

Lista 2

Paweł Karwecki

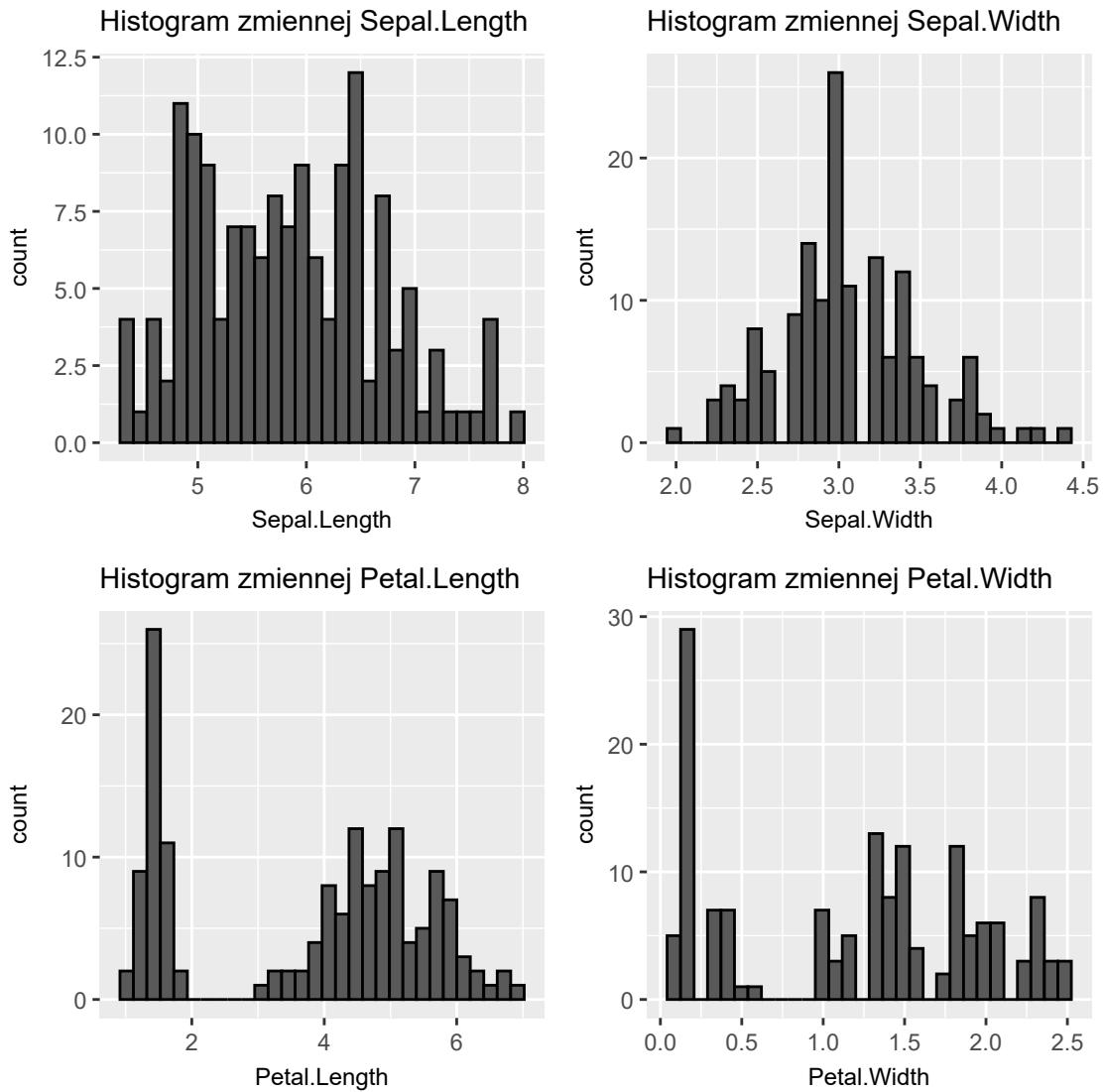
30.04.2025

Spis treści

1 Zadanie 1	2
1.1 Wybór cech	2
1.2 Porównanie nienadzorowanych metod dyskretyzacji	5
1.3 Podsumowanie	10
2 Zadanie 2	11
2.1 Opis poszczególnych cech, przygotowanie danych i ich standaryzacja	11
2.2 Wyznaczenie składowych głównych	14
2.3 Zmienna odpowiadająca poszczególnym składowym	16
2.4 Wizualizacja danych wielowymiarowych	18
2.5 Korelacja zmiennych	19
2.6 Podsumowanie	21
3 Zadanie 3	21
3.1 Opis poszczególnych cech i przygotowanie danych	21
3.2 Redukcja wymiaru na bazie MDS	22
3.3 Wizualizacja danych	23

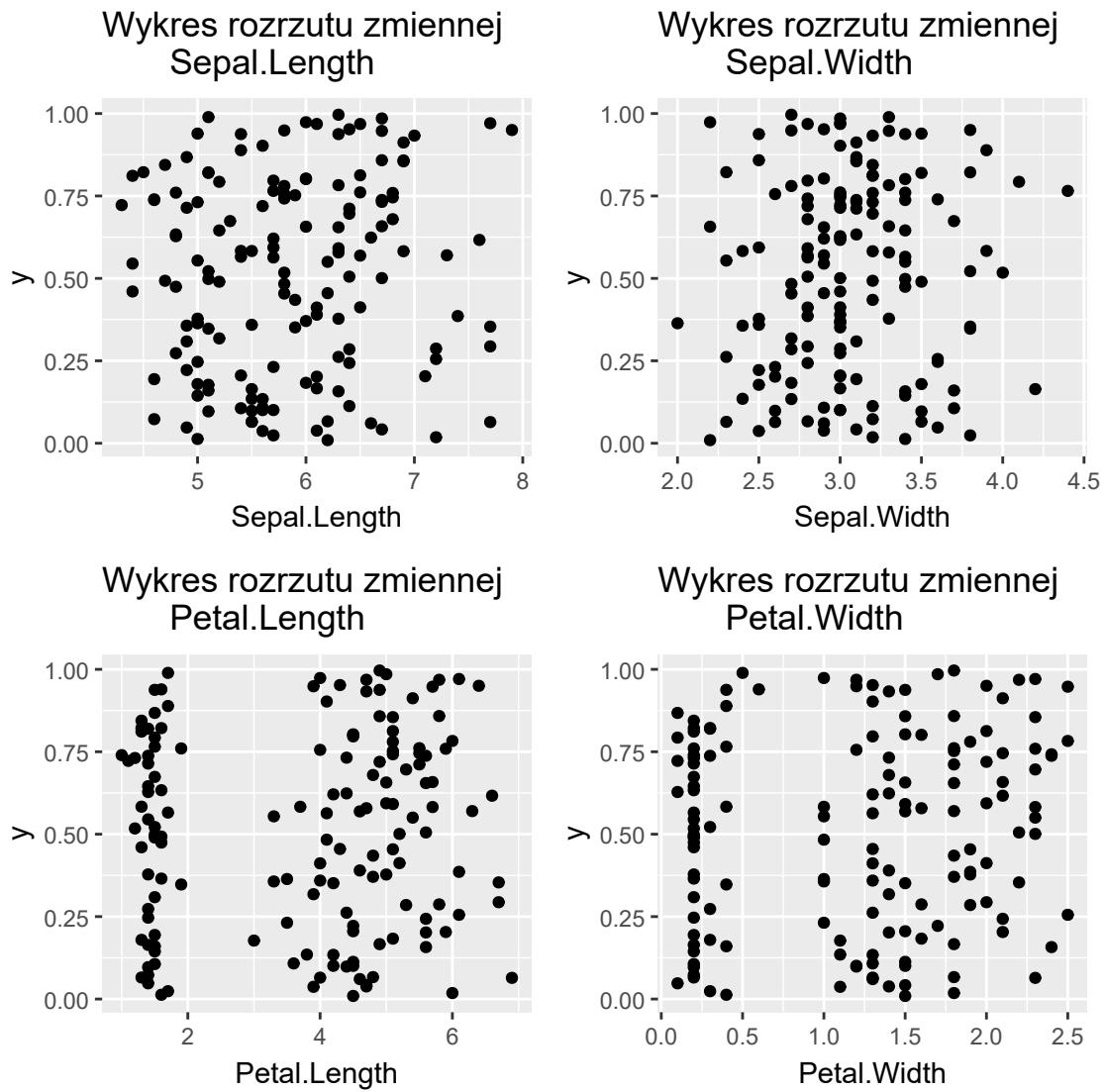
1 Zadanie 1

1.1 Wybór cech



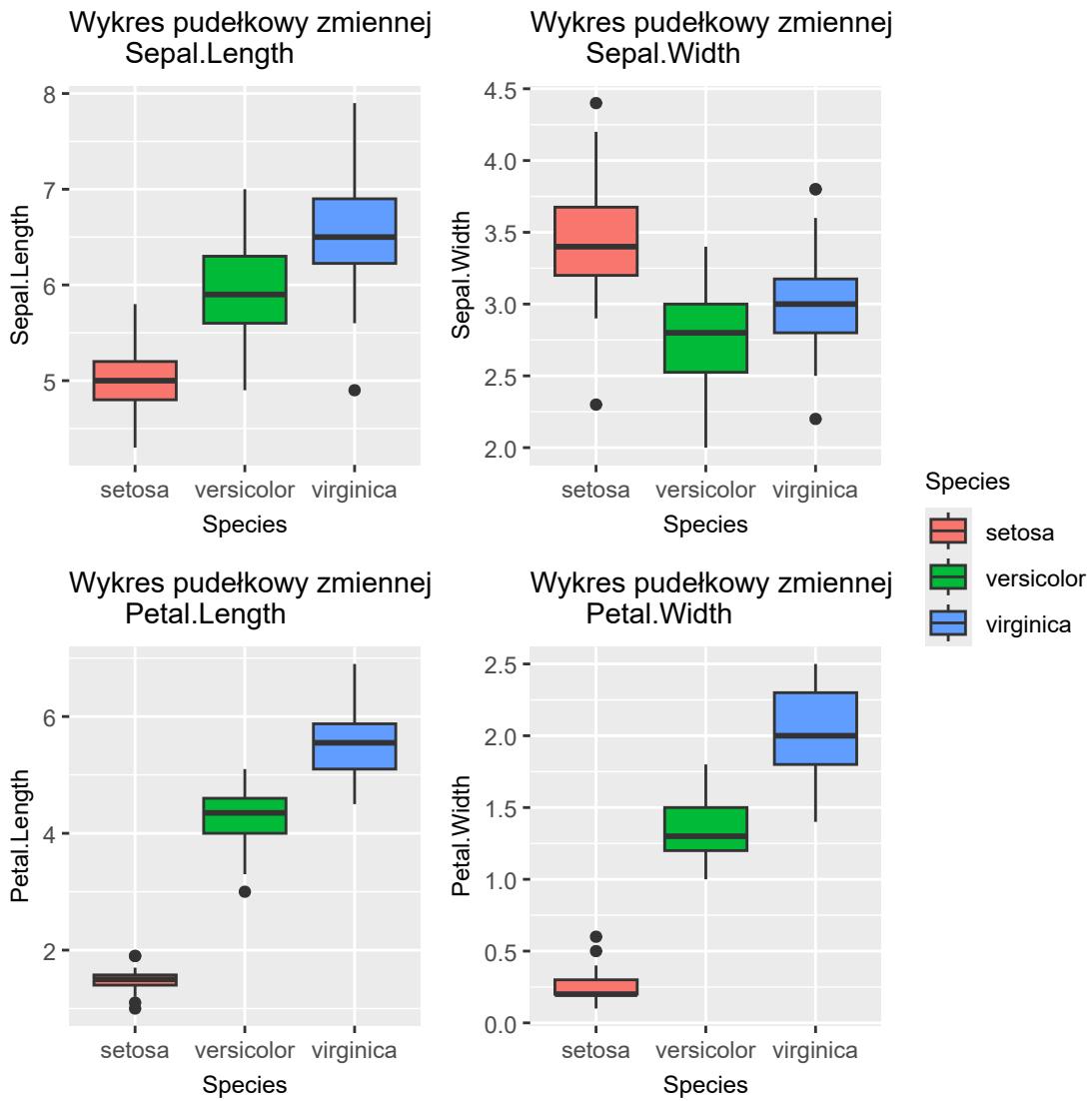
Rysunek 1: Histogramy zmiennych

Na rysunku 1 widać, że w cechach `Sepal.Length` oraz `Sepal.Width` wartości są blisko siebie, słupki są zagęszczone. Z kolei na dolnych wykresach widać szansę na znalezienie jakiejś zależności między danym gatunkiem a wartością cechy.



