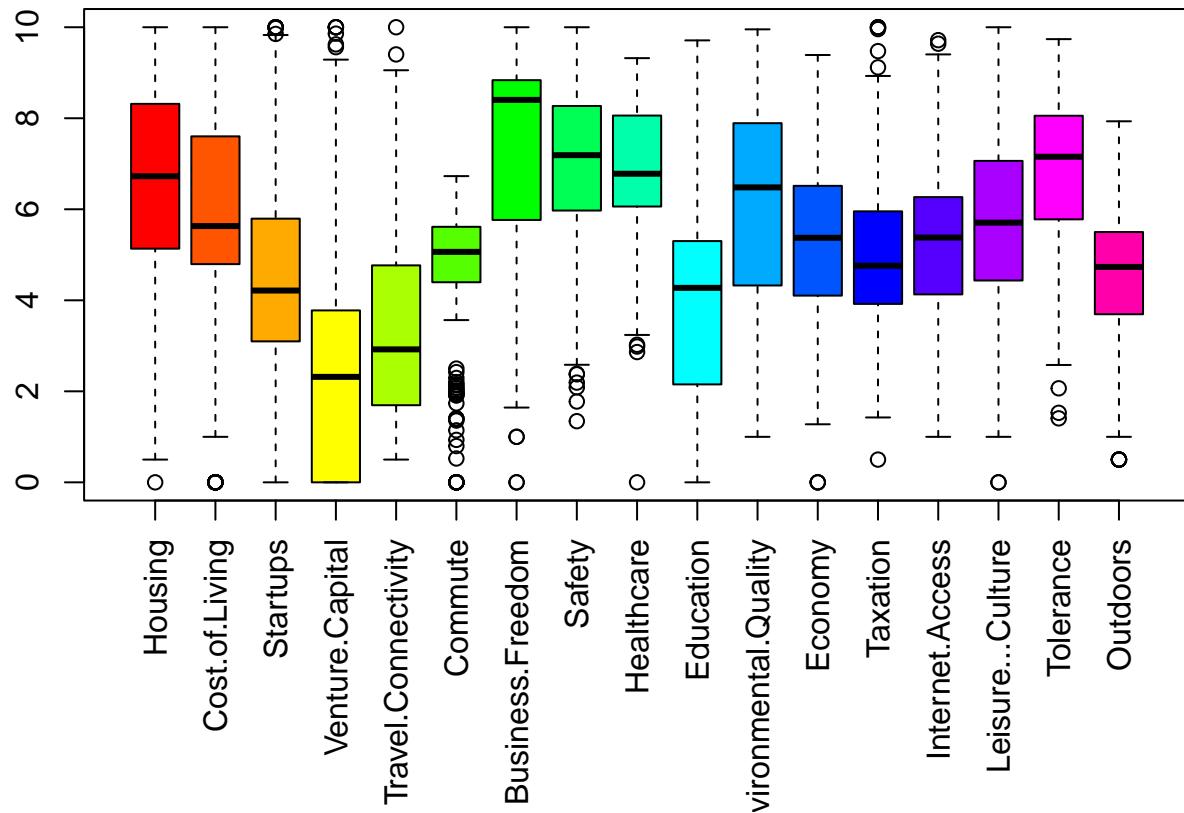
2 PCA - analiza składowych głównych

2.1 Krótki opis zagadnienia

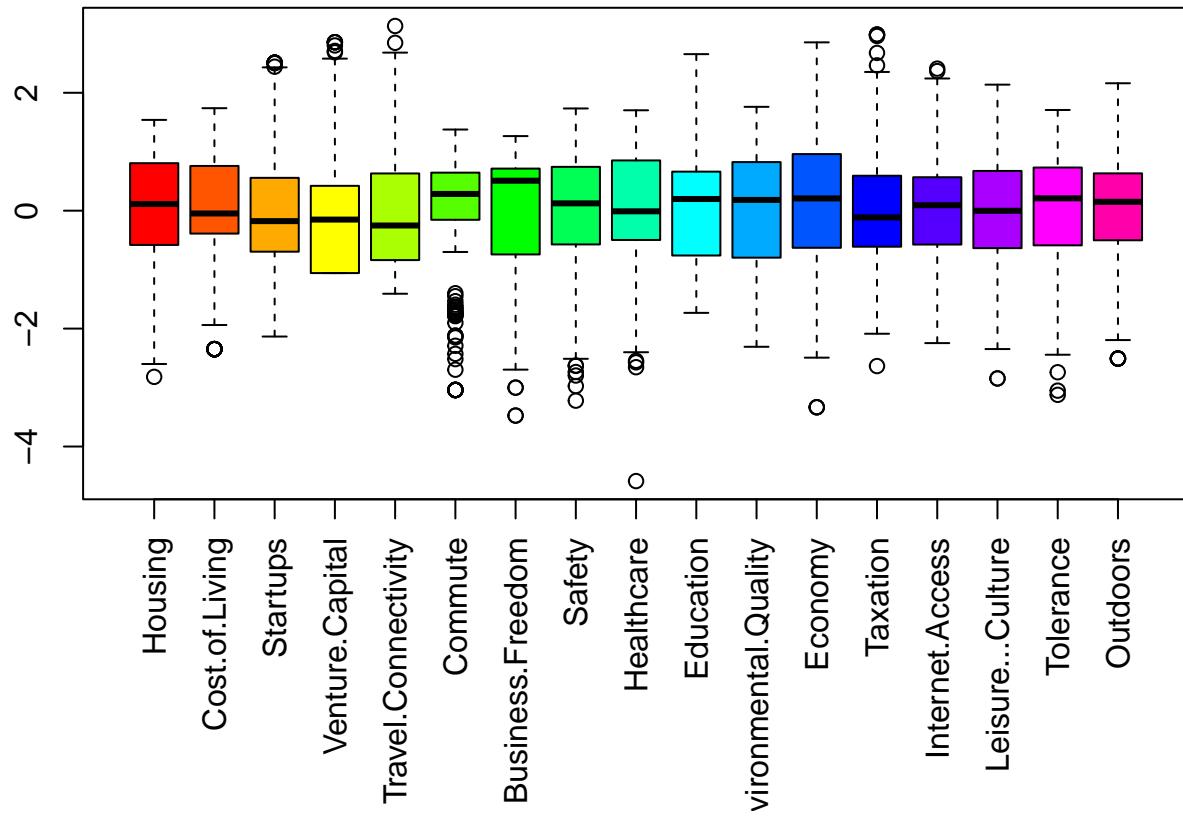
Analizowany zbiór danych zawiera wskaźniki jakości życia dla wybranych miast na całym świecie. Dane pochodzą z ze strony Kaggle (źródło: <https://www.kaggle.com/datasets/orhankaramancode/city-quality-of-life-dataset>) i obejmują różne kategorie, takie jak: bezpieczeństwo, opieka zdrowotna, jakość powietrza, 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.



