

Sprawozdanie z listy 2

Eksploracja danych

Marta Stankiewicz (282244)

Paweł Nowak (282223)

2025-04-14

Spis treści

1	Ocena zdolności separacyjnych zmiennych, dyskretyzacja zmiennych ciągłych	2
1.1	Ocena zdolności dyskryminacyjnych zmiennych ciągłych.	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.	5
2.3	Wyznaczanie składowych głównych.	6

Spis rysunków

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

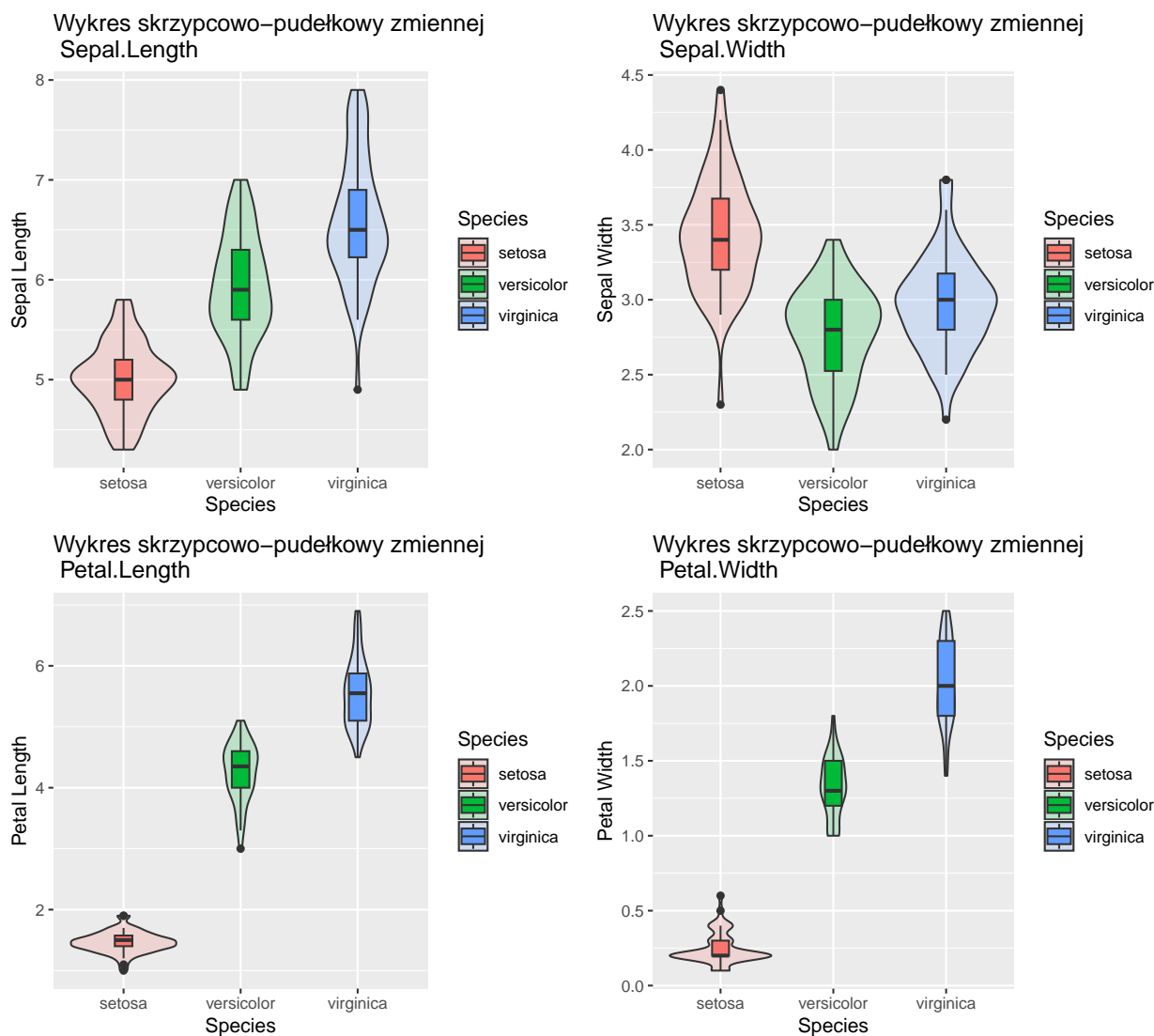
Spis tabel

1	Skuteczność wybranych metod dyskretyzacji dla zmiennej Sepal Width . . .	3
2	Skuteczność wybranych metod dyskretyzacji dla zmiennej Petal Width . . .	3