Rysunek 2: Wykresy rozrzutu zmiennych

Na rysunku 2 również możemy zauważać trudność w skategoryzowaniu zmiennych z górnego wykresów. Na dolnych wykresach, podobnie jak w przypadku histogramów, również można zauważać ‘przerwy’ w obserwacjach.

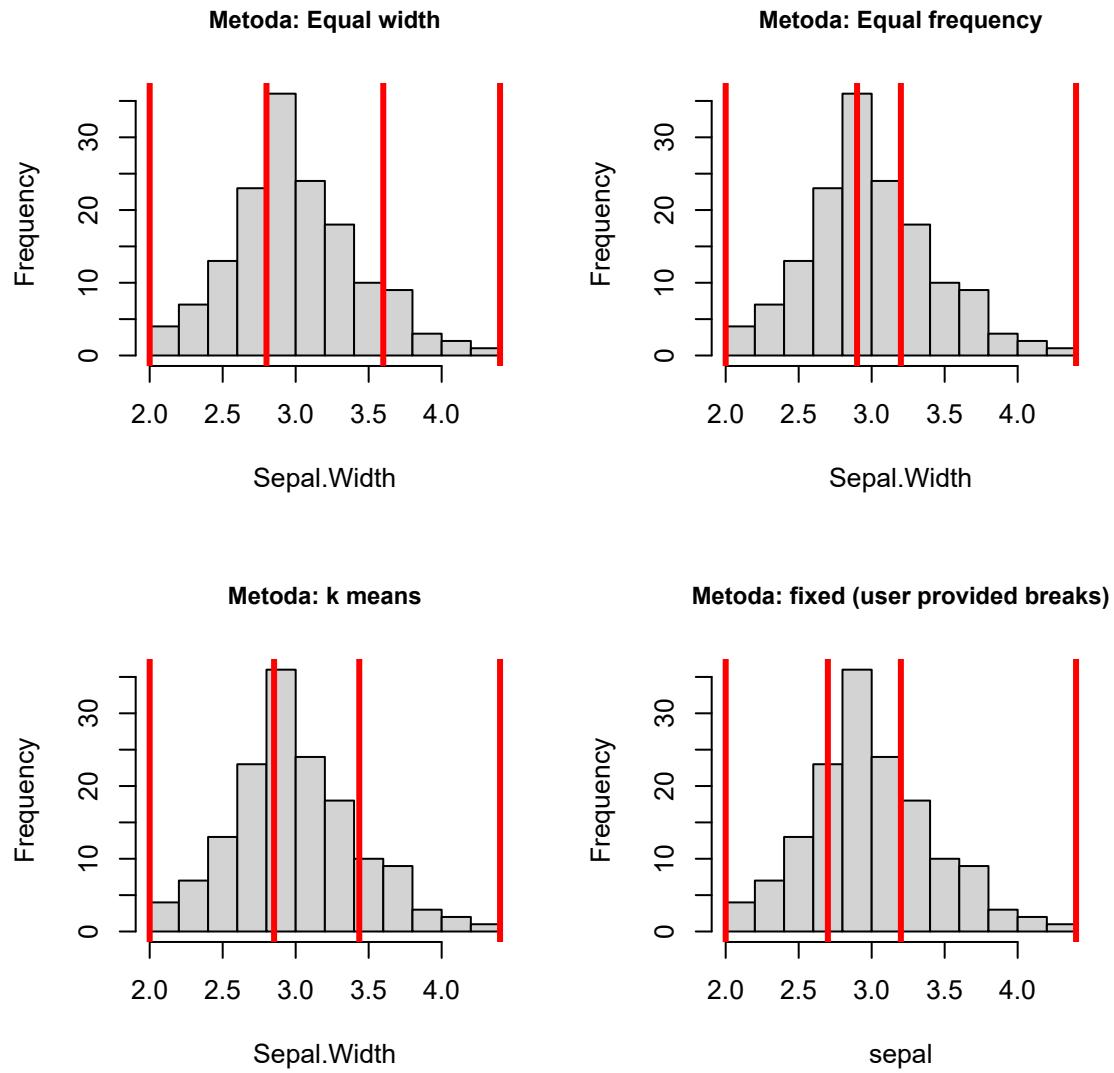


Rysunek 3: Wykresy pudełkowe zmiennych

Wykresy pudełkowe z rysunku 3 pokazują, że w `Sepal.Width` mamy większą wariancję zmiennych w zależności od ich gatunku i więcej wartości odstających niż w przypadku `Sepal.Length`. Z kolei na wykresie pudełkowym zmiennej `Petal.Length` widać, że każdy gatunek ma mniejszą wariancję niż w `Petal.Width`. Tak więc można przyjąć, że `Sepal.Width` będzie **najtrudniejszą** w dyskretyzacji zmienną, a `Petal.Length` **najłatwiejszą**.

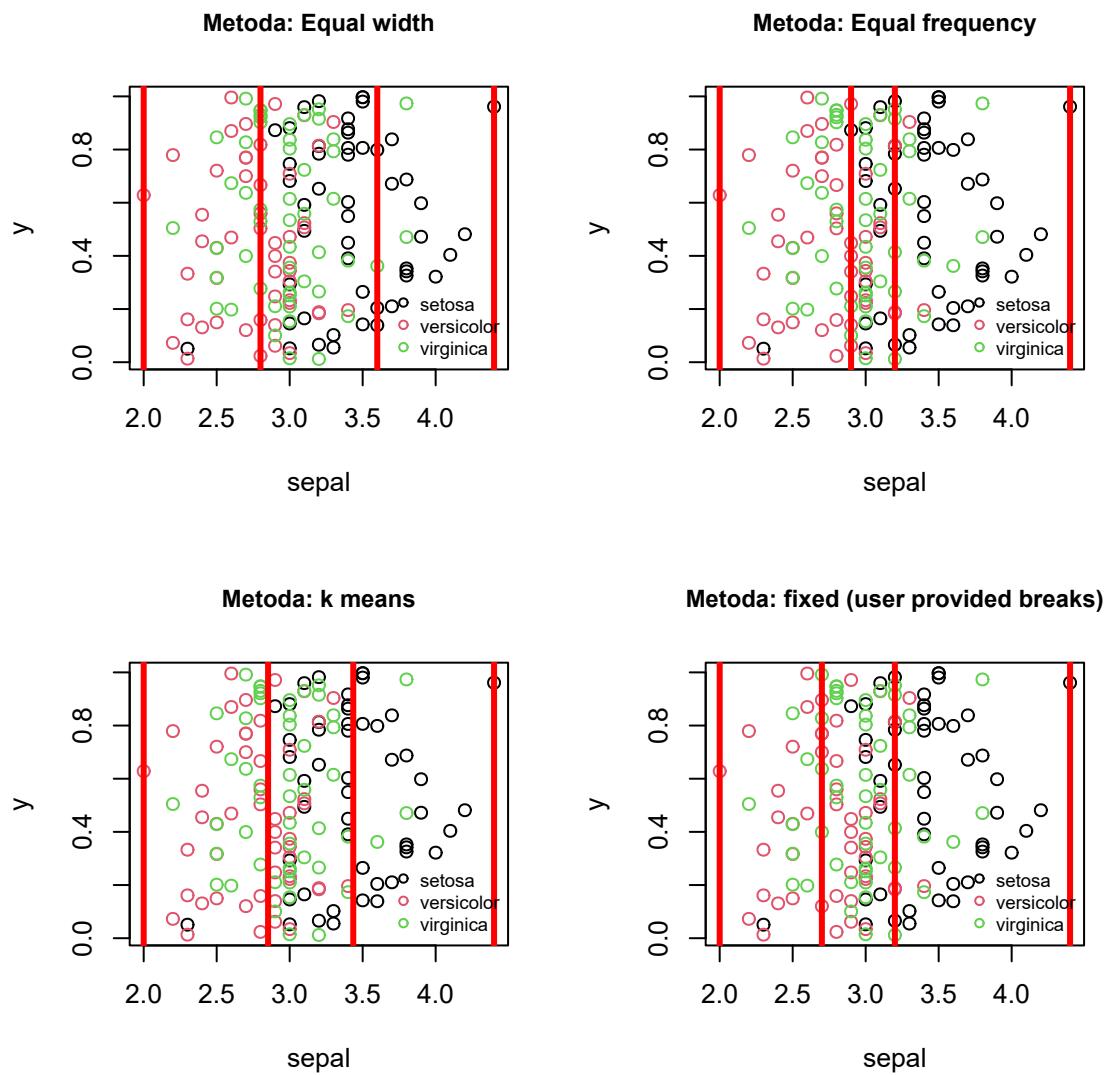
1.2 Porównanie nienadzorowanych metod dyskretyzacji