Jak widać na powyższym wykresie kolejne cechy nie wykazują bardzo zróżnicowanych statystyk, jednak warto je zestandardyzować tak aby nasze PCA było lepszej jakości.



2.3 Wyznaczanie składowych głównych

Tabela 14: Wektory ładunków (PC1, PC2 i 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

	PC1	PC2	PC3

Powyzsza tabela przedstawia wektory ładunków dla PC1, PC2 oraz PC3 Z powyzszych danych moześmy zauważyć następujące wnioski

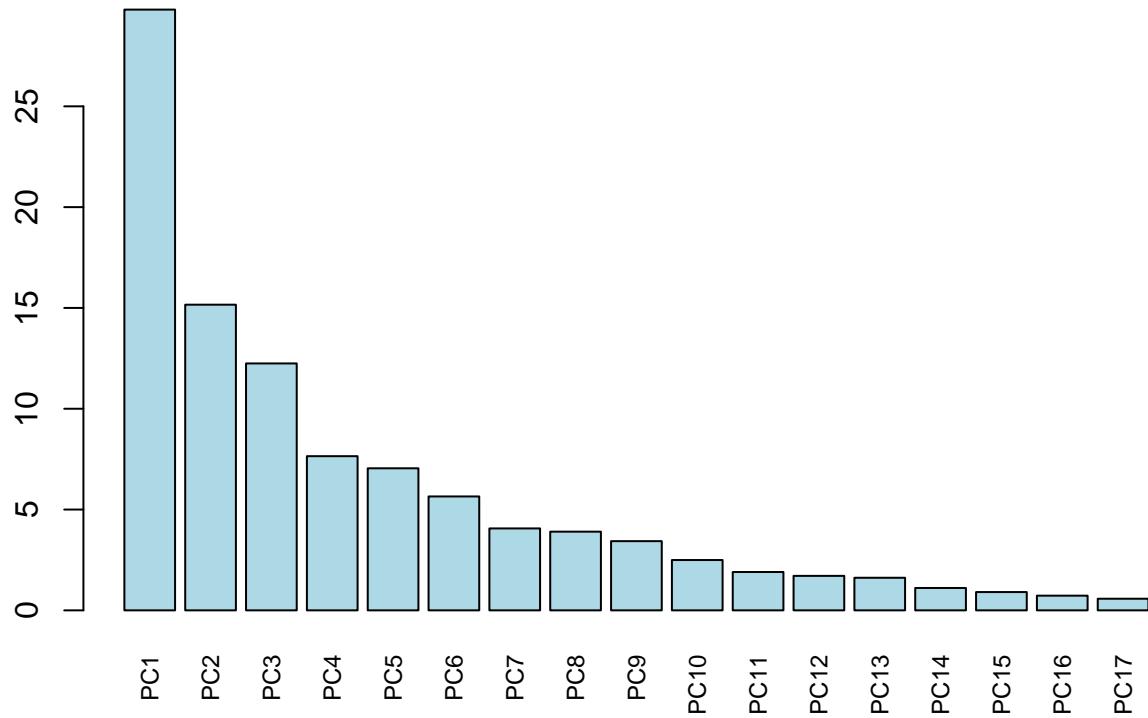
Pierwsza składowa wskazuje wysokie wartości: Housing (0.308), Cost of Living (0.260) – wysokie koszty zamieszkania i utrzymania. Ujemne wartości: Business Freedom (-0.377), Education (-0.403), Healthcare (-0.280) – oznaczają lepszą infrastrukturę biznesową, edukacyjną i zdrowotną. Zatem PC1 oddziela miasta bogate, ale drogie od tańszych, ale z gorszą infrastrukturą.

Druga składowa wskazuje ujemne wartości: Startups (-0.483), Venture Capital (-0.427), Leisure & Culture (-0.365) – wskazują na rozwiniętą scenę startupową i życie kulturalne. Dodatnie wartości: Tolerance (0.355), Safety (0.287) – odzwierciedlają otwartość społeczną i bezpieczeństwo. PC2 oddziela miasta pod kątem innowacyjności vs. tolerancji.

Trzecia składowa posiada ujemne wartości: Commute (-0.506), Travel Connectivity (-0.340) – wskazują na dobre połączenia transportowe i krótki czas dojazdów. Dodatnie wartości: Economy (0.309) – oznaczają silną gospodarkę. PC3 pokazuje zależność między gospodarką a mobilnością

