Sprawozdanie z listy 2

Eksploracja danych

Marta Stankiewicz (282244)
Paweł Nowak (282223)

2025-04-15

Spis treści

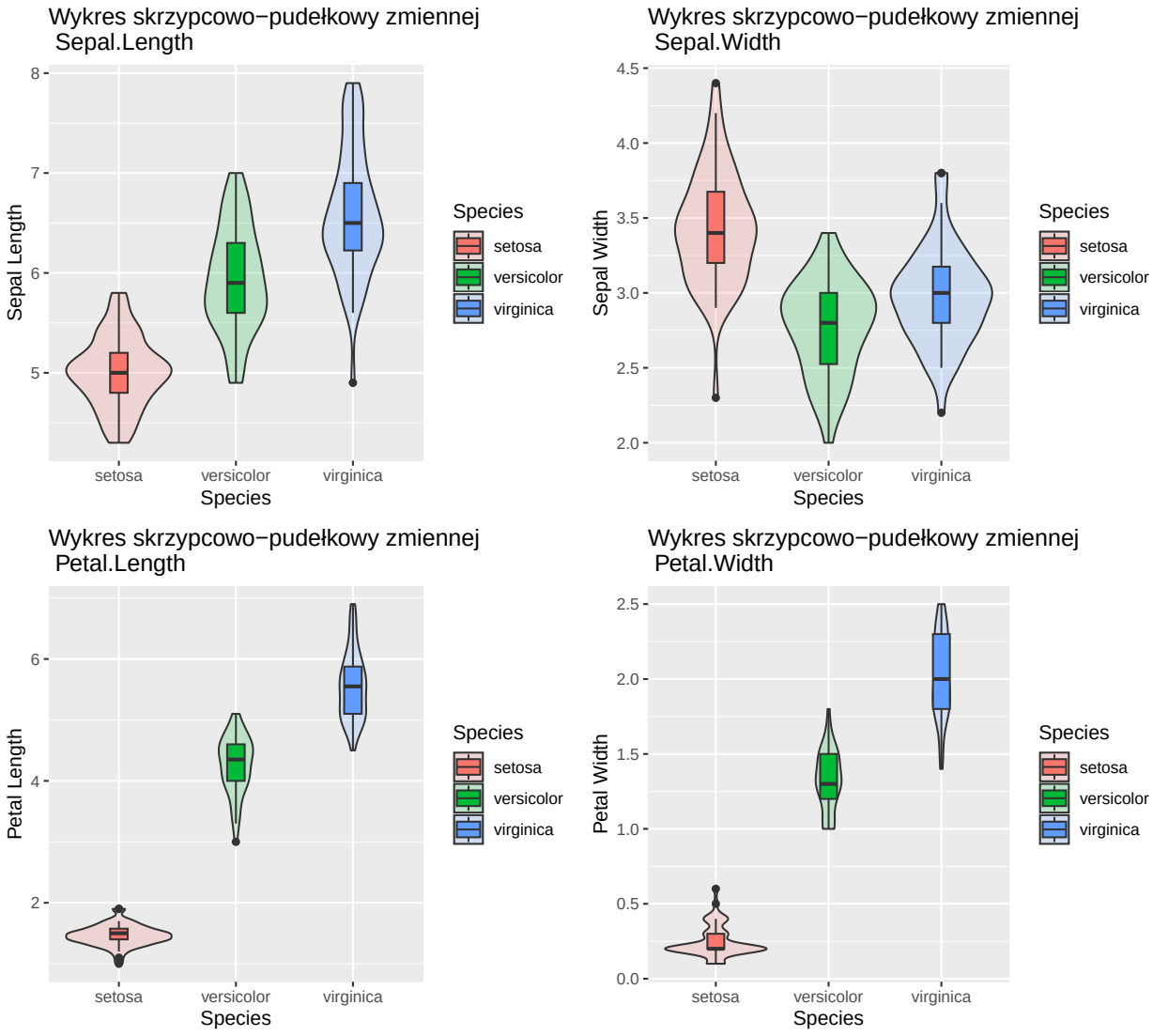
Spis rysunków

Spis tabel

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 róznych 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 róznych metod przedziałowania zostały przedstawione poniżej