1.2.1 Sepal.Width



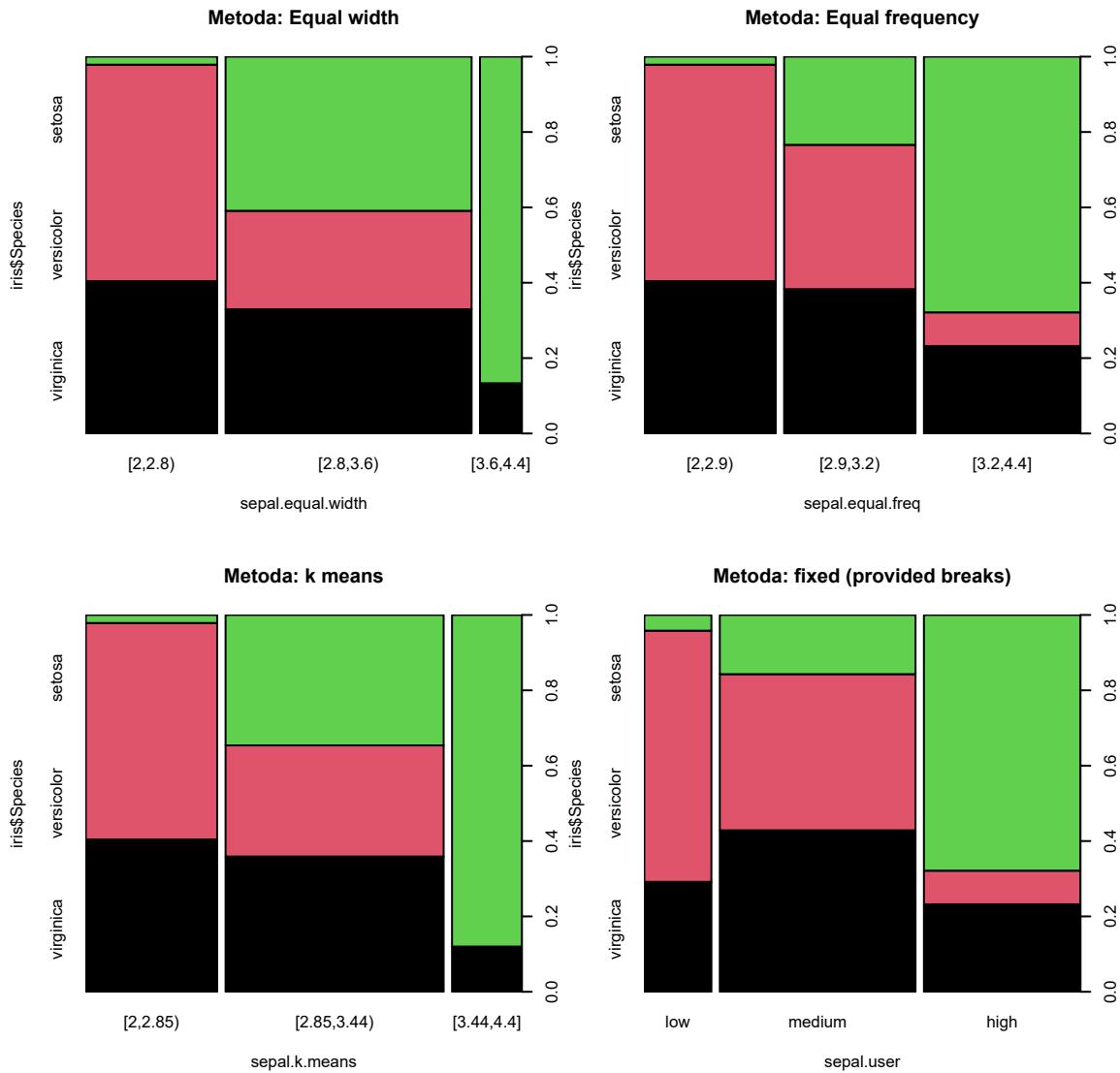
Rysunek 4: Histogramy zmiennej Sepal.Width z podziałem na 3 zbiory według danych algorytmów

Po podziałach na części z histogramów na rysunku 4 widać, że różne metody mają różne końce zbiorów. Najmniejszą częścią (pod względem długości przedziału) jest część środkowa dla metody Equal Frequency, a największą jest trzecia część w algortymie Equal Frequency i fixed (user provided breaks).



Rysunek 5: Wykresy rozrzutu zmiennej Sepal.Width z podziałem na 3 zbiory według danych algorytmów

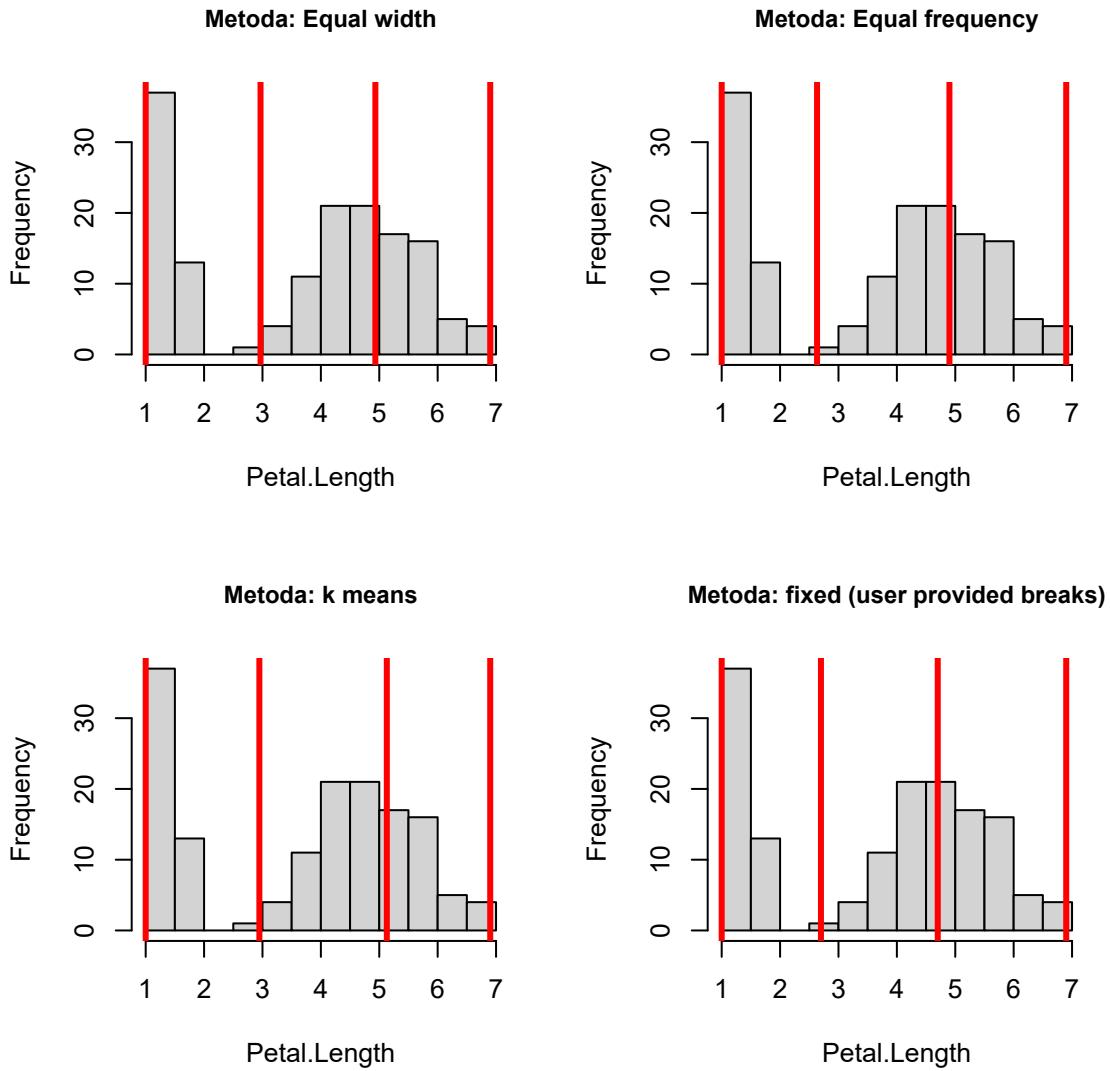
Na wykresach rozrzutu z rysunku 5 można dostrzec, że podział zmiennej na 3 zbiory jest dosyć ciężki. W każdej z metod w środkowym zbiorze mamy rekordy różnych gatunków, co zdecydowanie utrudnia dyskretyzację.



Rysunek 6: Tabela kontyngencji w formie wykresu dla różnych metod dyskretyzacji Sepal.Width

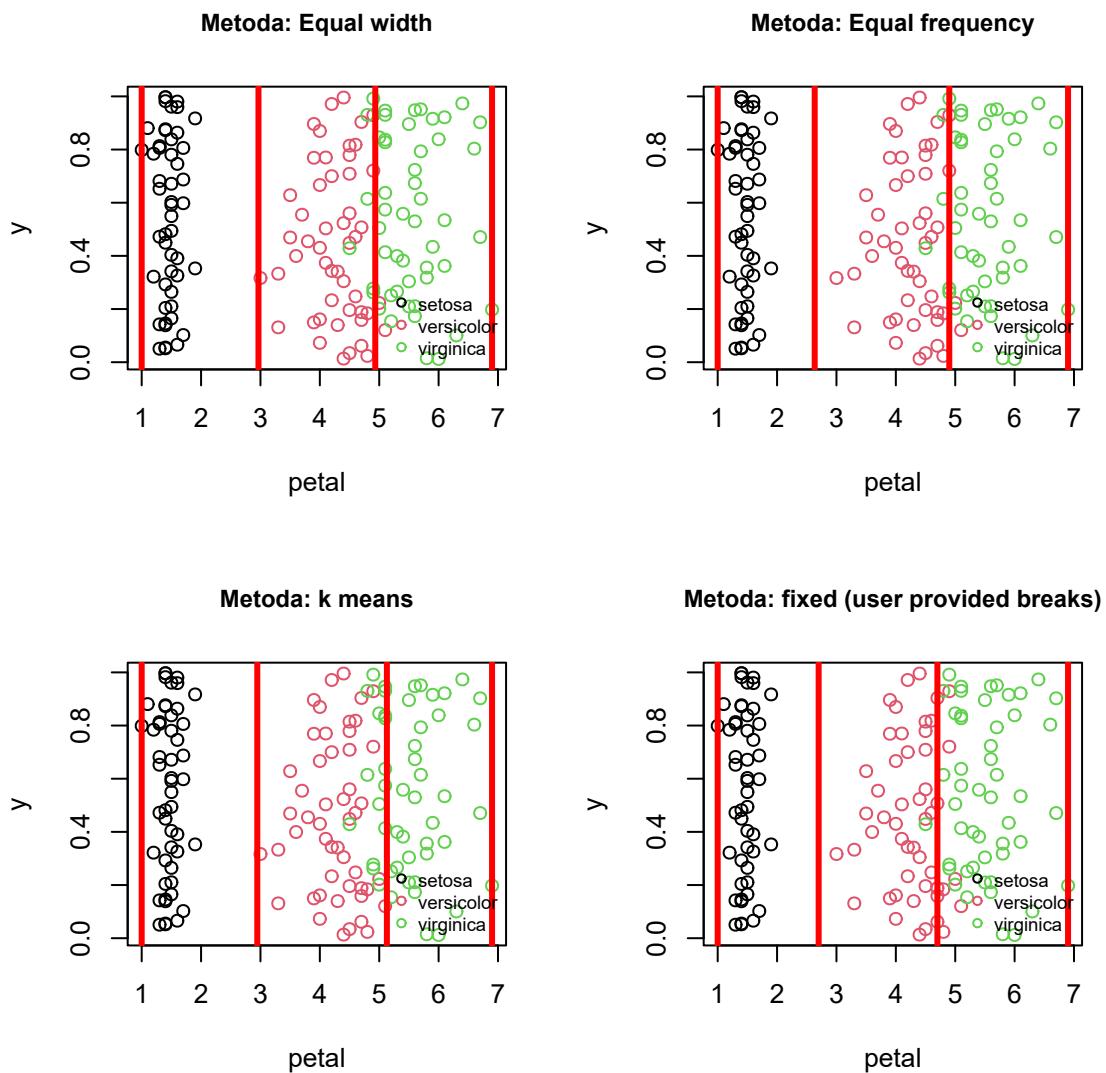
Rysunek 6 potwierdza trudność w dyskretyzacji zmiennej `Sepal.Width`. W praktyczne każdym słupku (poza 3. w `Equal Width` i `k means`) mamy różne kolory, co pokazuje zróżnicowanie rekordów. Najlepszą metodą wydaje się być `Equal Width`, ze względu na mały udział gatunku `setosa` w pierwszym zbiorze i brak `versicolor` w trzecim zbiorze.