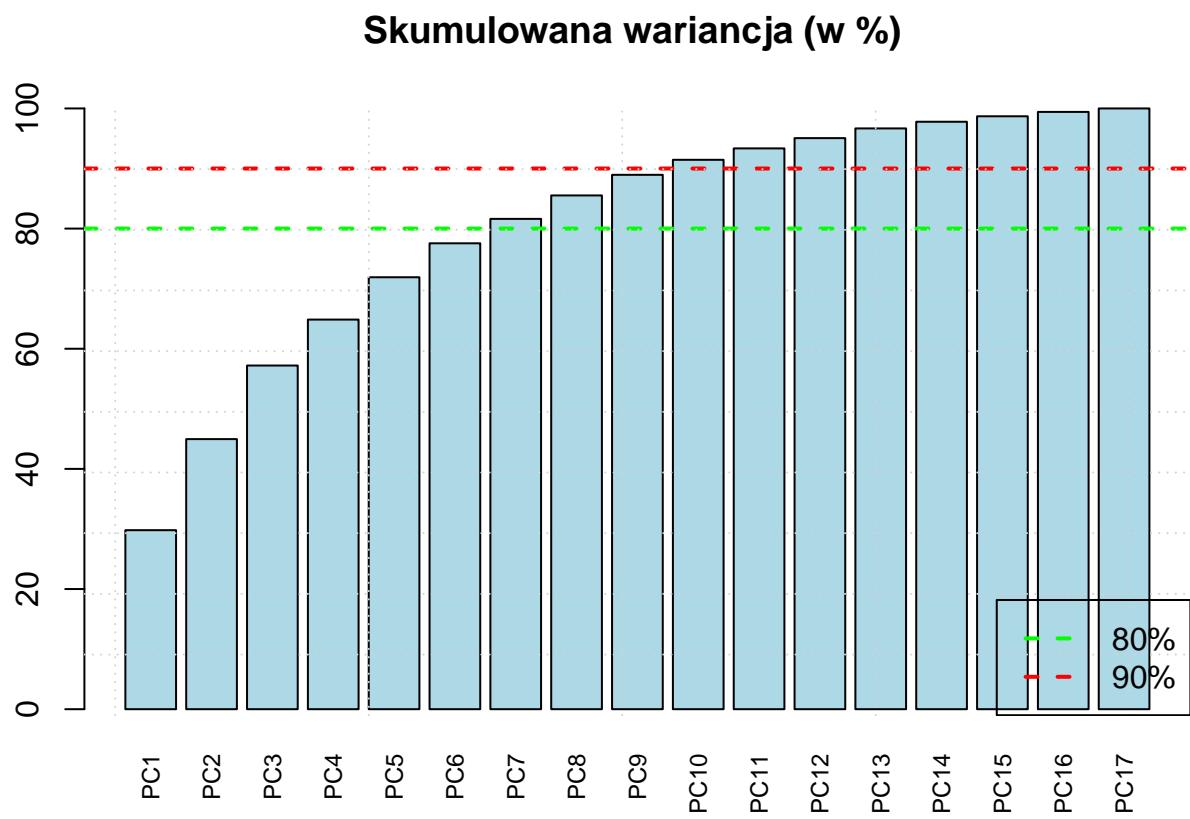
2.4 Zmienna odpowiadająca poszczególnym składowym

Wariancja odpowiadająca poszczególnym składowym (w %)



Rysunek 1: Wykres wariancji

Na powyższym wykresie widzimy wariancję dla poszczególnych zmiennych składowych, przedstawioną w procentach, pokazuje jaki procent zmienności odpowiada poszczególnym składowym głównym. Zgodnie z intuicją największe znaczenie mają pierwsze składowe, a kolejne coraz mniej.

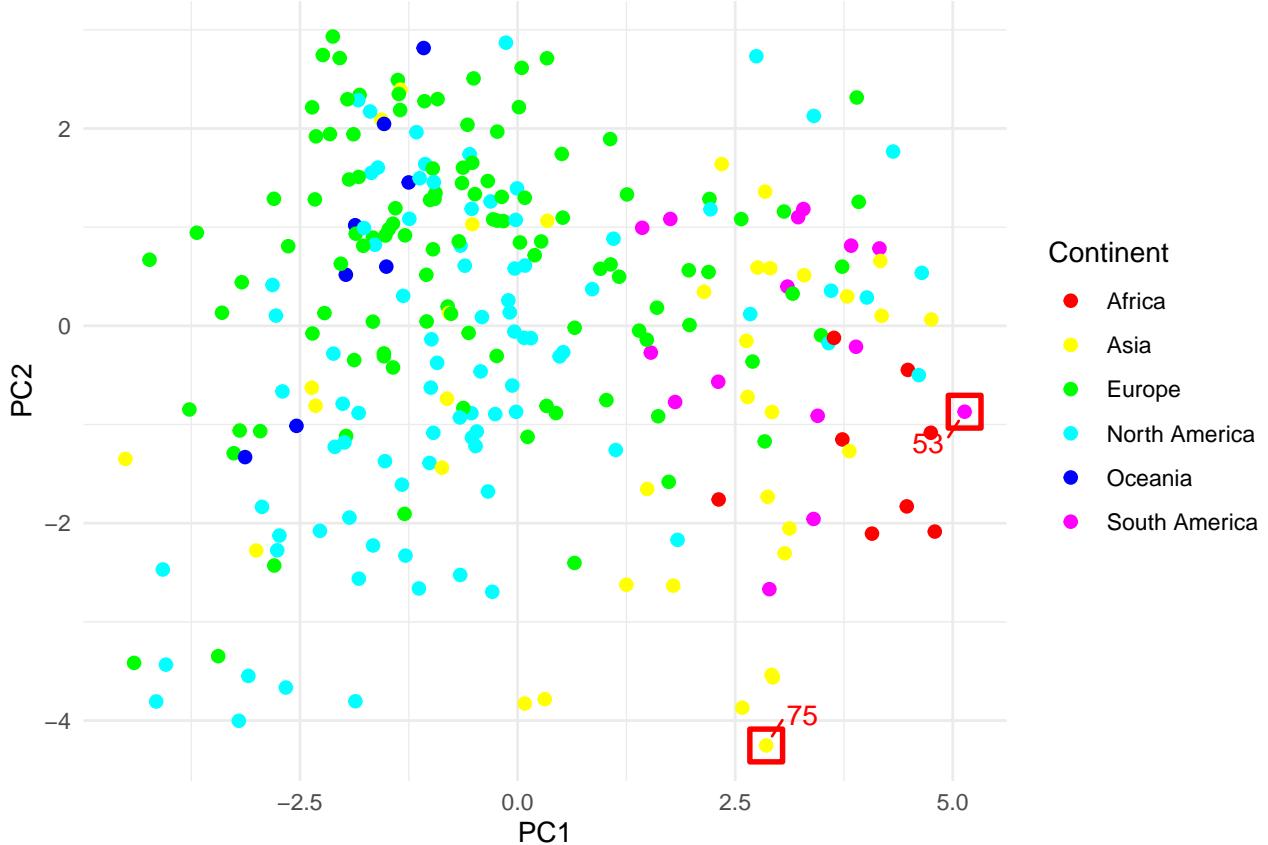


Rysunek 2: Wariancja skumulowana

Powyższy wykres przedstawia skumulowaną wariancję. Możemy zauważyć, że do wyjaśnienia 80% całkowitej zmienności potrzebujemy 7 pierwszych składowych. Natomiast do wyjaśnienia 90% potrzebne jest 10 pierwszych składowych (tj. PC1-PC10). Im dalsza składowa, tym mniej nam daje.

2.5 Wizualizacja danych wielowymiarowych

Dane – wykres rozrzutu 2D



Na podstawie Rysunku 5. 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.