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 zagregowaną klasą. Tę liczbę podzieliliśmy przez liczbę wszystkich przypadków, aby uzyskać procent zgodności danej metody dyskretyzacji. Porównanie różnych metod przedziałowania zostały przedstawione poniżej

Tabela 1: Skuteczność wybranych metod dyskretyzacji dla zmiennej *Sepal.Width*

Przedziałowanie według równej częstości	Przedziałowanie według równej szerokości	Dyskretyzacja oparta na algorytmie K-średnich	Stałe granice przedziału
79.33	93.33	72	84.67

Tabela 2: Skuteczność wybranych metod dyskretyzacji dla zmiennej *Petal.Width*

Przedziałowanie według równej częstości	Przedziałowanie według równej szerokości	Dyskretyzacja oparta na algorytmie K-średnich	Stałe granice przedziału
97.33	100	100	86

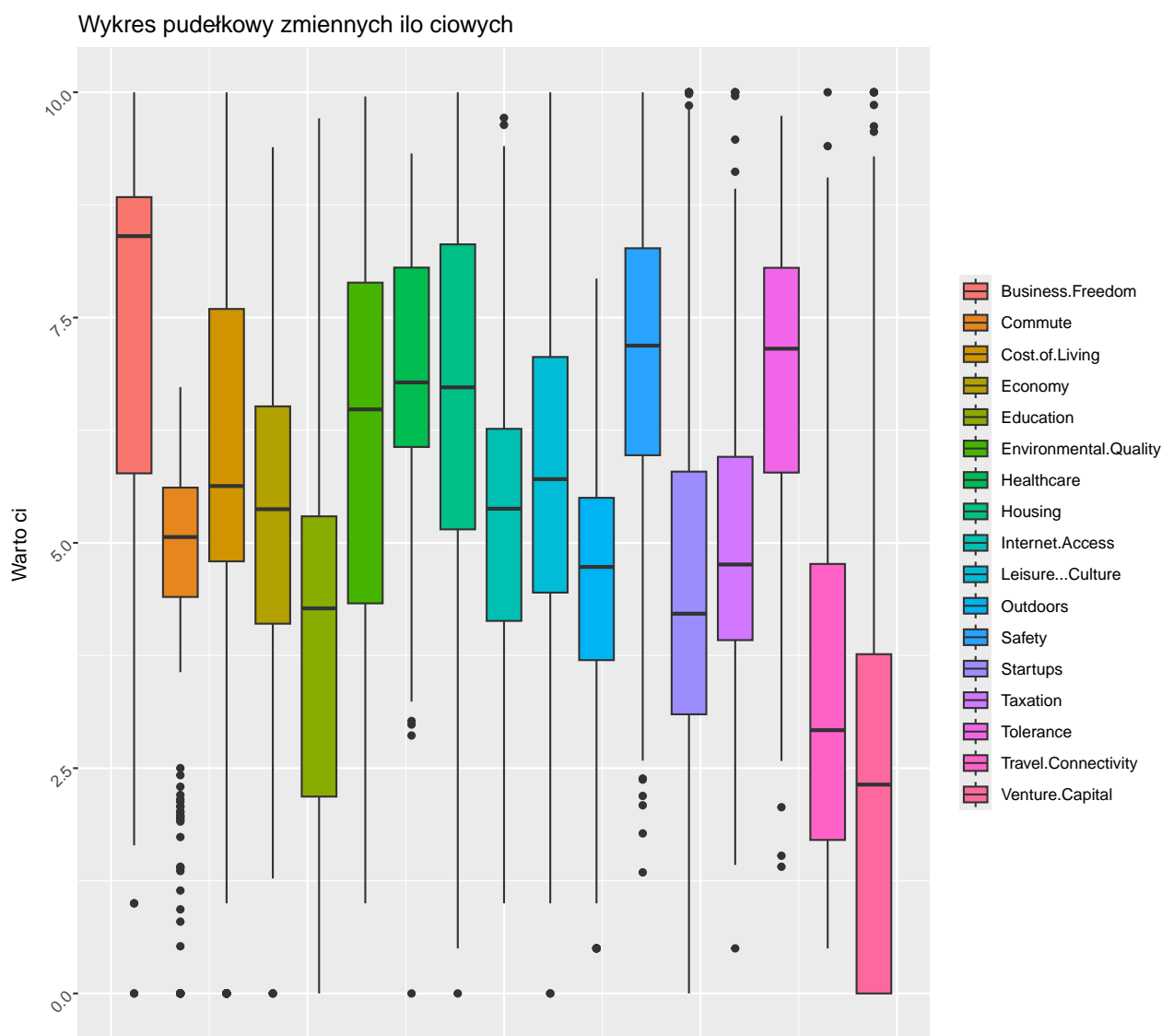
1.2.2 Wnioski dotyczące skuteczności przedziałowania