1.2.2 Petal.Length



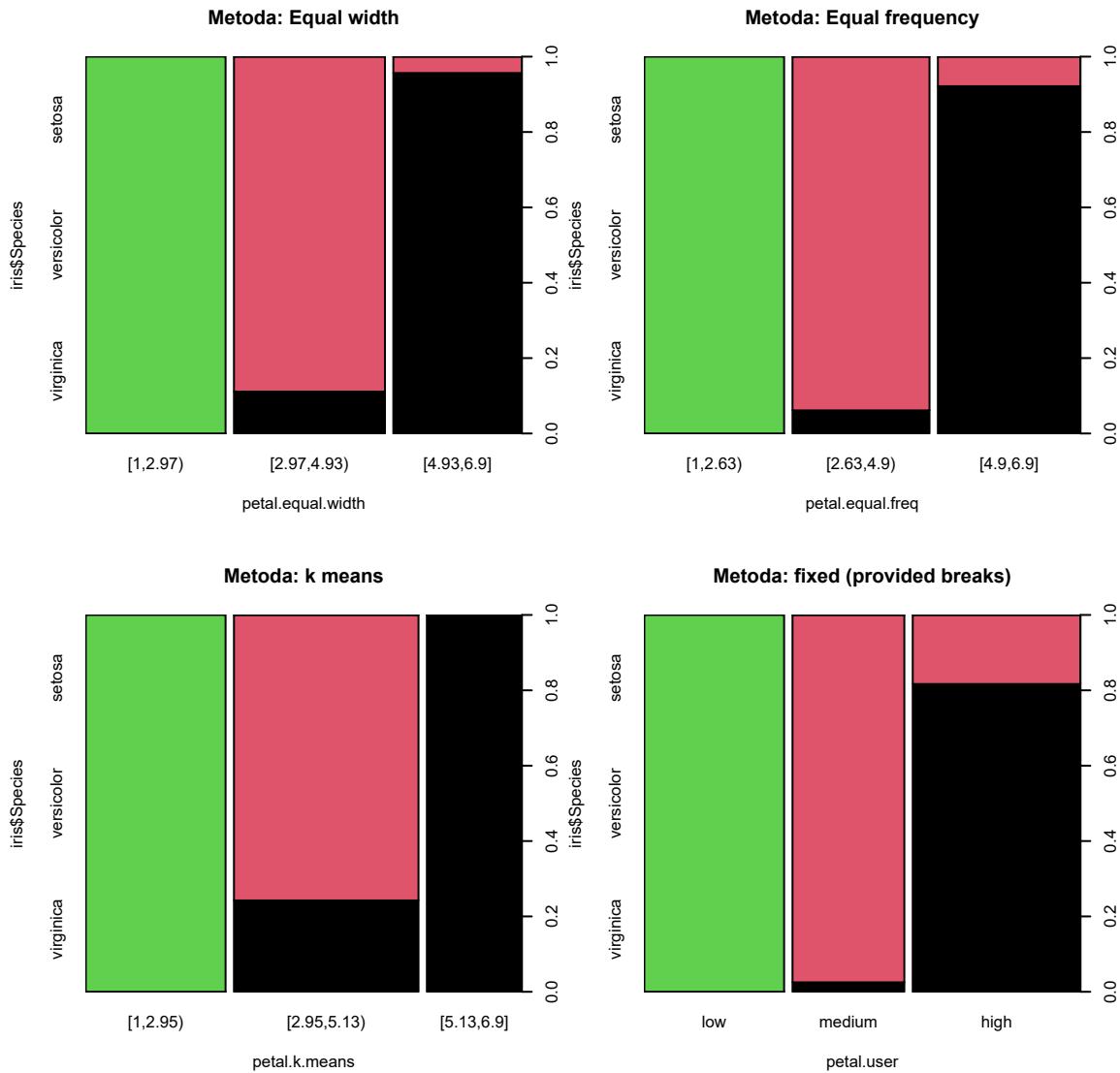
Rysunek 7: Histogramy zmiennej Petal.Length z podziałem na 3 zbiory według danych algorytmów

Histogramy z rysunku 7 pokazują, że granice zbiorów są zdecydowanie bardziej do siebie zbliżone, niż w przypadku poprzedniej zmiennej.



Rysunek 8: Wykresy rozrzutu zmiennej Petal.Length z podziałem na 3 zbiory według danych algorytmów

Wykresy rozrzutu na rysunku 8 potwierdzają, że zmienną Petal.Length już łatwiej zdyskretyzować. Na każdym z wykresów widać, że pierwszy zbiór zawiera jedynie elementy z gatunku **setosa**, drugi zbiór w znacznej części zawiera rekordy z grupy **versicolor**, a w trzecim zbiorze większość stanowi gatunek **virginica**.



Rysunek 9: Tabela kontyngencji w formie wykresu dla różnych metod dyskretyzacji Petal.Length

Wykres na podstawie tabeli kontyngencji z rysunku 9 pokazuje, że w tym przypadku faktyczna dykretyzacja jest łatwiejsza niż dla `Sepal.Width`. Wszystkie metody przyporządkowały gatunek `setosa` do pierwszego zbioru. W drugim i trzecim zbiorze już jest trochę gorzej. W najgorszym przypadku, `versicolor` stanowi mniej niż 80% drugiego zbioru (metoda `k means`), a w najlepszym około 95% (metoda `fixed (provided breaks)`). Z drugiej strony, te metody przyporządkowują gatunek `virginica` w trzecim zbiorze odpowiednio w najlepszy i najgorszy sposób. Złotym środkiem wydaje się być metoda `Equal frequency`, która dobrze dzieli 2. i 3. zbiór na gatunki z ponad 90% skutecznością.

1.3 Podsumowanie

W przypadku zmiennej `Sepal.Width`, dyskretyzacja faktycznie wydaje się być trudnym zadaniem. Żadna z metod nie daje jednoznacznej odpowiedzi na to, jak podzielić zbiory, aby utożsamić je z danym gatunkiem. Najlepiej poradziła sobie metoda `Equal Width`, która i tak ma dosyć równy udział wszystkich gatunków w drugim zbiorze. Z kolei `Petal.Length` jest zmienną, którą łatwiej skategoryzować. Najlepiej będzie wybrać

metodę **Equal frequency**, która z 90% skutecznością dzieli zmienne **versicolor** i **virginica** na dwa osobne zbiorы, a wszystkie rekordy z gatunku **setosa** przyporządkowuje pierwszemu zbiorowi.