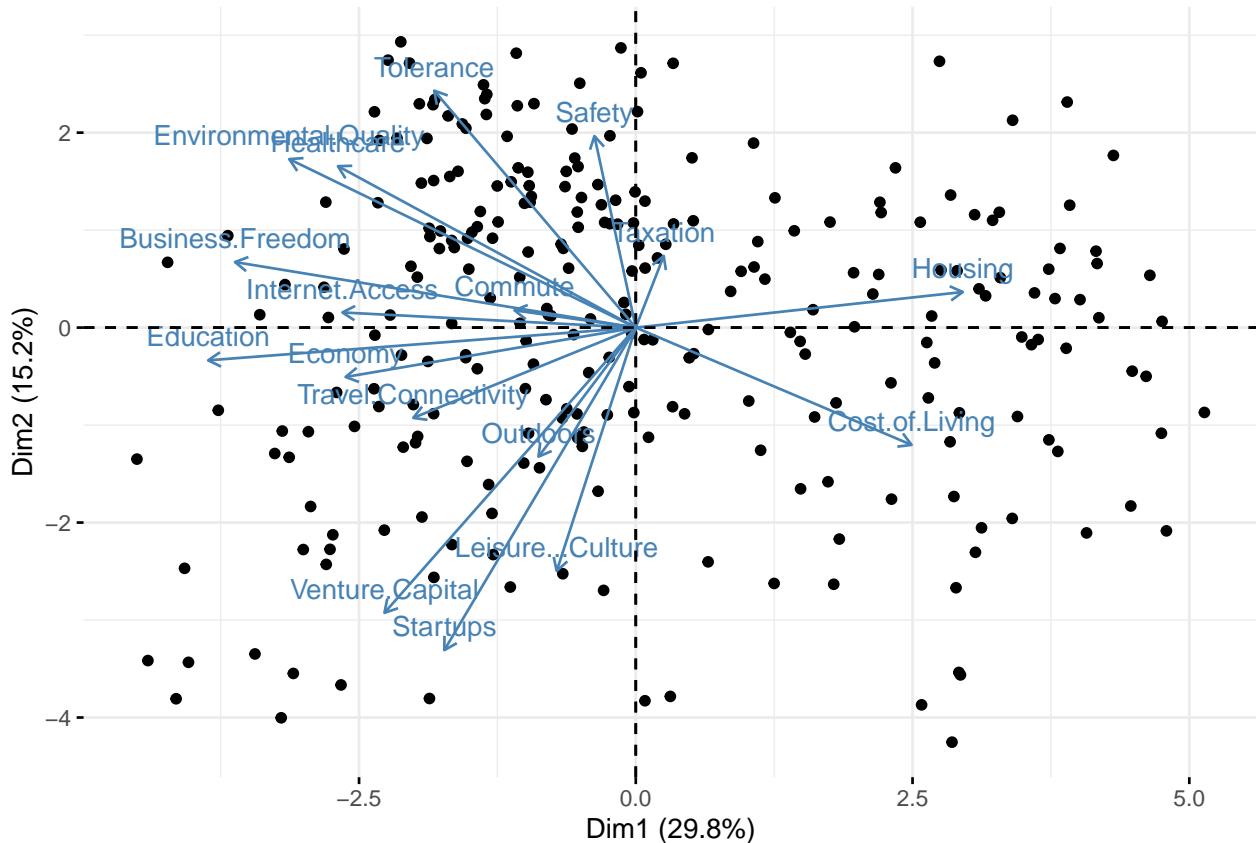
Na Rysunku można zauważać, że miasta z niektórych kontynentów (np. Oceania czy Ameryka Północna) grupują się w bardziej zwarte klastry. Może to sugerować, że kraje z tych regionów mają dość zbliżone warunki pod kątem analizowanych zmiennych – np. rozwoju technologicznego, infrastruktury czy innowacyjności. Z drugiej strony, dane z Afryki czy Azji są bardziej rozproszone, co może wskazywać na większe zróżnicowanie wewnętrzne – od krajów bardzo rozwiniętych po te, które są jeszcze w fazie rozwoju.

Da się też zauważać pewne naturalne skupiska miast. Przykładowo – miasta z Europy, Oceanii i Ameryki Północnej znajdują się w podobnych rejonach wykresu, co może potwierdzać ich podobny poziom rozwoju czy strategii inwestycyjnych. Z kolei miasta z Azji, Ameryki Południowej i Afryki częściej występują na obrzeżach tego głównego skupiska, co może wskazywać na pewne różnice względem bardziej rozwiniętych regionów.

Spośród wszystkich punktów najbardziej wyróżniają się dwa: punkt o najwyższej wartości PC1 (75) - Singapur oraz ten z najniższym PC2 (53) - Delhi. Singapur to prawdopodobnie jedno z najbardziej rozwiniętych miast, charakteryzujące się bardzo wysokim poziomem zaawansowania technologicznego, otwartości na innowacje i dużym udziałem kapitału inwestycyjnego. Delhi może reprezentować miasto, które mocno wybija się na tle swojego kontynentu – może być to miasto o dynamicznym rozwoju startupów czy silnym ekosystemie innowacji.