Z tabel ?? oraz ?? 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.

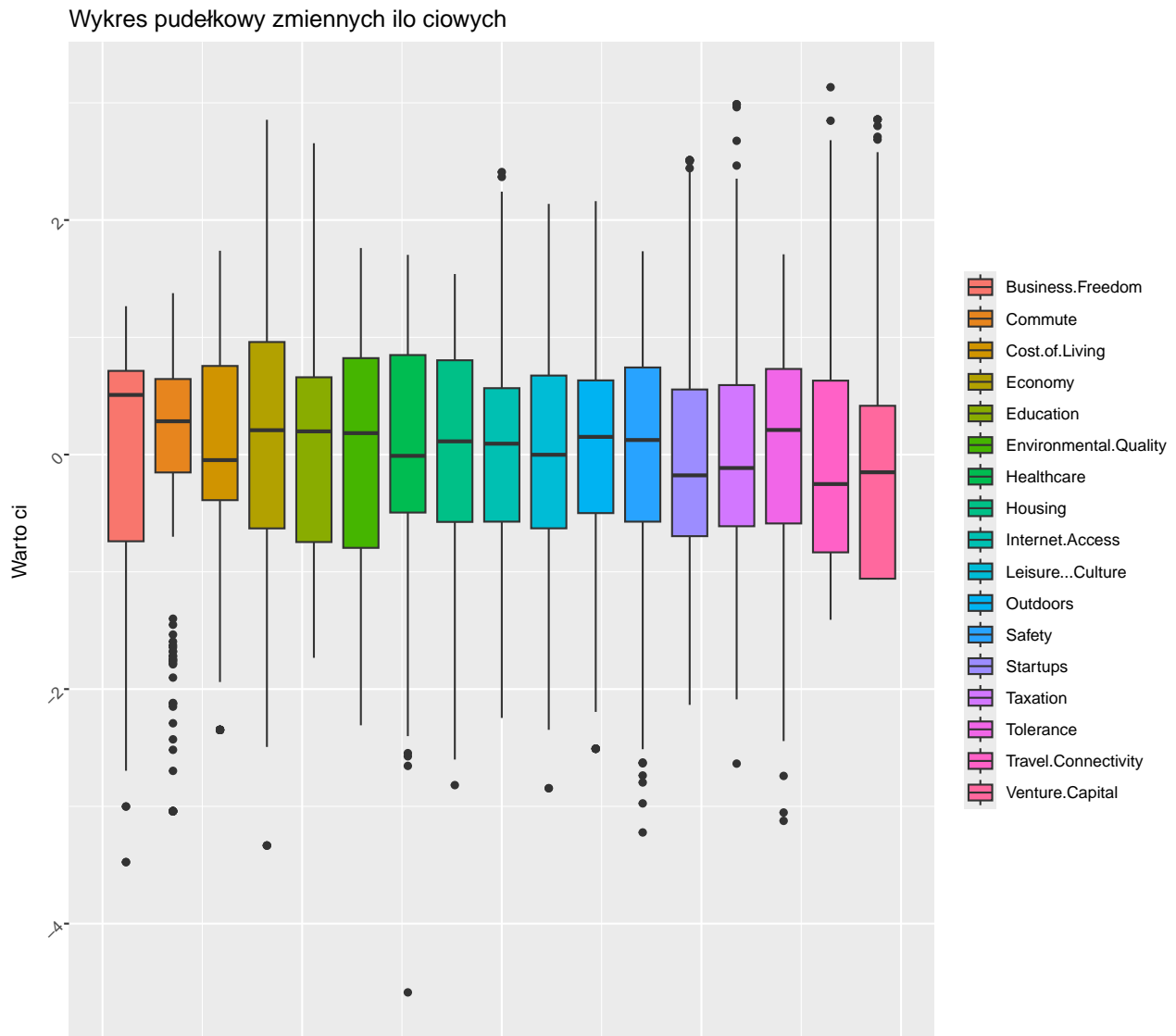
W celu porównania wariancji wszystkich zmiennych ilościowych ze zbioru *uaScoresDataFrame*, posłużymy się wykresami pudełkowymi.