Tabela 1: Skuteczność wybranych metod dyskretyzacji dla zmiennej Sepal Width

Przedziałowanie według równej częstotliwość	Przedziałowanie według równej szerokości	Dyskretyzacja oparta na algorytmie K-średnich	Stałe granice przedziału
91.33	81.33	90.67	72.67

Tabela 2: Skuteczność wybranych metod dyskretyzacji dla zmiennej Petal Width

Przedziałowanie według równej częstotliwość	Przedziałowanie według równej szerokości	Dyskretyzacja oparta na algorytmie K-średnich	Stałe granice przedziału
97.33	100	98.67	86

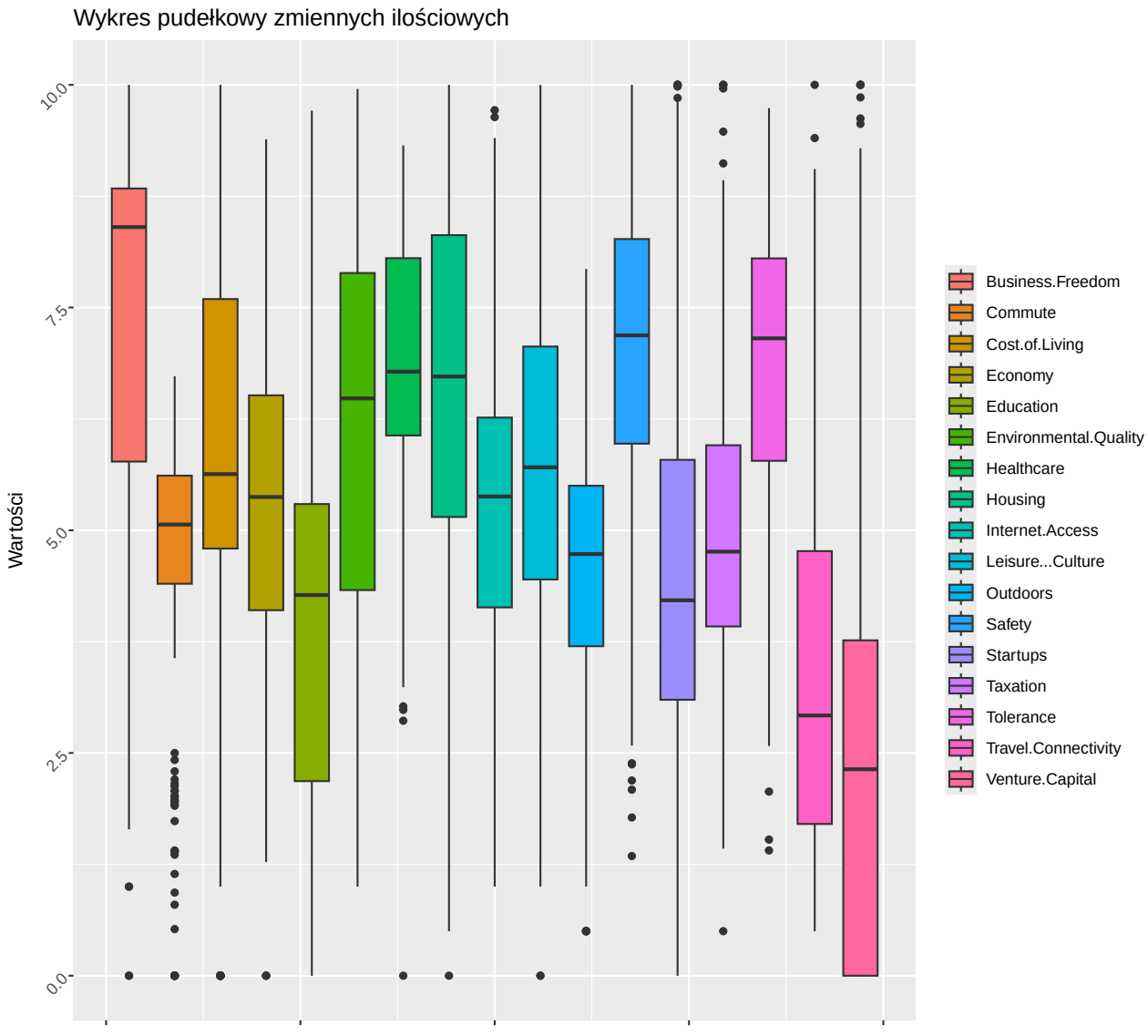
1.2.2 Wnioski dotyczące skuteczności przedziałowania