2.6 Korelacja zmiennych

PCA – Biplot

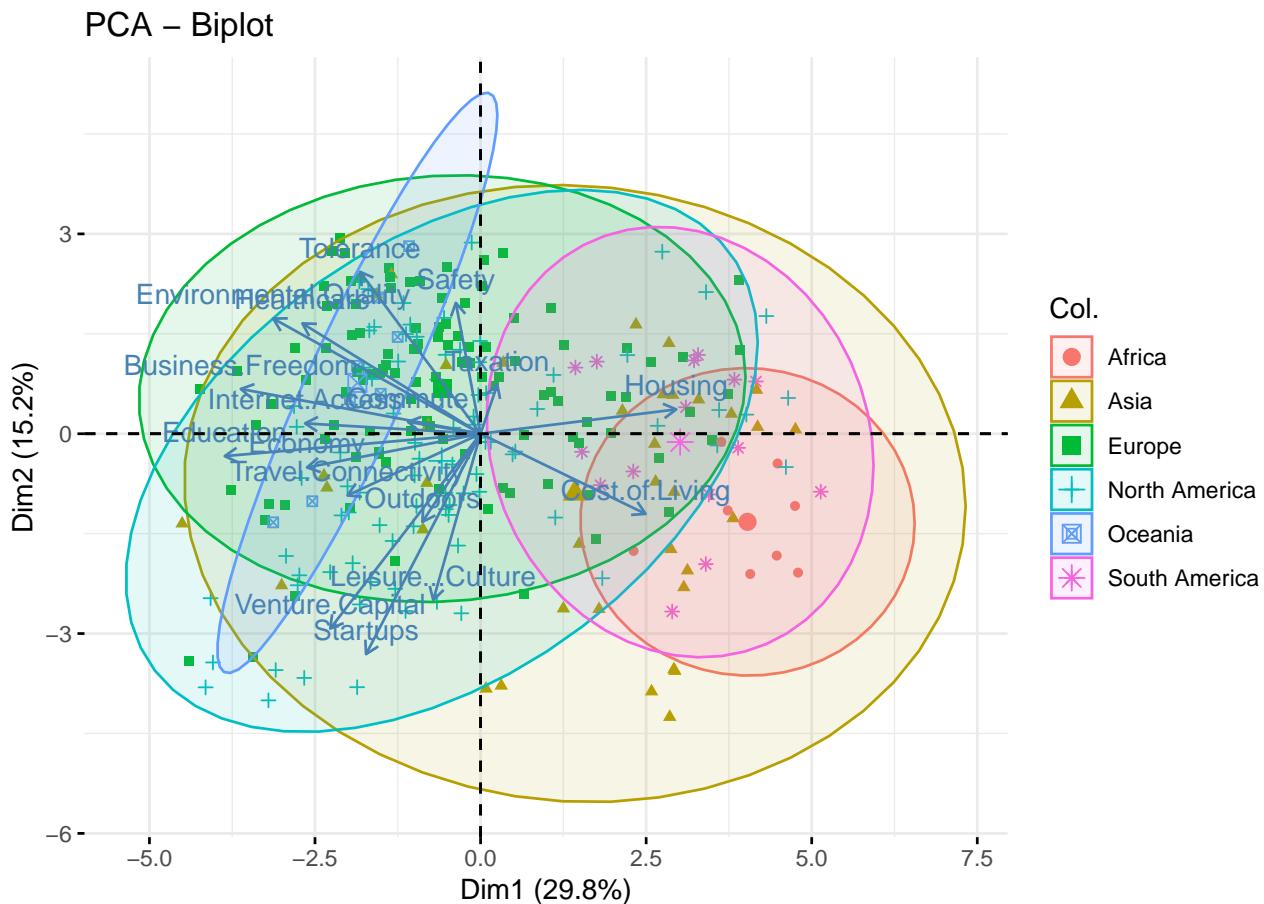


Na dwuwymiarowym biplocie można zauważać, jak poszczególne zmienne wpływają na dwa pierwsze główne komponenty oraz jak są względem siebie skorelowane. Im dłuższa strzałka, tym dana zmienna ma większy wpływ na daną cechę.

Zdecydowanie w oczy rzucają się zmienne takie jak Startups, Culture, Leisure i Venture Capital – wszystkie one są skierowane w podobną stronę, co oznacza, że są ze sobą dodatnio skorelowane. Można to interpretować tak, że miasta z silnym środowiskiem startupowym zwykle oferują też więcej możliwości spędzania wolnego czasu i mają łatwiejszy dostęp do kapitału inwestycyjnego.

Z kolei Cost of Living i Housing są zmiennymi mocno „oddzielonymi” od reszty – ich wektory skierowane są prawie przeciwnie do wcześniej wspomnianych zmiennych rozwoju. To sugeruje ujemną korelację – czyli tam, gdzie koszty życia i mieszkani są wysokie, niekoniecznie mamy sprzyjające warunki do prowadzenia start up-u czy szeroką ofertę kulturalną.

Są też zmienne, które nie wykazują silnej korelacji z żadną inną – ich wektory są krótkie i/lub ustawione pod kątem bliskim prostemu względem pozostałych.

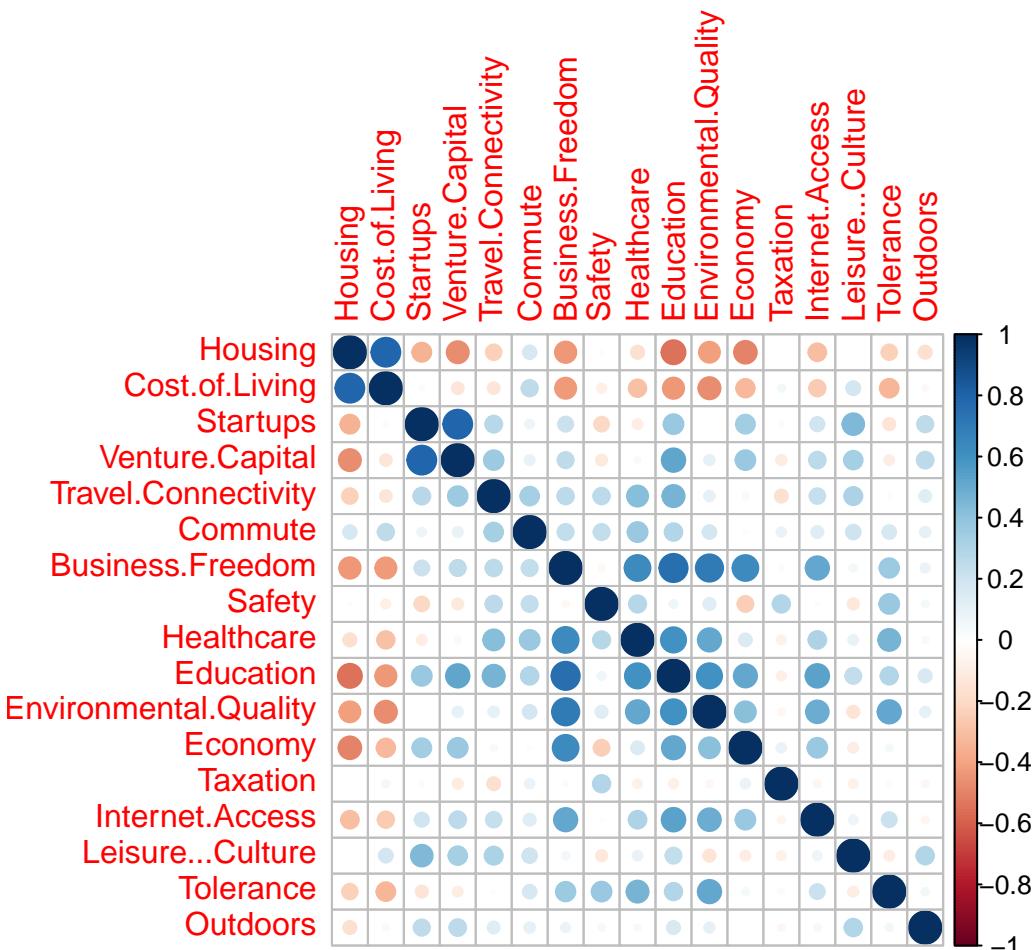


Na powyższym rysunku mamy ponownie dwuwykres PCA, ale tym razem pokazano też podział punktów według kontynentów. Widać wyraźnie, że miasta z Oceanii, Europy i Ameryki Północnej skupiają się w podobnym rejonie wykresu – to może sugerować, że są one na zbliżonym poziomie rozwoju gospodarczego, technologicznego i jakości życia. Świadczy

o tym obecność tych punktów w obszarze, gdzie kierują się wektory takich zmiennych jak Startups, Venture Capital czy Internet Access.