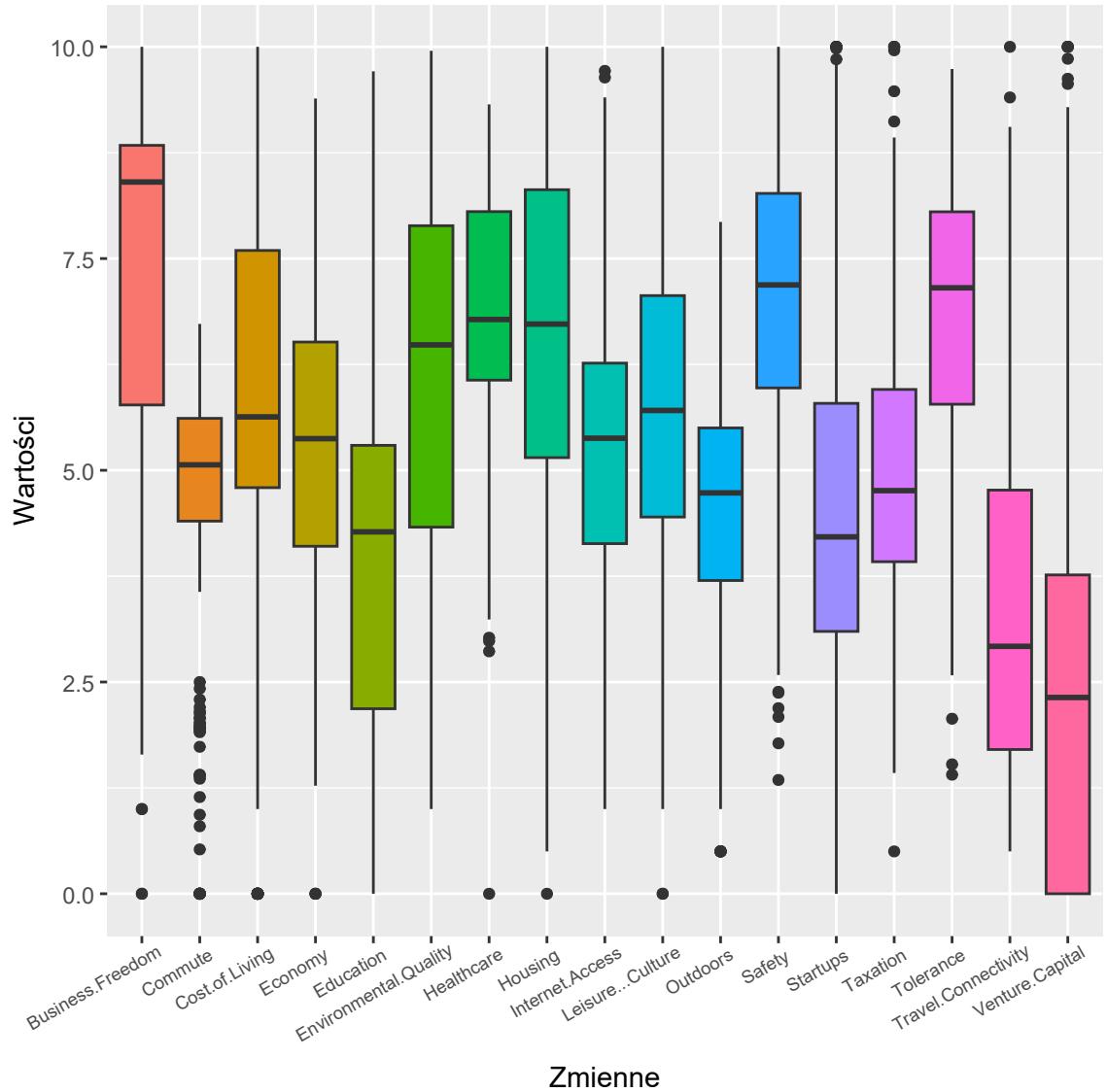
2 Zadanie 2

2.1 Opis poszczególnych cech, przygotowanie danych i ich standaryzacja

Tabela 1: Rodzaj zmiennych i wytlumaczone nazwy cech

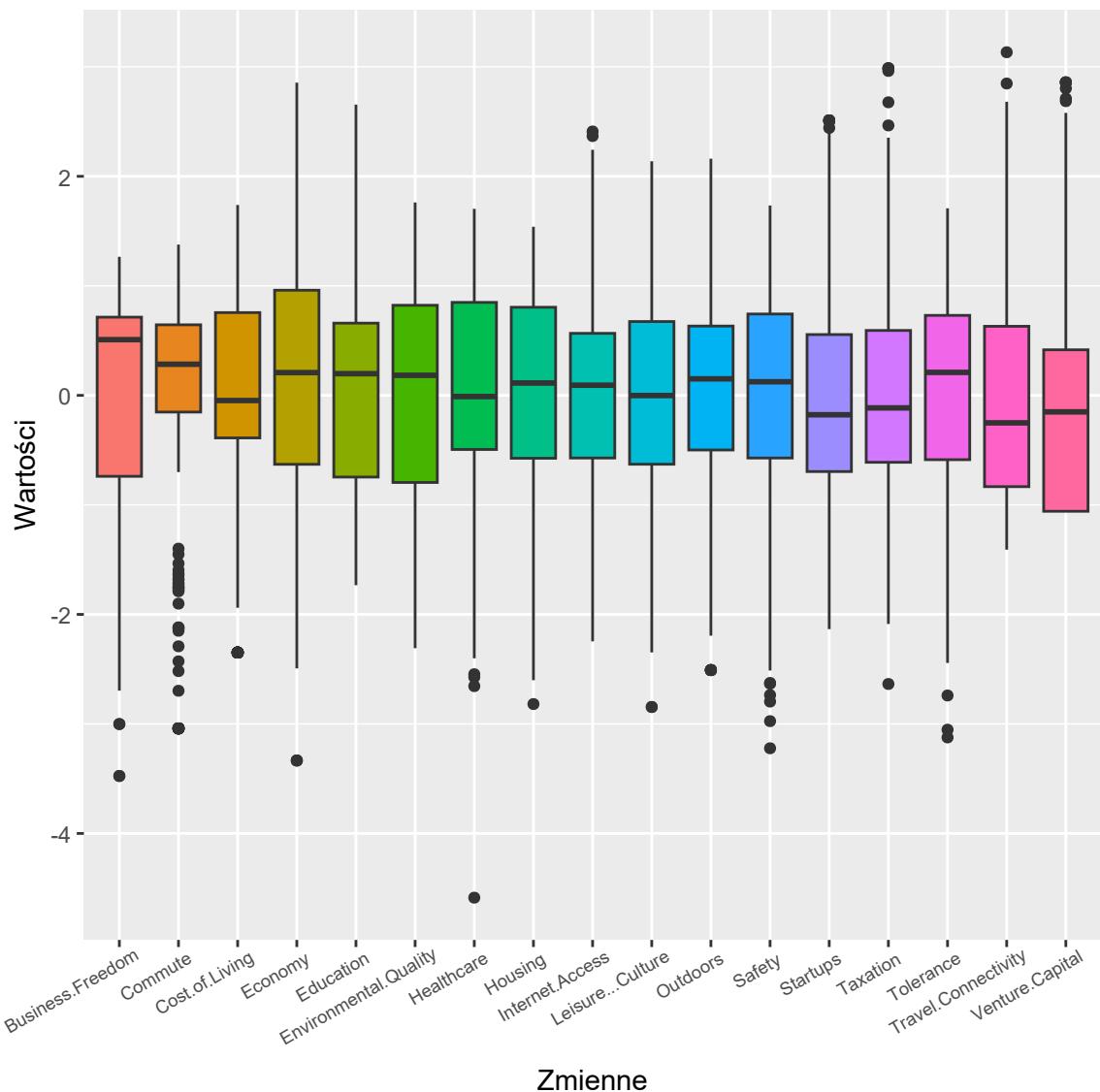
Nazwa zmiennej	Typ zmiennej	Opis zmiennej
X	integer	Identyfikator wiersza
UA_Name	character	Nazwa miasta
UA_Country	character	Nazwa kraju
UA_Continent	character	Nazwa kontynentu
Housing	numeric	Koszty wynajmu mieszkania
Cost.of.Living	numeric	Koszty życia
Startups	numeric	Możliwość założenia startupu
Venture.Capital	numeric	Finansowanie małych i średnich firm (np. startupów)
Travel.Connectivity	numeric	Połączenia komunikacji miejskiej
Commute	numeric	Jakość transportu publicznego
Business.Freedom	numeric	Wolność gospodarcza
Safety	numeric	Bezpieczeństwo
Healthcare	numeric	Służba zdrowia
Education	numeric	Edukacja
Environmental.Quality	numeric	Jakość środowiska
Economy	numeric	Ekonomia
Taxation	numeric	Podatki
Internet.Access	numeric	Dostęp do internetu
Leisure...Culture	numeric	Rozrywka i kultura
Tolerance	numeric	Tolerancja
Outdoors	numeric	Dostęp do rekreacji

Jak widać z tabeli 1, mamy 3 zmienne, które nie są ilościowe: **UA_Name**, **UA_Country**, **UA_Continent**. Nie przydadzą się one w analizie PCA, ale mogą się przydać w późniejszej identyfikacji m.in. punktów na wykresie rozrzutu. W analizie w ogóle nie przyda się zmienna **X**, więc ją też można usunąć.



Rysunek 10: Wykresy pudełkowe zmiennych ilościowych

Jak widać na rysunku 10, cechy mają bardzo różne wartości i wariancje. Spowoduje to faworyzację niektórych cech, tzn. **Buisiness Freedom** ma na ogół większą wartość niż **Venture Capital**, więc będzie bardziej wpływać na wyniki analizy. Potrzebna zatem będzie standaryzacja.

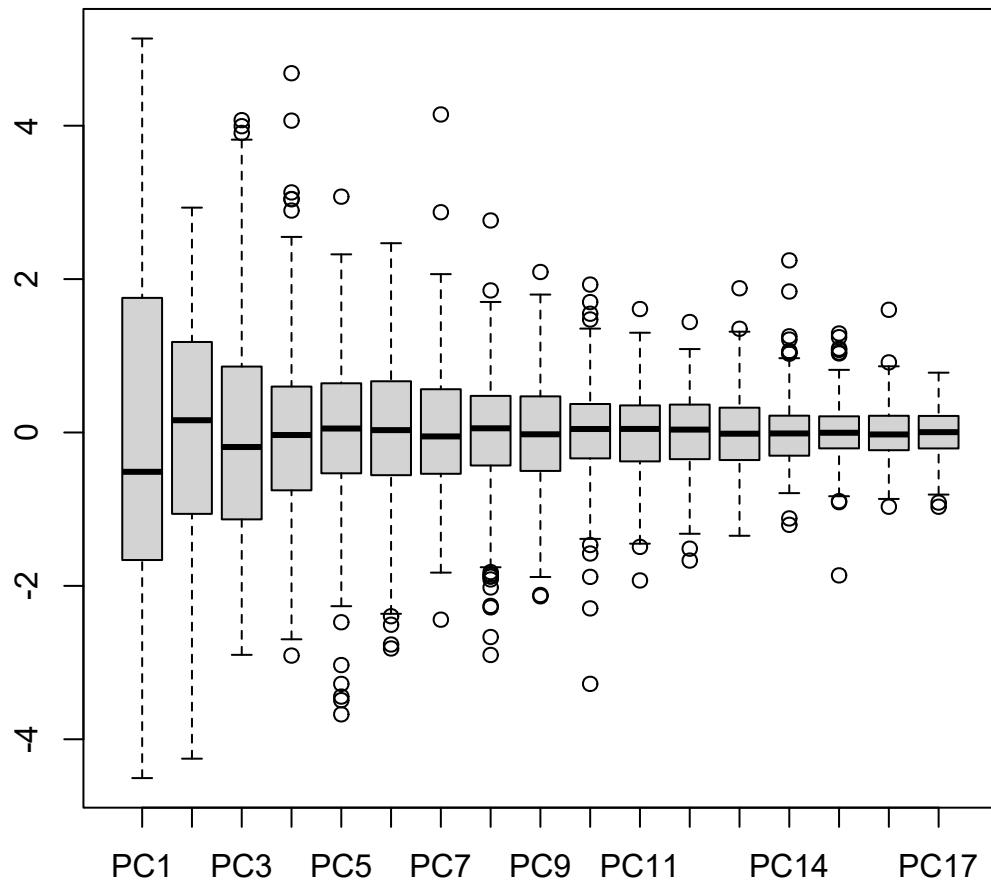


Rysunek 11: Wykresy pułapkowe zmiennych ilościowych po standaryzacji

Standaryzacja pomogła w ustaleniu mniej więcej podobnych wartości i wariancji cech, co widać na rysunku 11.

2.2 Wyznaczenie składowych głównych

Wykresy pudłkowe dla poszczególnych składowych głównych



Rysunek 12: Wykres pudełkowy składowych głównych

Na rysunku 12 widać, że początkowe składowe główne charakteryzują się wysoką wariancją, która maleje wraz z kolejnymi składowymi. W późniejszych składowych (PC3 i dalej) widoczne są również wartości odstające.

Tabela 2: Wektory ładunków dla trzech pierwszych składowych głównych

	PC1	PC2	PC3
Housing	0.3078251	0.0533534	-0.3135465
Cost.of.Living	0.2596091	-0.1757815	-0.3305352
Startups	-0.1802385	-0.4834415	0.0061000
Venture.Capital	-0.2365974	-0.4274509	0.0148768

Tabela 2: Wektory ładunków dla trzech pierwszych składowych głównych (*continued*)

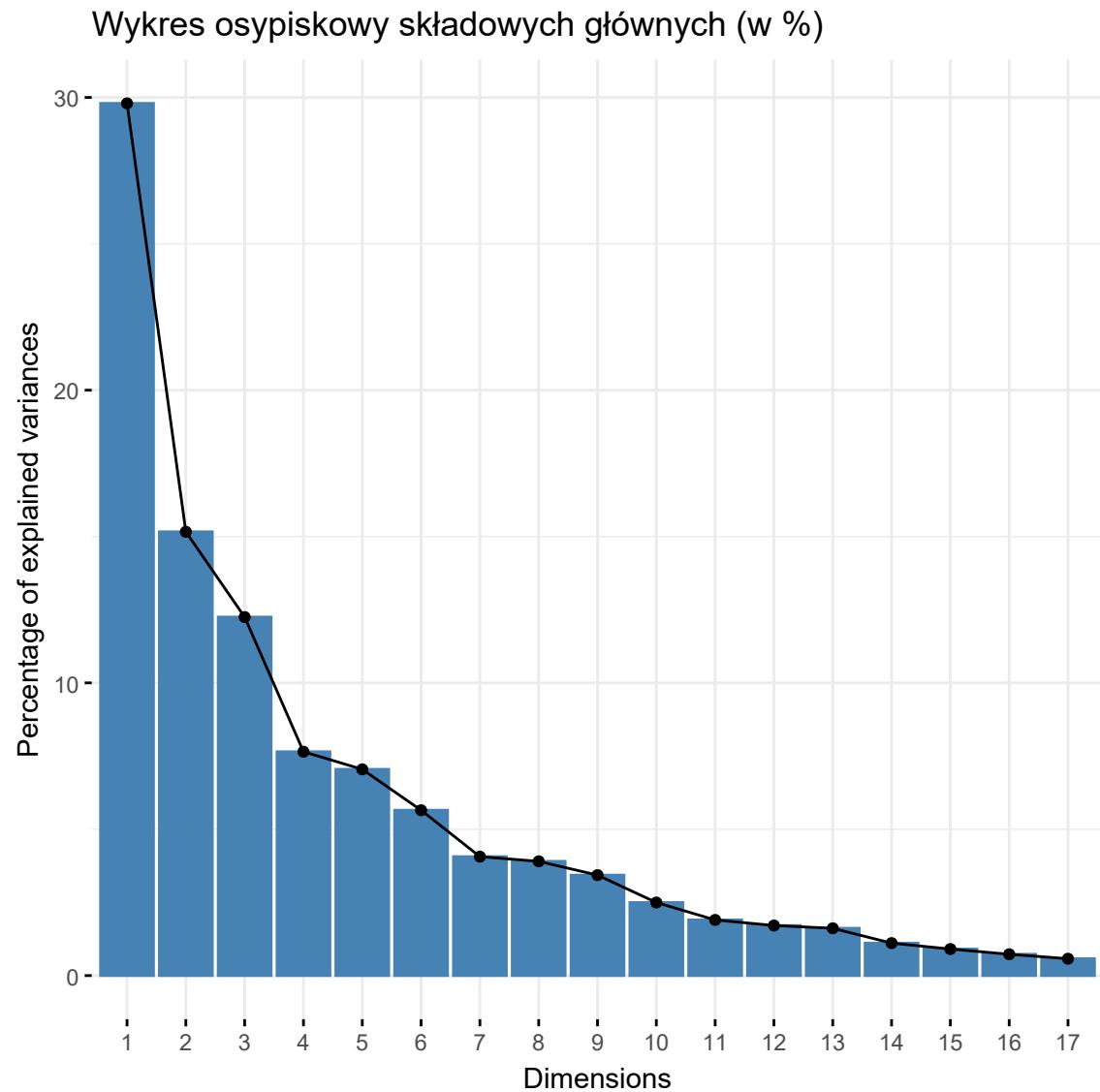
	PC1	PC2	PC3
Travel.Connectivity	-0.2094543	-0.1353067	-0.3397760
Commute	-0.1142045	0.0259310	-0.5057359
Business.Freedom	-0.3772809	0.0982196	0.0241046
Safety	-0.0389355	0.2871039	-0.3330100
Healthcare	-0.2803590	0.2419482	-0.2810248
Education	-0.4025620	-0.0490795	-0.0738645
Environmental.Quality	-0.3262220	0.2525355	0.0535717
Economy	-0.2731752	-0.0740033	0.3086705
Taxation	0.0262992	0.1074151	-0.0201849
Internet.Access	-0.2761922	0.0227056	0.0284416
Leisure...Culture	-0.0744466	-0.3647324	-0.3050545
Tolerance	-0.1897496	0.3550911	-0.1027251
Outdoors	-0.0915866	-0.1933825	-0.1485868