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* znaczącą rożnią 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.

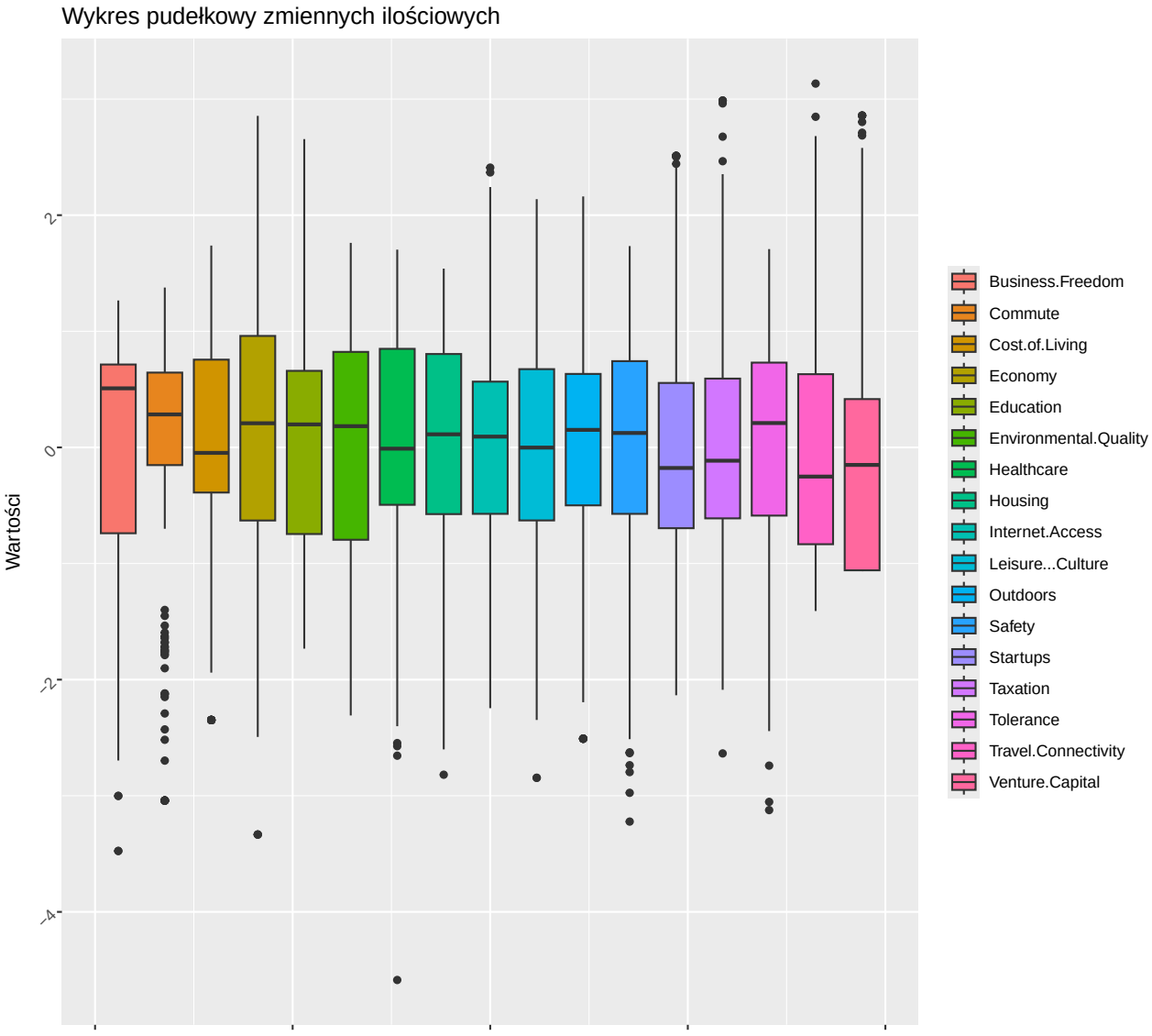
W celu porównania wariancji wszystkich zmiennych ilościowych ze zbioru *uaScoresDataFrame*, posłużymy się wykresami pudełkowymi.