Rysunek 2: Wykresy pudełkowe zmiennych ciągłych przed zastosowaniem standaryzacją

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.

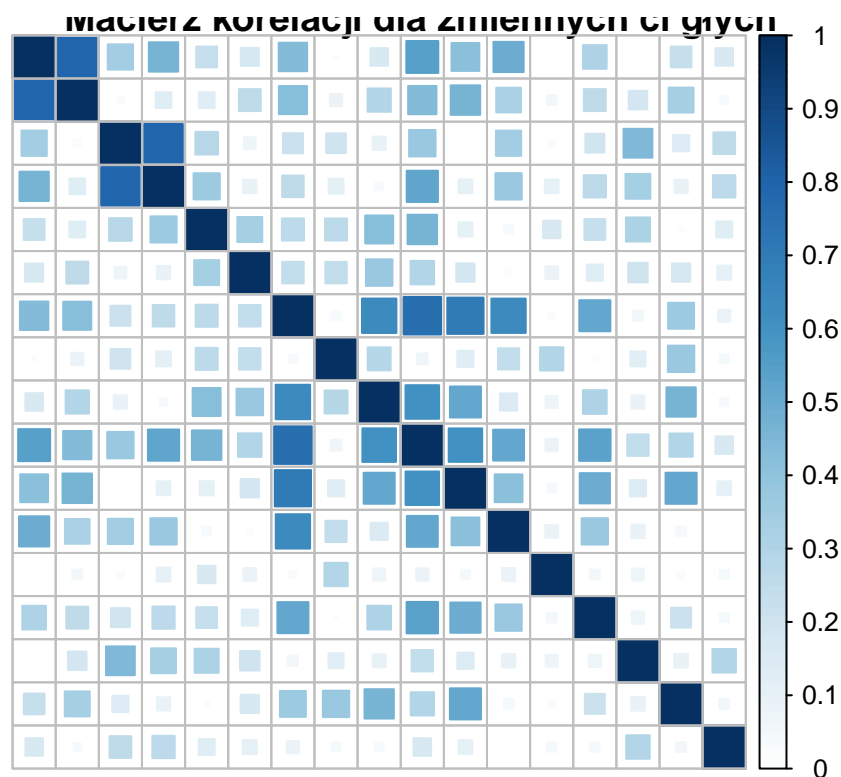
Na poniższym wykresie 3 pudełkowym widoczne są efekty standaryzacji zastosowane dla zmiennych ze zbioru danych.



Rysunek 3: Wykresy pudełkowe 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 $[0;1]$.