Z kolei miasta z Afryki i Ameryki Południowej przeważnie znajdują się po lewej stronie wykresu. Może to oznaczać niższe koszty życia i mieszkani (bo w tym kierunku zmierzają wektory Cost of Living i Housing), ale jednocześnie słabsze wyniki w obszarach takich jak edukacja, gospodarka, kultura czy ochrona zdrowia.

Co ciekawe, Azja jest mocno rozproszona po całym wykresie. To pokazuje, jak bardzo różnicowane są azjatyckie miasta – znajdziemy tu zarówno bardzo rozwinięte metropole (np. Singapur czy Tokio), jak i miasta z niższym poziomem rozwoju infrastruktury i usług.



Na podstawie powyższego wykresu można wyciągnąć wnioski analogiczne do tych z dwuwykresu. Zaobserwowane wcześniej zależności znajdują potwierdzenie w danych z macierzy korelacji. Możemy zauważyć wysoką korelację między zmiennymi Startups, Culture, Leisure i Venture Capital. Również widzimy ujemną korelację między zmiennymi Startups i housing co znowu potwierdza nasze wcześniejsze wnioski.

Jednak widzimy tutaj także koreacje które wynikały z poprzedniego rysunku, jednak był on dosyć nieczytelny. Tutaj dodatkowo możemy dostrzec wysoką korelację między Education i Business.Freedom, oraz niską np. między Education a Housing.

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

2.7 Wnioski końcowe

Podczas analizy PCA udało się wyciągnąć sporo ciekawych wniosków dotyczących miast z różnych części świata. Po przekształceniu danych do postaci głównych składowych, zauważymy, że PC1 jest silnie związana z poziomem rozwoju gospodarczego i kosztami życia – miasta z wysokim PC1 to zazwyczaj te bogatsze, ale też droższe do życia. PC2 pokazała ciekawy kontrast między nowoczesnością i innowacyjnością (np. obecnością startupów czy dostępnością kapitału wysokiego ryzyka) a takimi cechami jak bezpieczeństwo i tolerancja. Z kolei PC3 dobrze obrazuje relację między siłą gospodarki a komfortem życia, czyli dostępnością transportu i czasem dojazdów.

Z czysto technicznego punktu widzenia, do uzyskania naprawdę dobrej reprezentacji danych wystarczy już 7 składowych, które tłumaczą 80% zmienności. A jeśli chcemy podejść do tematu bardzo dokładnie – 10 składowych daje aż 90%, co pokazuje, że dalsze komponenty nie wnoszą już nic szczególnie przełomowego. Można więc powiedzieć, że PCA faktycznie pomogło w sensownym uproszczeniu danych, bez straty najważniejszych informacji.

Jeśli chodzi o wykresy rozrzutu, to bardzo dobrze było widać pewne klastry. Europa, Oceania i Ameryka Północna trzymają się razem, co wskazuje na podobny poziom rozwoju gospodarczego i infrastrukturalnego. Z kolei Azja, Afryka i Ameryka Południowa były bardziej rozproszone, co pokazuje, że tam różnice między miastami są znacznie większe. Ciekawym przypadkiem był Singapur, który mocno wyróżniał się na osi PC1 – to potwierdza jego bardzo wysoki poziom rozwoju. Po drugiej stronie był Delhi, które miało najniższy wynik na PC2, co może sugerować słabsze wskaźniki nowoczesności i innowacyjności.

Analiza korelacji też rzuciła trochę światła na zależności między zmiennymi. Na przykład środowisko startupowe było mocno powiązane z życiem kulturalnym, z kolei wysokie koszty życia i mieszkań były negatywnie skorelowane z dostępnością edukacji i jakością infrastruktury – czyli często tam, gdzie drogo, niekoniecznie jest wygodnie.

Na koniec warto podkreślić znaczenie standaryzacji – bez niej wyniki PCA mogłyby być zupełnie inne. Jedna zmienna mogłaby zdominować cały obraz, co mocno zniekształciłoby interpretację. Dzięki standaryzacji dane były bardziej porównywalne, wykresy przejrzyste, a wnioski – wiarygodne.

Podsumowując, PCA pozwoliło wyciągnąć spójne i logiczne wnioski dotyczące głównych czynników różnicujących miasta. Zarówno analiza składowych, jak i korelacji dała jasny obraz tego, co wpływa na rozwój, jakość życia czy innowacyjność w różnych częściach świata.

3 MSD - skalowanie wielowymiarowe

3.1 Wczytanie i przygotowanie da1nych

W tej części sprawozdania będziemy korzystać ze zbiotu danych *Titanic*. Zawiera on wybrane charakterystyki opisujące pasażerów Titanica (w tym m.in. takie zmienne jak: wiek, płeć, miejsce rozpoczęcia podróży czy klasa pasażerska) wraz z informacją czy dana osoba przeżyła katastrofę (zmienna *Survived*).