Rysunek 2: Wykresy pudelkowe 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 brakuję atrybutów o wysokiej wariancji. W celu ujednolicenia wariancji zmiennych, konieczne będzie zastosowanie standaryzacji.

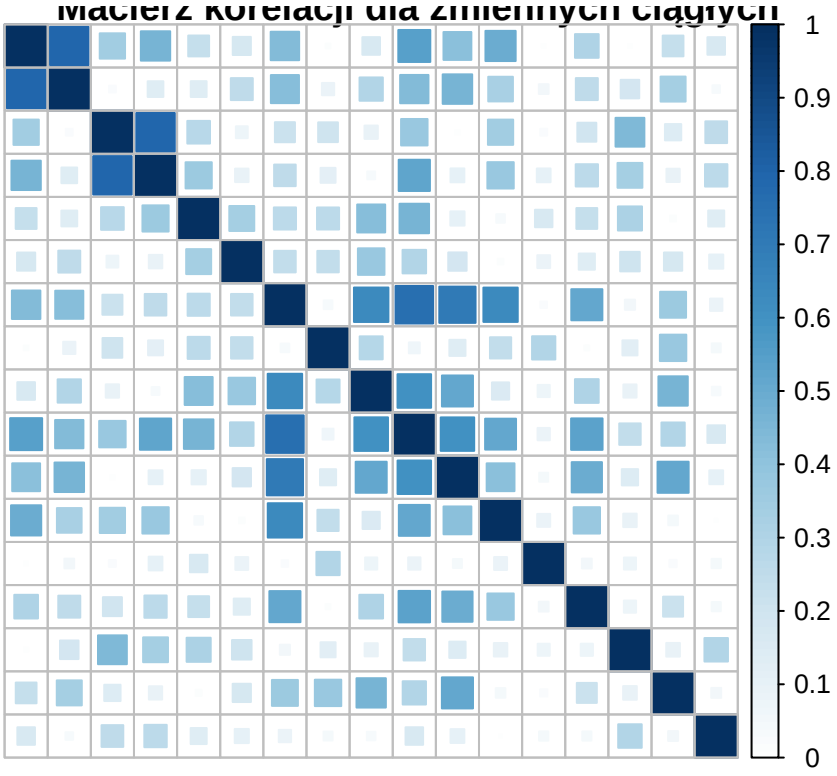
Na poniższym wykresie 3 pudełkowym widoczne są efekty standaryzacji zastosowane dla zmiennych ze zbioru danych.



Rysunek 3: Wykresy pudelkowe 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 reduntantnych zmiennych. Taka redundancja może zostać wyelimonowana za pomocą analizy składowych głównych. Aby poprawić czytelność wykresu, nazwy zmiennych zostały pominięte, a wartość współczynnika została przeskalowana do przedzialu [0;1].