Tabela 2 przedstawia wektory ładunków dla trzech pierwszych składowych głównych. W PC1 dużą wagę (powyżej 0.25) stanowią: **Business Freedom, Healthcare, Education, Environmental Quality, Economy, Internet Access** (na minus) i **Housing, Cost of Living** (na plus). Zmienna pokazuje pewien kontrast - im lepiej się żyje w danym mieście, tym większe są ceny. Jeśli wartość PC1 będzie wysoka, to koszty utrzymania się będą atrakcyjne, ale za to jakość życia będzie niska.

W PC2 dużą wagę mają cechy **Safety, Environmental Quality, Tolerance** (na plus) i **Startups, Venture Capital, Leisure Culture** (na minus), czyli w miastach o wysokiej wartości w PC2 jest bezpiecznie, ale ciężko jest rozpocząć biznes.

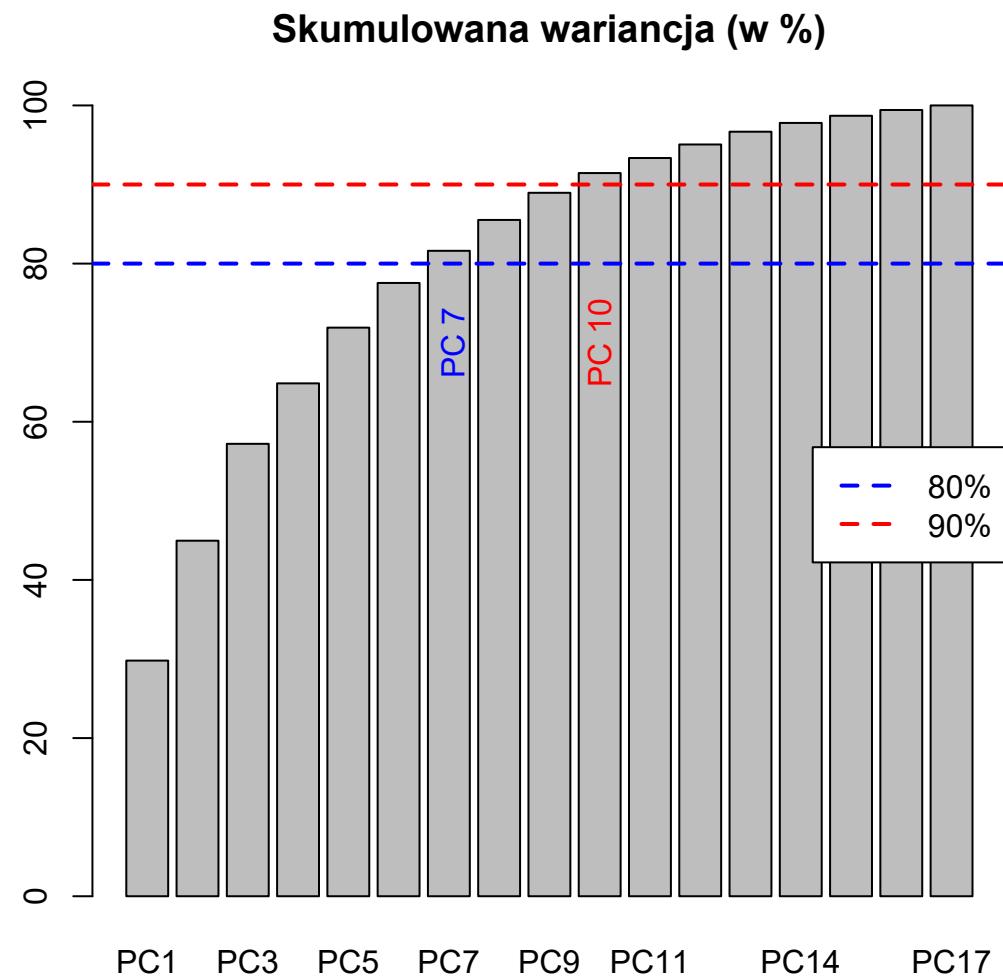
Z kolei w PC3 wysoką wagę mają zmienne **Housing, Cost of Living, Travel Connectivity, Commute, Safety, Leisure Culture** (minus) oraz **Economy** (plus). W miastach o niskiej wartości w tej składowej koszty życia są niskie, jest dobra komunikacja miejska i jest bezpiecznie, ale ekonomia jest w słabej kondycji.

2.3 Zmienna odpowiadająca poszczególnym składowym



Rysunek 13: Wykres osypiskowy składowych głównych

Jak widać na rysunku 13, prawie 30% wariancji jest wyjaśnione przez PC1, PC2 tłumaczy około 15% wariancji, a PC3 ponad 10%. Kolejne składowe tłumaczą mniejszą część wariancji, w szczególności PC15, PC16 i PC17 tłumaczą mniej niż 1% (każda z nich, nie w sumie).

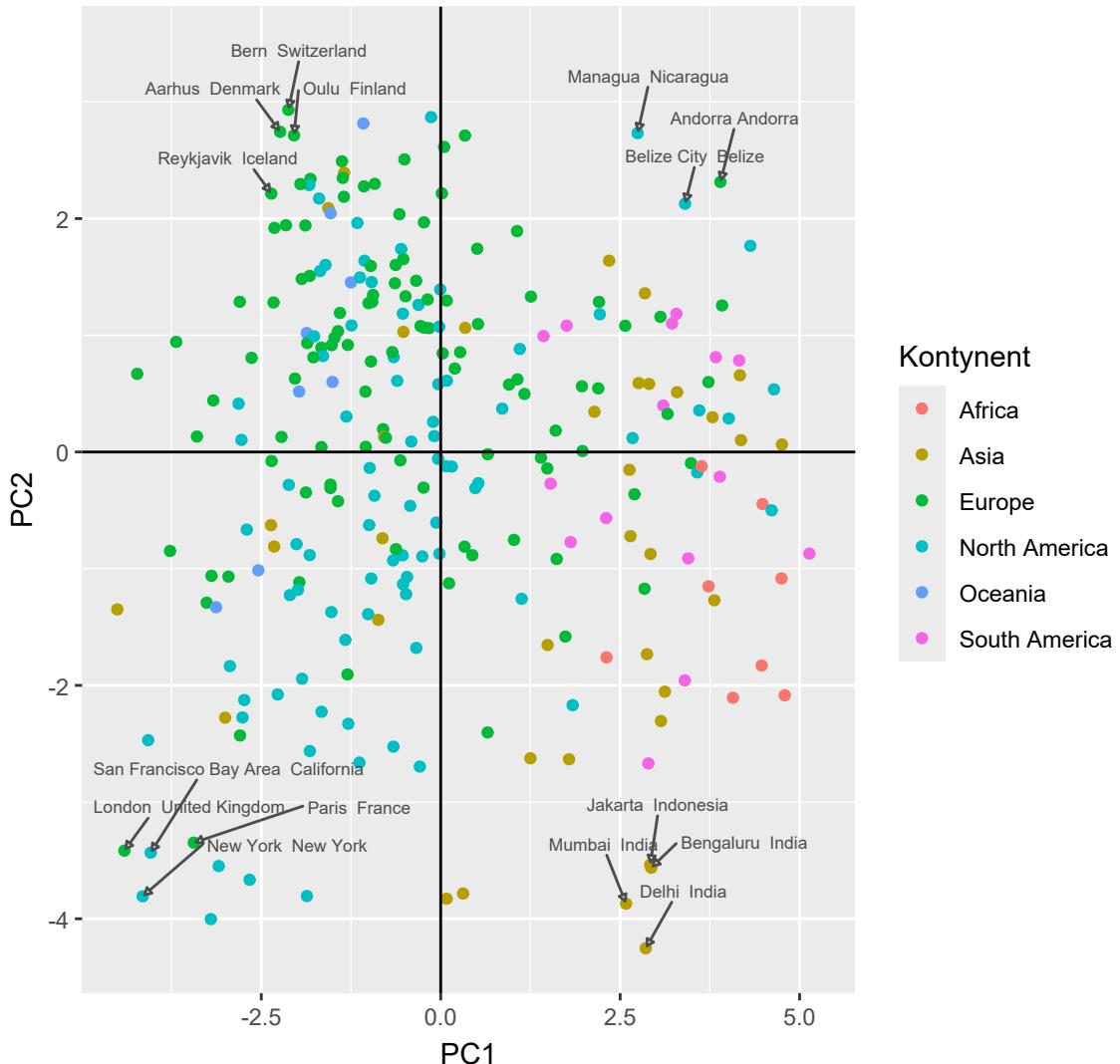


Rysunek 14: Wykres skumulowanej wariancji

Wykres słupkowy na rysunku 14 pokazuje, że do wyjaśnienia 80% wariancji potrzebujemy aż 7 składowych głównych, a w przypadku 90% - 10 składowych głównych.