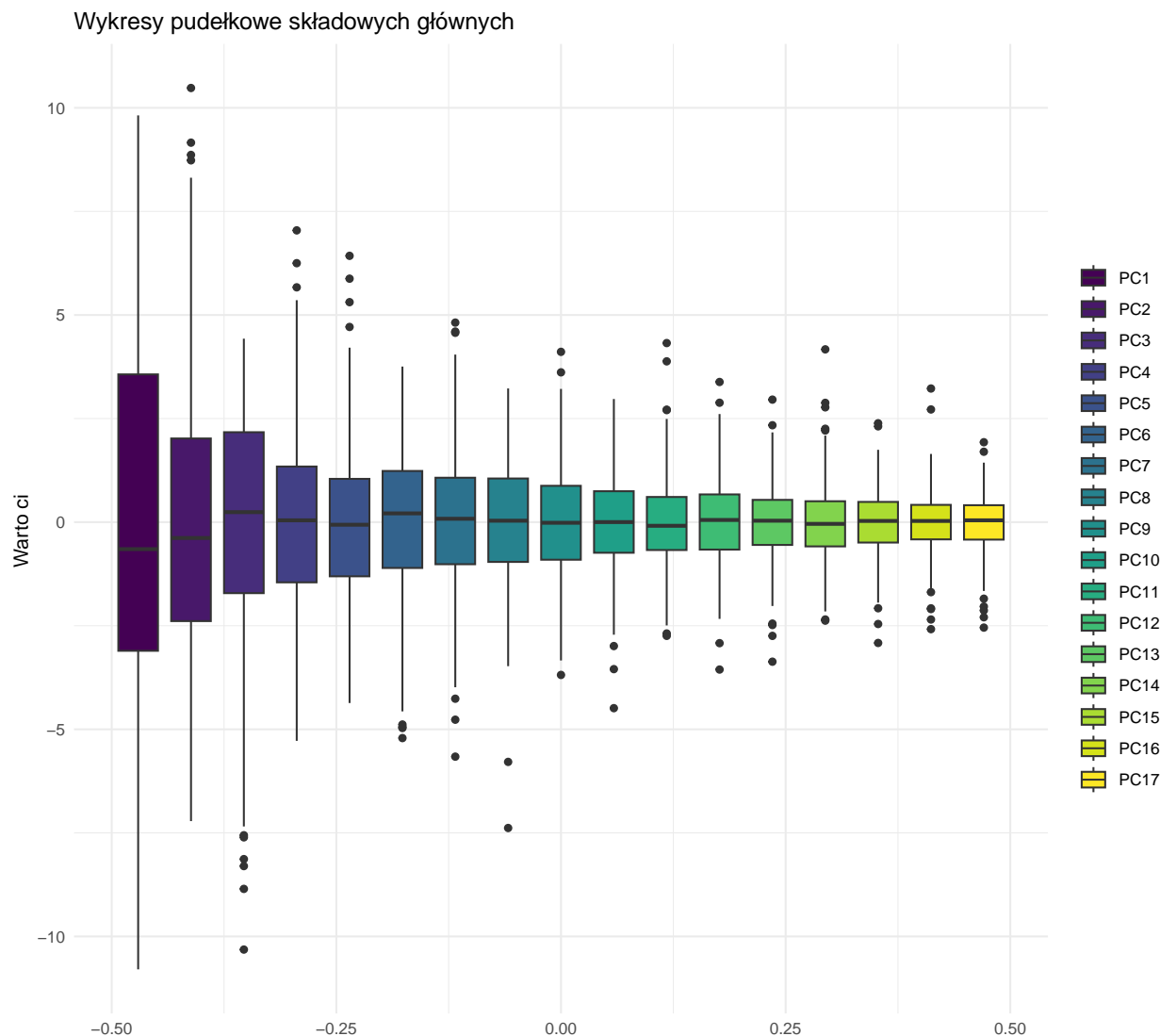


Rysunek 4: Macierz korelacji dla zmiennych ciągłych

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 oczekiwać, że redukcja wymiarowości będzie wymagała uwzględnienia relatywnie dużej liczby składowych głównych, aby osiągnąć zakładaną frakcję wyjaśnianej wariancji.