Tabela 15: Struktura danych Titanic

Typ zmiennej	
PassengerId	integer
Survived	integer
Pclass	integer
Name	character
Sex	character
Age	numeric
SibSp	integer
Parch	integer
Ticket	character
Fare	numeric
Cabin	character
Embarked	character

Tabela 16: Brakujące dane

	x
Age	177

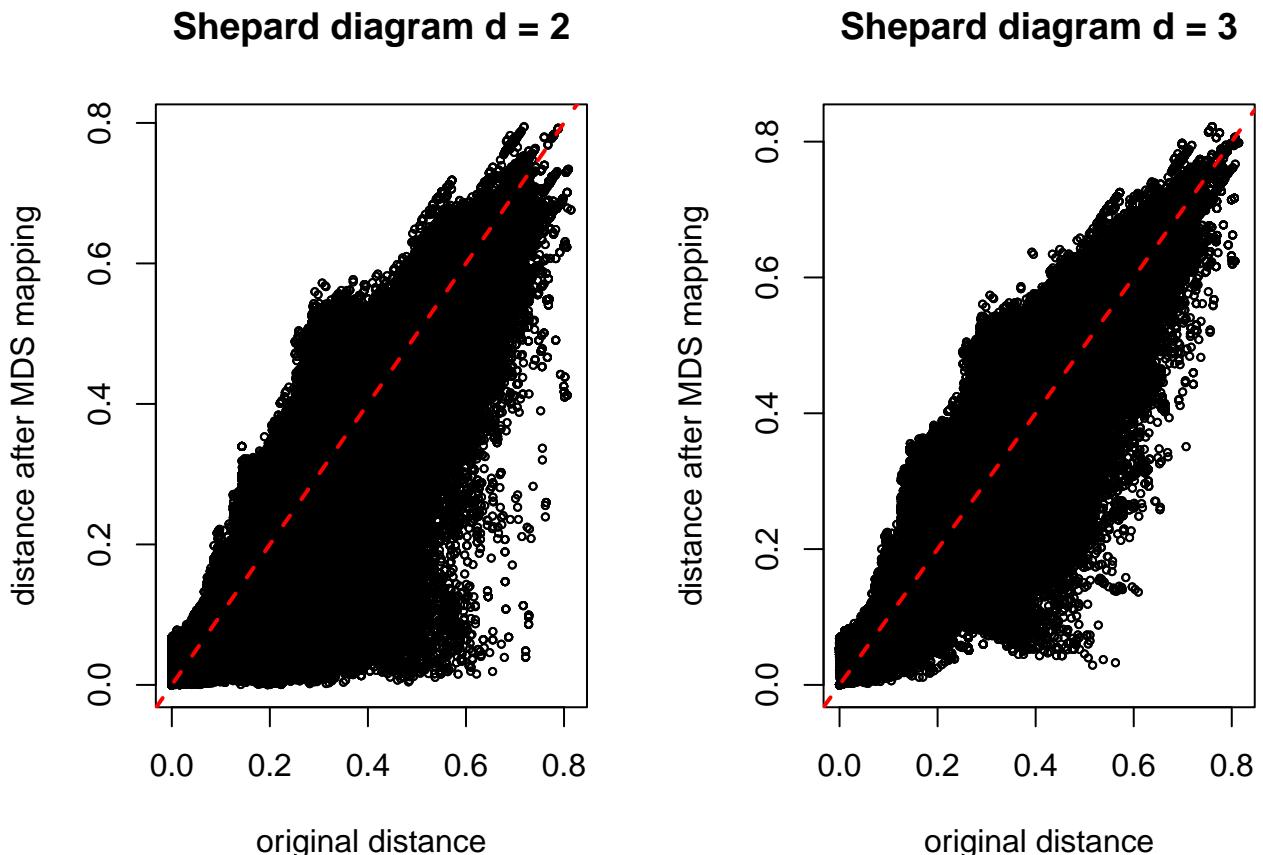
Nasze dane mają 891 przypadków i 12 cech. W powyższej tabeli możemy zobaczyć wszystkie cechy oraz ich typy. Nasze dane mają 7 cech numerycznych oraz 5 tekstowych. Widzimy, że nie wszystkie zmienne zostały poprawnie rozpoznane. Zmienne *Survived*, *Pclass*, *Sex*, *Embarked* powinny mieć typ factor. Zmienne *PassengerId*, *Name*, *Ticket* i *Cabin* pełnią role identyfikatorów, zatem powinny zostać usunięte. Widzimy także, że mamy braki danych występujące w kolumnie *Age*.

Po czyszczeniu danych zostają nam 4 zmienne ilościowe oraz 4 jakościowe.

3.2 Redukcja wymiaru na bazie MDS

W celu redukcji wymiaru zostanie wykonane skalowanie wielowymiarowe (MDS) do 2 i 3 wymiarów.

Wartość STRESS dla 2 wymiarów: 6069.187 Wartość STRESS dla 3 wymiarów: 3523.95



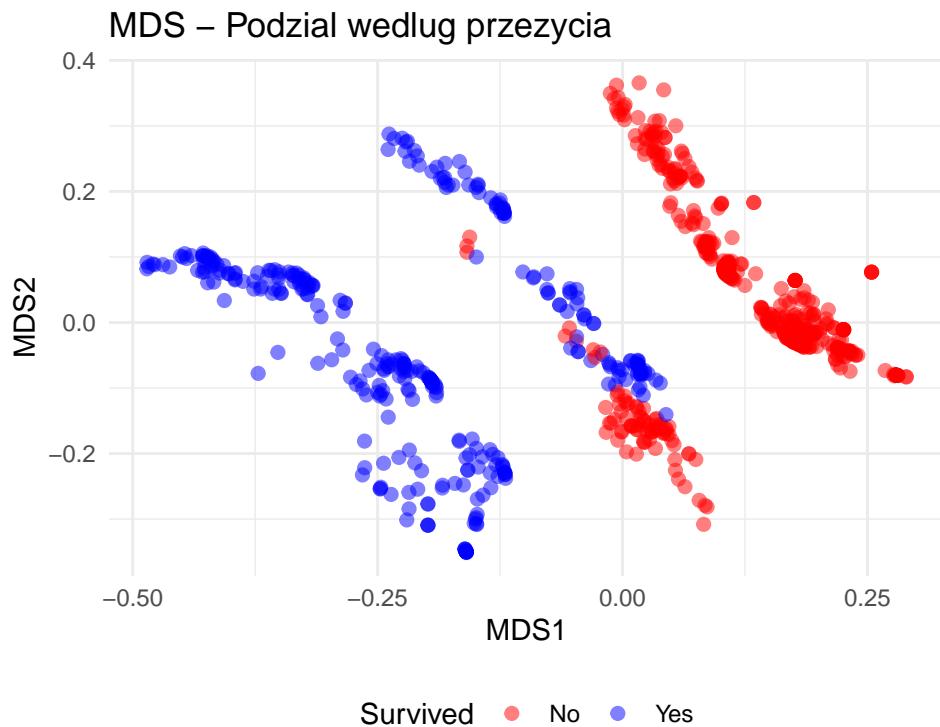
Dla danych bez zmiennej Survived, wyznaczono macierz odmiенноśc, uwzględniając odpowiednio typy zmiennych oraz standaryzację. Następnie wykonano zarówno klasyczne (metryczne) MDS (`cmdscale()`), redukując dane do 2 i 3 wymiarów.