2.4 Wizualizacja danych wielowymiarowych

Wykres rozrzutu dla dwóch pierwszych składowych głównych



Rysunek 15: Wykres rozrzutu dla dwóch pierwszych składowych głównych

Z wykresu rozrzutu na rysunku 15 można odczytać kilka ciekawych rzeczy. Po pierwsze, duża część europejskich miast ma wartości mniejsze od 0 w PC1 i większe od 0 w PC2 (II ćwiartka), co podkreśla wysoką jakość życia i bezpieczeństwo w Europie, ale również wysokie koszty i trudności z założeniem własnej firmy. Wyróżniają się tu takie miasta jak duński Aarhus, islandzki Reykjavik, fińskie Oulu (miasta ze skandynawskich państw) i Berno, stolica Szwajcarii.

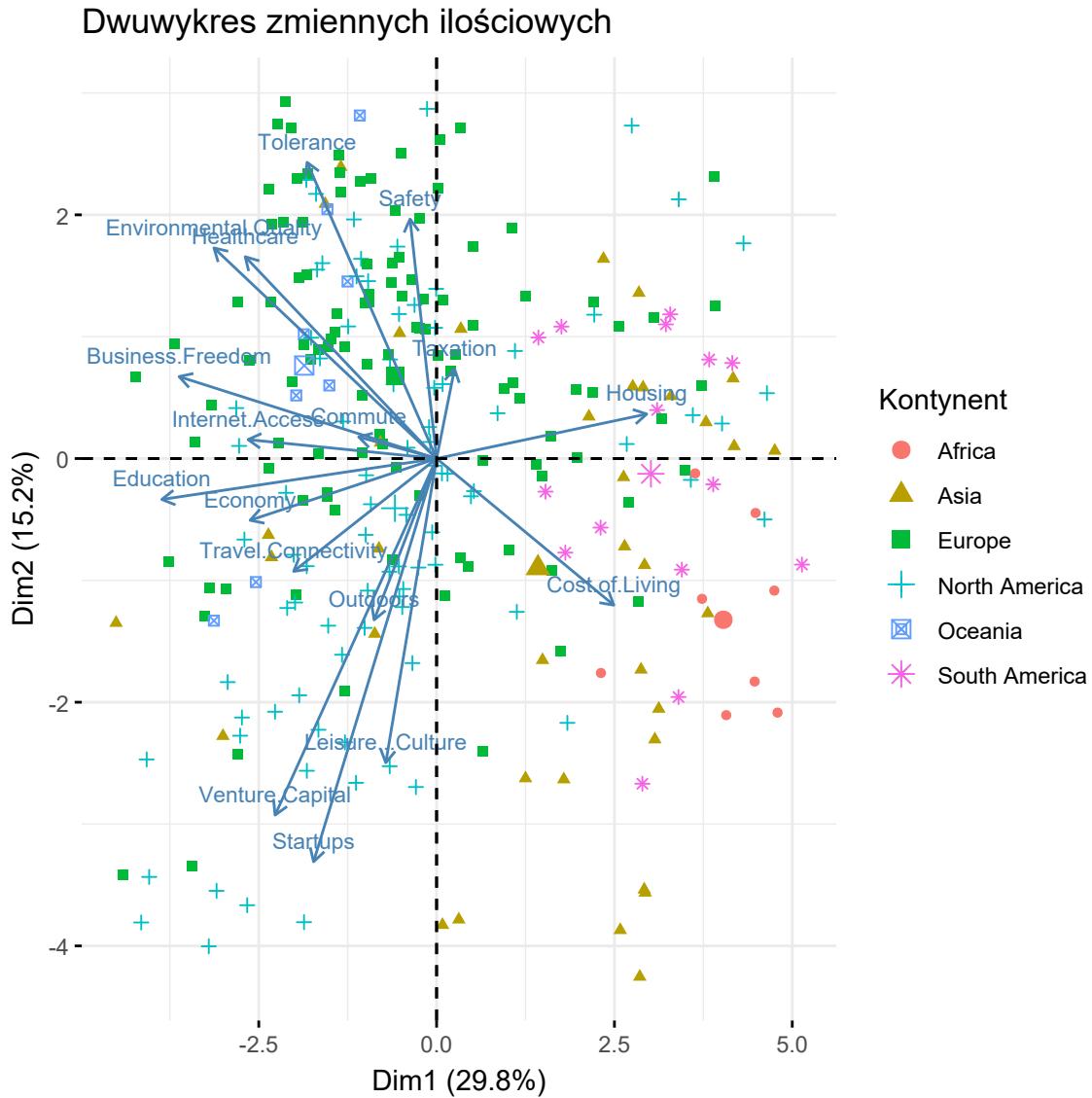
W I ćwiartce wykresu również są miasta bezpieczne, w których ciężko o swój biznes. Tutaj jednak koszty utrzymania są niskie, ale również jakość życia jest niska. Z odstających wartości, takich jak Andorra, Belize City czy Managua możemy wywnioskować, że w tej ćwiartce są prawdopodobnie miasta z małych państw czy też miasta, które są raczej celami podróży na wakacje.

Przechodząc do IV ćwiartki, są tu miasta z niskimi kosztami utrzymania, w których łatwo założyć startup, ale jakość życia i bezpieczeństwo są na niskim poziomie. Są tu wszystkie miasta afrykańskie oraz sporo

azjatyckich, w tym wyróżniające się aglomeracje Indii czy też stolica Indonezji - Jakarta. Wspomniane aglomeracje są dużymi miastami, które są bardzo zróżnicowane pod względem społecznym, skąd może wynikać ich umiejscowienie na wykresie.

W III części większość punktów stanowią miasta z Ameryki Północnej, szczególnie z USA. Nie powinno więc dziwić, że w tych miastach łatwo założyć swój biznes, jakość życia jest wysoka, ale bezpieczeństwo i koszty utrzymania się są również niemałe. W tej części wykresu najprawdopodobniej są duże miasta z rozwiniętych państw, jak chociażby wyróżnione San Francisco i Nowy Jork, ale również europejski Londyn i Paryż.

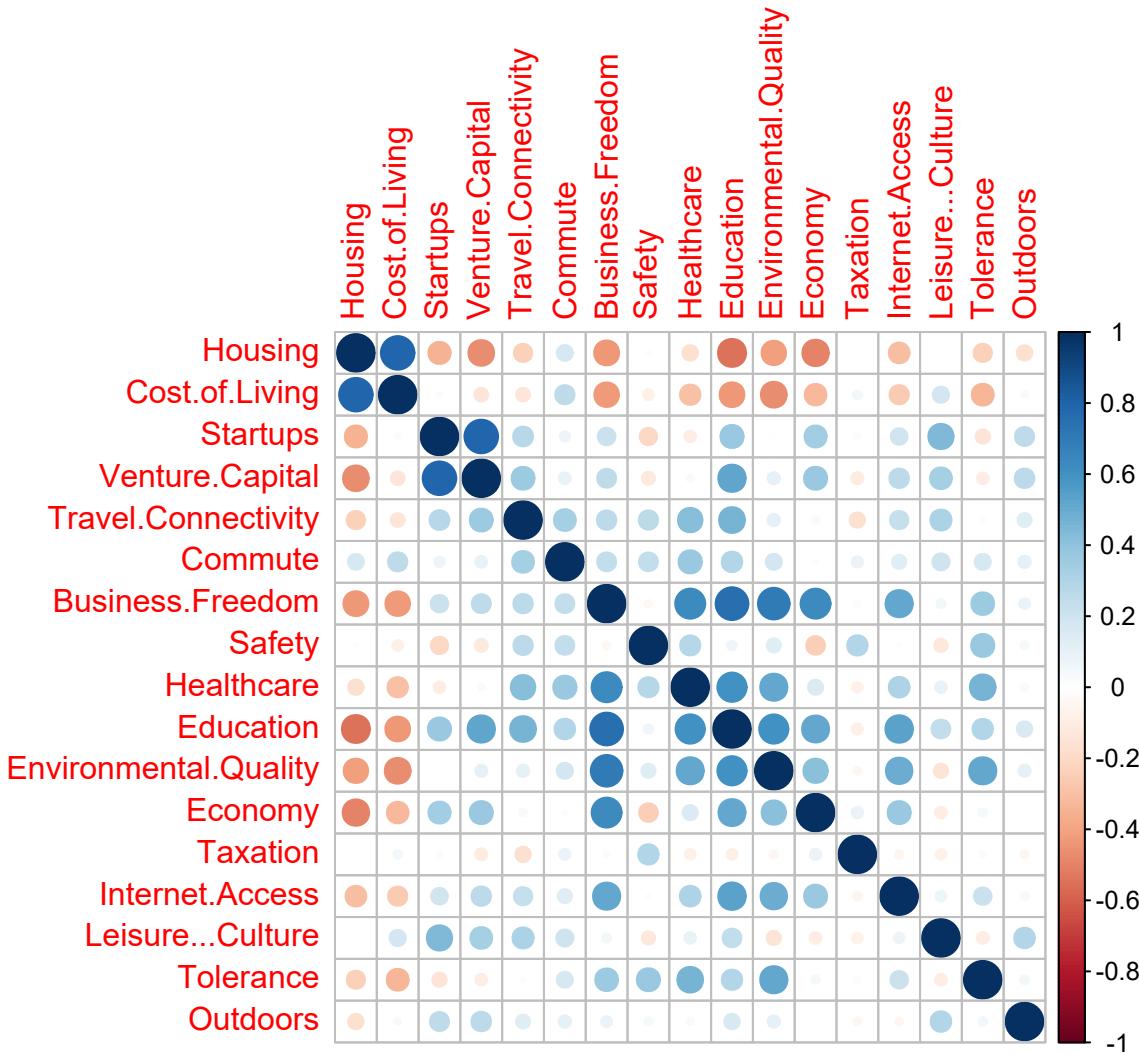
2.5 Korelacja zmiennych



Rysunek 16: Dwuwykres zmiennych ilościowych

Z rysunku 16 można odczytać, że zmienne o tym samym kierunku będą miały silną, dodatnią korelację, a te o przeciwnym kierunku - silną, ujemną korelację. Widać, że zmienne Venture Capital i Startups mają silną, dodatnią korelację. Wyróżnia się również zmienne Cost of Living i Housing, które mają ujemną

korelację z cechami świadczącymi o jakości życia (II ćwiartka), a te są ze sobą silnie skorelowane. Widać również zależność między miastami z danych kontynentów, a kierunkami zmiennych. W okolicach zmiennych świadczących o jakości życia, mamy dużo miast europejskich. W ćwiartce, w której jest Cost of Living, jest sporo miast z Afryki i Azji. Z kolei w kierunku zmiennych dotyczących biznesu, znajdują się miasta z Ameryki. Sprawdźmy, jak faktycznie wygląda macierz korelacji.



Rysunek 17: Macierz korelacji zmiennych ilościowych

Wcześniej przewidywania są mniej więcej zgodne z rysunkiem 17. Widoczna jest silna korelacja: Startups z Venture Capital; Business Freedom z Education, Environmental Quality, Economy i Healthcare; Housing z Cost.of.Living.

Ujemna, silna korelacja jest chociażby między cechą Housing i zmiennymi Education, Environmental Quality, Economy, Business Freedom i Venture Capital.