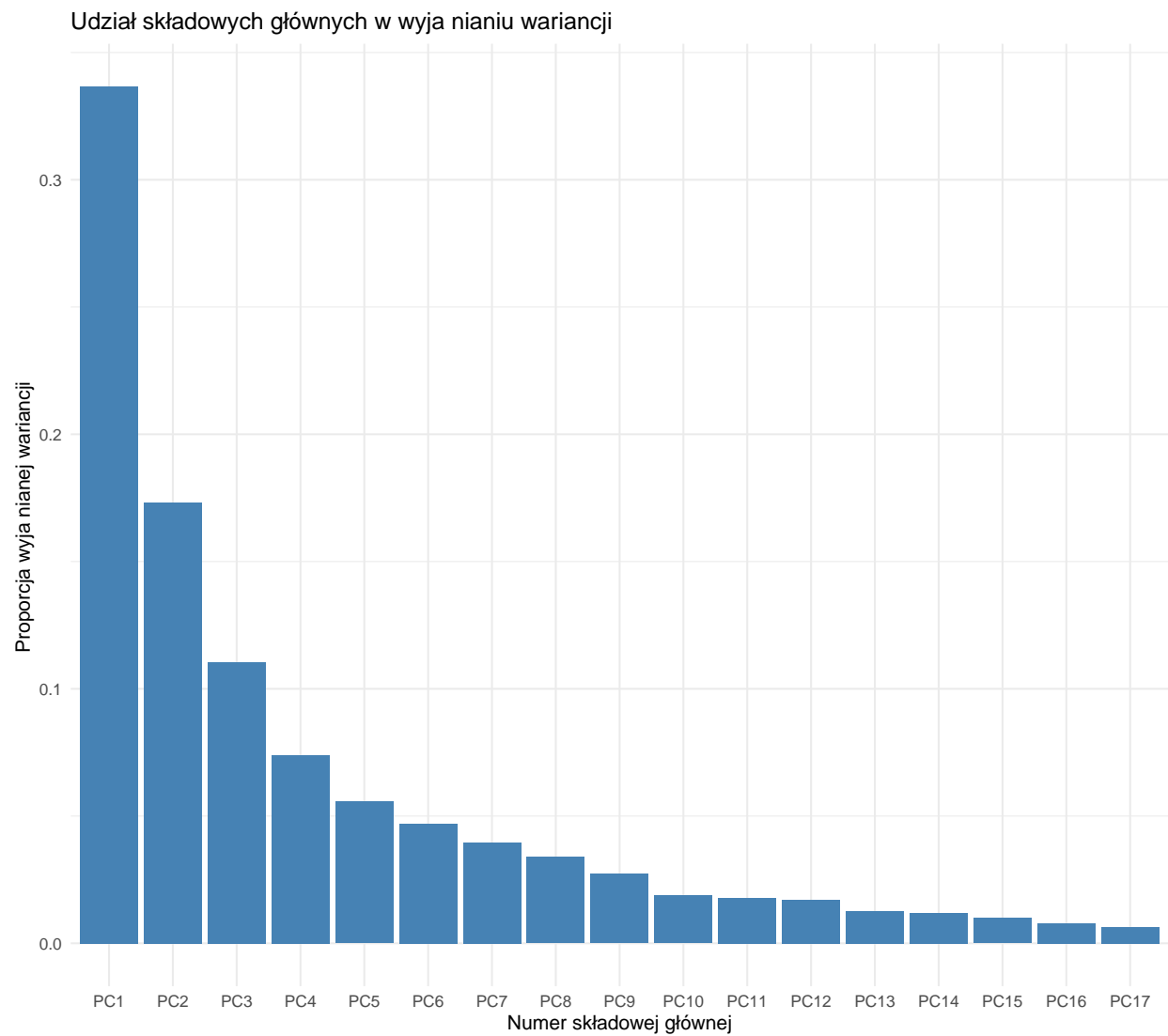
2.3 Wyznaczanie składowych głównych.

Dla analizowanego zbioru zmiennych ciągłych 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.