Na powyższych wykresach możemy zobaczyć diagramy Sheparda dla dwóch i trzech wymiarów. Analiza wykazała, że punkty dobrze układają się wzdłuż linii prostej, co wskazuje na zachowanie struktury danych. Wartości STRESS dla dwóch wymiarów są wyższe niż dla trzech, gdzie osiągneliśmy wartość ok. 0.16 co wskazuje na dosyć dobre odwzorowanie. W dalszej analizie będziemy brali pod uwagę właśnie te skalowanie.

Wykonane zostało również skalowanie niemetryczne, jednak ze względu na dobrą strukturę danych nie poprawiło ono znacząco rezultatów, zatem zostaniemy przy metrycznym MDS.

3.3 Wizualizacja danych

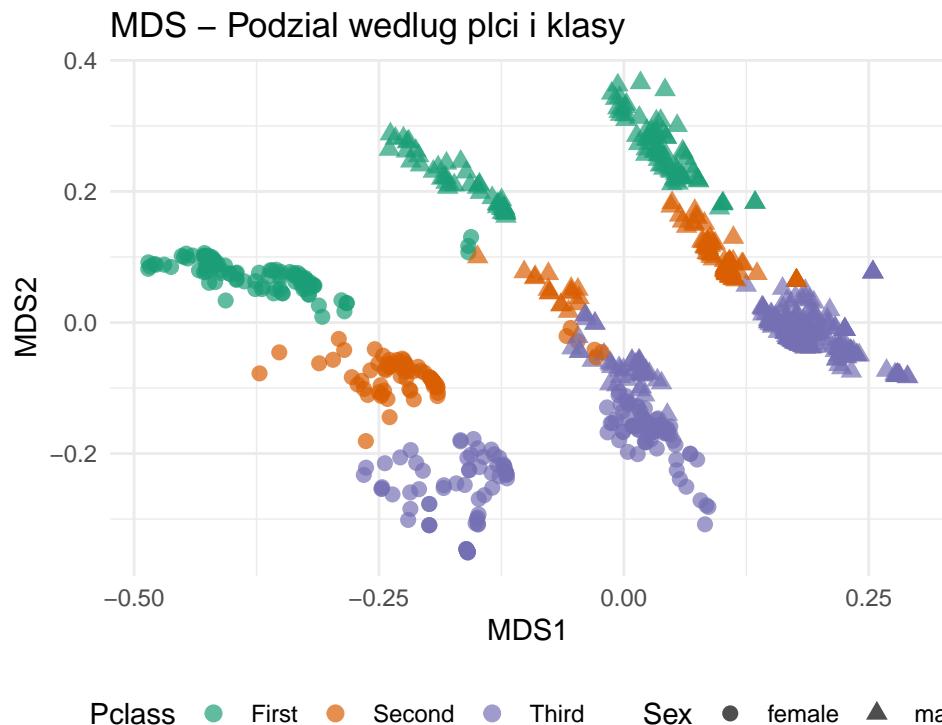
MDS dla 3 wymiarów dał lepsze rezultaty jeśli chodzi o odległości między punktami, jednak ze względu na wygodę i większą czytelność, zwizualizujemy dane na wykresach 2D, w których będziemy porównywać MDS1 z MDS2.



Analiza MDS wykazała wyraźny podział między grupami: osoby, które przeżyły (niebieskie) skupiają się w konkretnych regionach przestrzeni, podczas gdy te, które zginęły (czerwone), zajmują inny obszar, oczywiście nie jest to idealnie rozdzielone dlatego nie jesteśmy w stanie w pełni powiedzieć jakie cechy determinowały przeżycie, jednak wskazuje to na związek cech osób z przeżyciem i potwierdza, że dane zawierają informacje umożliwiające rozróżnienie tych grup.

Jest to zgodne z informacjami dot. przeżycia katastrofy ponieważ wiemy że większość ocalałych to kobiety i dzieci, które otrzymały pierwszeństwo podczas ewakuacji, a to są cechy które braliśmy pod uwagę w naszym MDS

Dodatkowo możemy zaobserwować nietypowe, odstające obserwacje. Obecność punktów oddalonych od głównych skupisk sugeruje, że niektóre przypadki mimo podobieństwa cech mogły przeżyć albo nie. Przykładem może tutaj być, że niektórzy mężczyźni się szybciej ewakuowali, zatem przeżyli.



Analiza MDS wykazała wyraźny podział między grupami: Pasażerowie podróżujący w różnych klasach (1., 2. i 3.) tworzą wyraźnie odseparowane skupiska, co wskazuje na znaczące różnice w ich cechach (np. wiek, lokalizacja na statku). Klasa pierwsza (oznaczona kolorem zielonym) skupia się głównie w górnej części wykresu, klasa druga (pomarańczowa) w części środkowej, natomiast klasa trzecia (fioletowa) w dolnej. Taki rozkład może odzwierciedlać nierówności między pasażerami, które prawdopodobnie przekładały się na ich szanse przeżycia – np. poprzez różny dostęp do szalup ratunkowych.

Analiza MDS pod względem płci ujawnia częściową separację płci w obrębie każdej klasy. Zarówno kobiety jak i mężczyźni w każdej klasie są skupieni w pewnych obszarach, co oznacza że mieli wiele cech wspólnych, jednak tylko w obrębie klas.

Przy porównaniu z wykresem z podziałem na przeżycie, możemy wysunąć ciekawe wnioski. Widzimy przede wszystkim, że zdecydowanie więcej kobiet przeżyło. Wynika to najprawdopodobniej z faktu, że miały one wraz z dziećmi pierwszeństwo w ewakuacji. Jeśli chodzi o mężczyzn, to widzimy że głównie przeżyli ci którzy byli w pierwszej klasie (jednak mimo wszelko sporo mężczyzn z tej klasy zginęło), oraz tylko niewielka część w klasie drugiej.

Podsumowując, z naszej analizy wynika silne zróżnicowanie względem zmiennych Pclass, Survived oraz Sex, który wynika z dobrze rozdzielonych grup punktów.