2.6 Podsumowanie

Analiza PCA pozwoliła na wyciągnięcie ciekawych wniosków. Z miast, w których poziom życia jest na wysokim poziomie, jest bezpiecznie, ale założenie biznesu jest ciężkie, większość stanowią miasta europejskie. Potencjalny przedsiębiorca szukałby miejsca zamieszkania w amerykańskich miastach, lecz musiałby się liczyć z niskim poziomem bezpieczeństwa i wysokimi kosztami życia. Niski koszt utrzymania się, ale i wysokie niebezpieczeństwo charakteryzuje kolejne miasta afrykańskie i azjatyckie. Mimo, że do wyjaśnienia 80% wariancji potrzebne jest aż 7 składowych głównych, to analiza z uwzględnieniem tylko dwóch pierwszych składowych dała raczej zrozumiałe wyniki w kwestii możliwości korelacji zmiennych. Zastosowanie standaryzacji było kluczowe, ze względu na zbyt duże zróżnicowanie w wariancji cech.

3 Zadanie 3

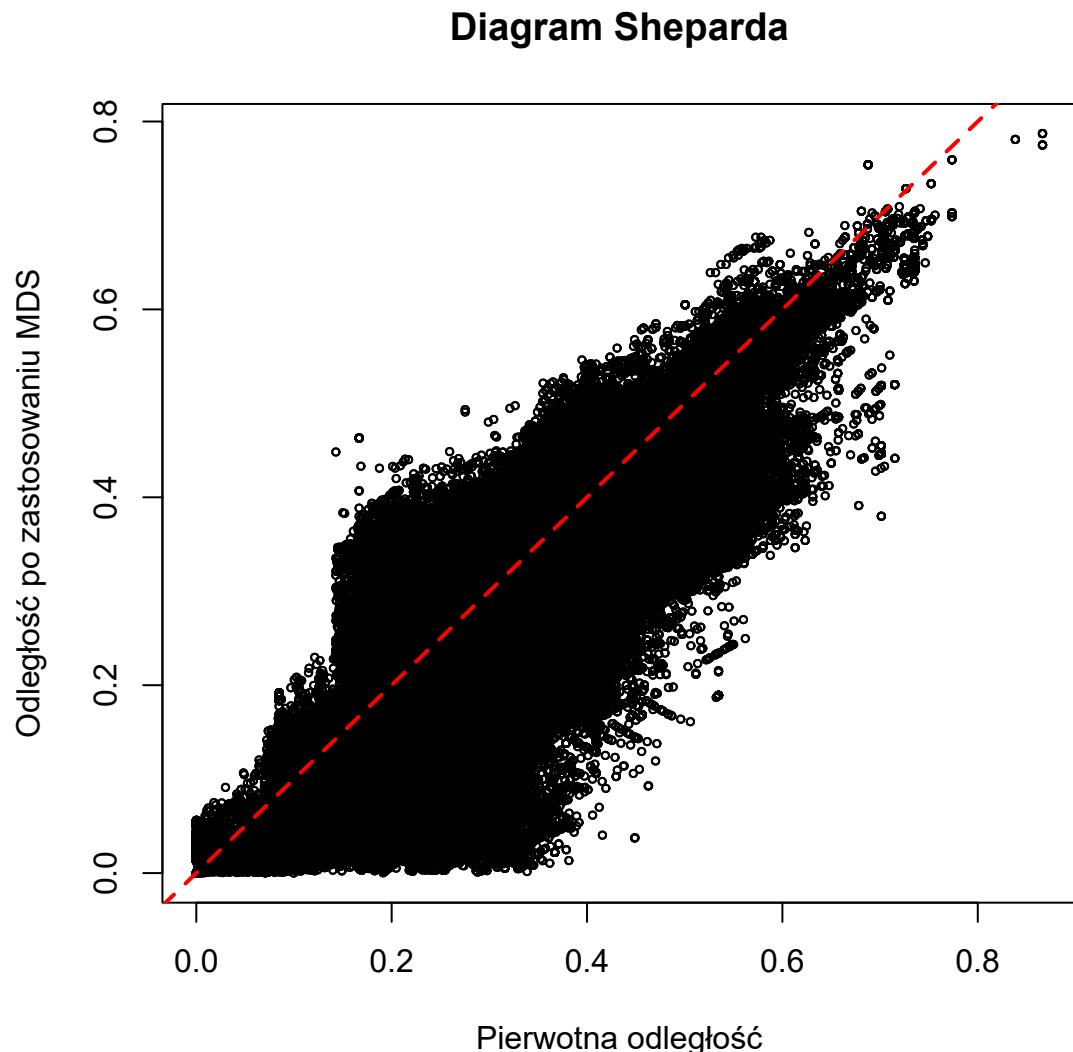
3.1 Opis poszczególnych cech i przygotowanie danych

Tabela 3: Rodzaj zmiennych i wyjaśnione nazwy cech

Nazwa zmiennej	Typ zmiennej	Opis zmiennej
PassengerId	integer	Numer ID pasażera
Survived	integer	Czy pasażer przeżył
Pclass	integer	Klasa biletu
Name	character	Imię i nazwisko pasażera
Sex	character	Płeć
Age	numeric	Wiek
SibSp	integer	Liczba rodzeństwa/współmałżonków na pokładzie
Parch	integer	Liczba rodziców/dzieci na pokładzie
Ticket	character	Numer biletu
Fare	numeric	Koszt biletu
Cabin	character	Numer kabiny
Embarked	character	Miasto wejścia na pokład

Na podstawie zawartości tabeli 3 można zauważyć, że zmienne `Survived`, `Pclass`, `Sex` i `Embarked` można zamienić na dane typu `factor`. W celu analizy MDS usunięte zostaną cechy: `PassengerId`, `Name`, `Ticket`, `Cabin`.

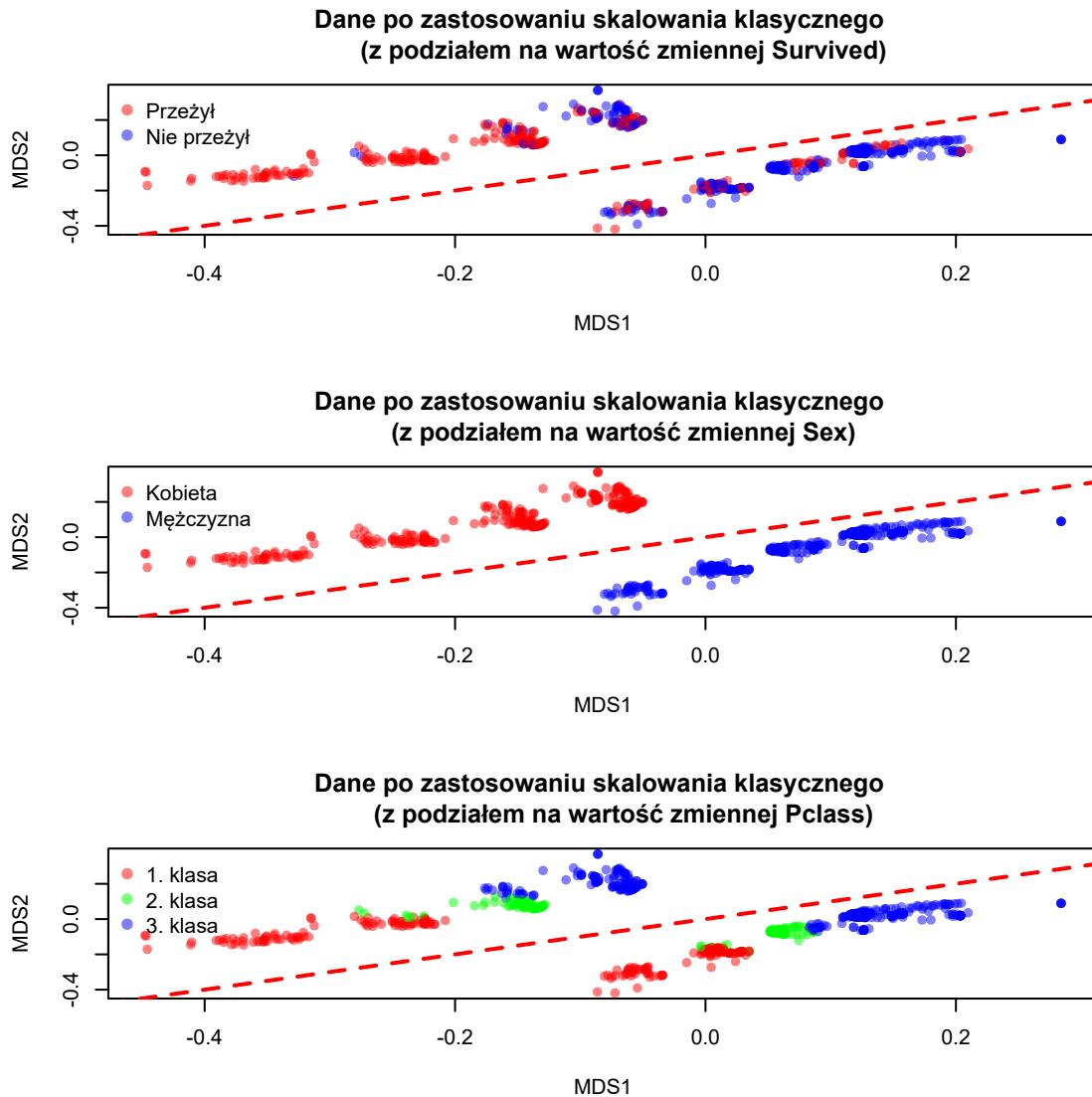
3.2 Redukcja wymiaru na bazie MDS



Rysunek 18: Diagram Sheparda dla zbioru danych

Punkty na rysunku 18 układają się wzduł prostej $y = x$, ale widać, że zbiór punktów jest dosyć “szeroki”, to znaczy mamy wiele punktów, które mają tę samą odległość po zastosowaniu skalowania MDS, ale w pierwotnej wersji są od siebie oddalone. Na przykład, punkty o odległości 0 mają w pierwotnej wersji odległość od 0 aż do wartości przekraczającej 0.3. Ciekawe jest również to, że 42.926% odległości zwiększyło się przez skalowanie MDS, 57.071% się zmniejszyło, a 0.002% nie zmieniło odległości.

3.3 Wizualizacja danych



Rysunek 19: Wykresy rozrzutu po redukcji wymiaru

Na rysunku 19 widać rozproszenie punktów po zastosowaniu skalowania wielowymiarowego. Na samej górze jest wykres z podziałem graficznym punktów ze względu na wartość zmiennej **Survived**. Widać, że punkty są podzielone na 2 grupy oddzielone czerwoną, przerywaną linią. Podział jest spowodowany wartością zmiennej **Sex**, co jest widoczne na środkowym wykresie. W grupie poniżej wspomnianej prostej duża część osób zginęła przez katastrofę. Nie można tego samego powiedzieć o grupie osób ponad prostą, w której większość stanowią osoby, które przeżyły. Wśród tych 2 grup można wyróżnić jeszcze podział na 3-4 podgrupy. Jest podział spowodowany zmienną **Pclass** (dolny wykres). Z odstających wartości: mamy kilka osób, które w MDS1 mają wartość mniejszą od -0.4 oraz jest osoba, która dla tej samej zmiennej ma wartość powyżej 0.2.

Podsumowując, większość ofiar w katastrofie Titanica stanowią mężczyźni. Wśród tych ofiar, większość z nich miała bilet 3. klasy (52% poszkodowanych). Kobiety z 3. klasy również poniosły śmiertelne konsekwencje wypadku, ale warto zwrócić uwagę na to, że prawie wszystkim kobietom z 1. i 2. klasy udało się przeżyć (tylko 3% z nich zginęło).