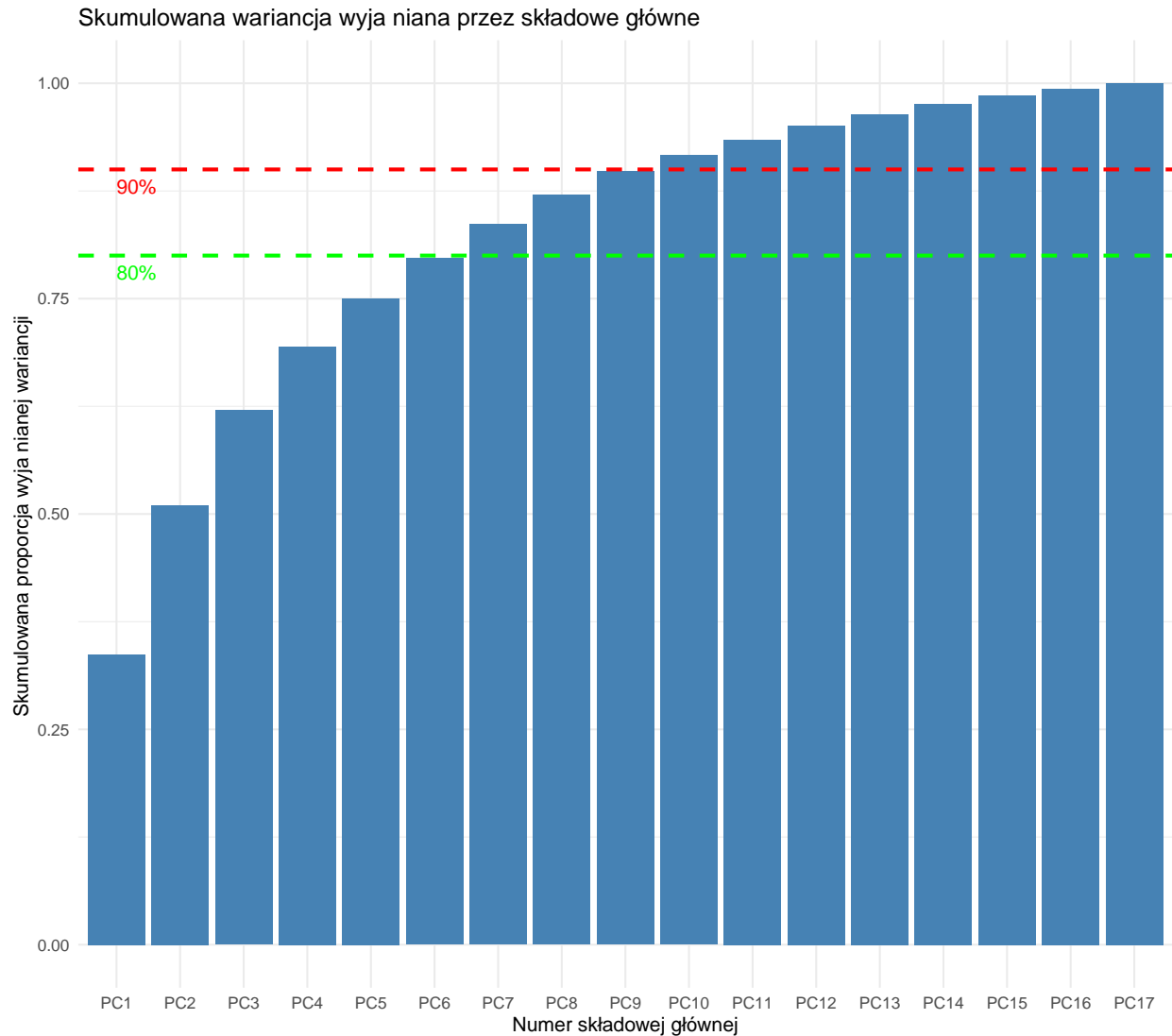


Rysunek 5: Wykresy pudełkowe dla składowych głównych

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 satysfakcjonującego poziomu odwzorowania struktury danych.



Rysunek 6: Porównanie udziału wariancji wyjaśnianej przez poszczególne 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 dziesięciu składowych.