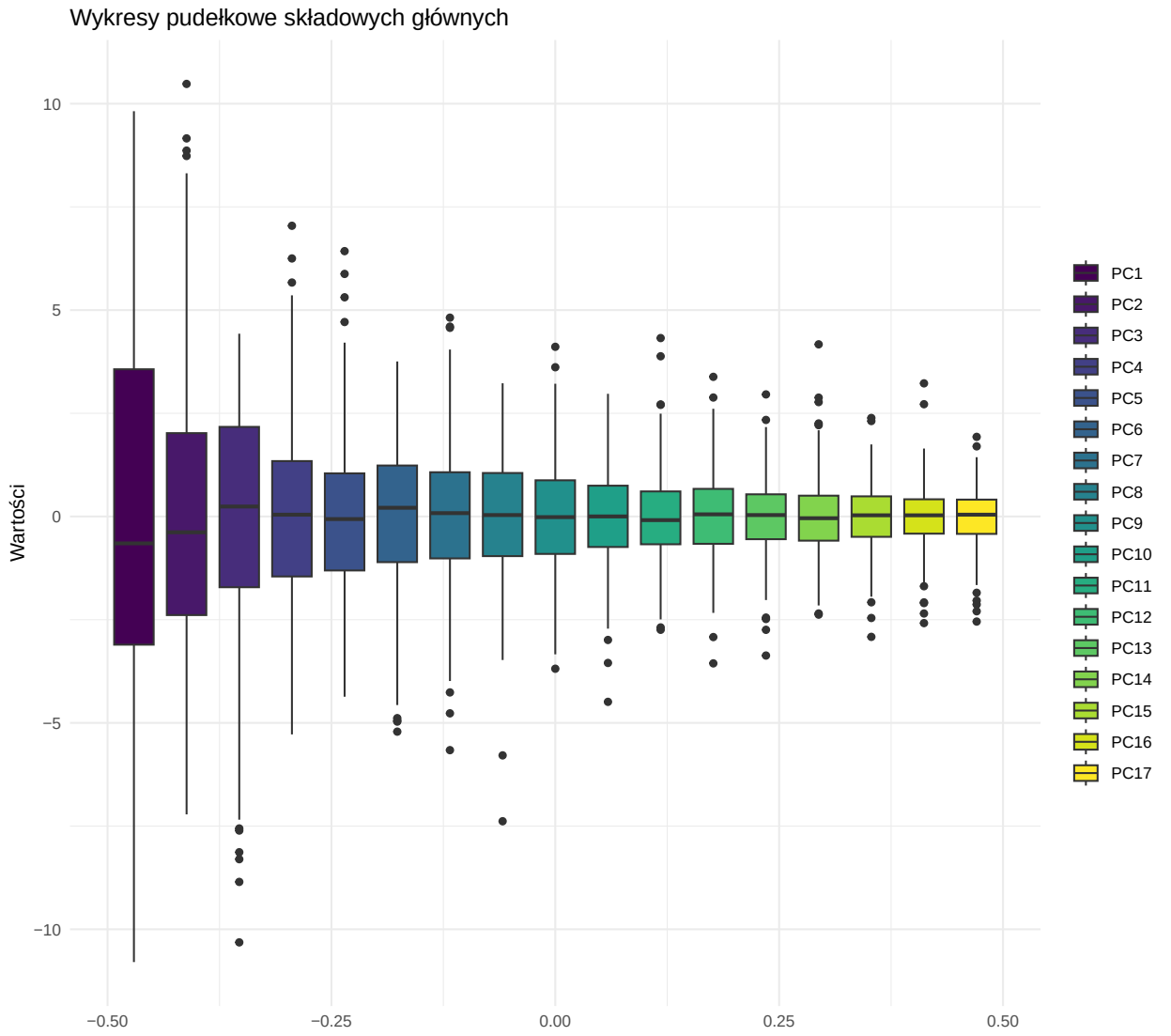


Rysunek 4: Macierz korelacji dla zmiennych ciągłych

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.

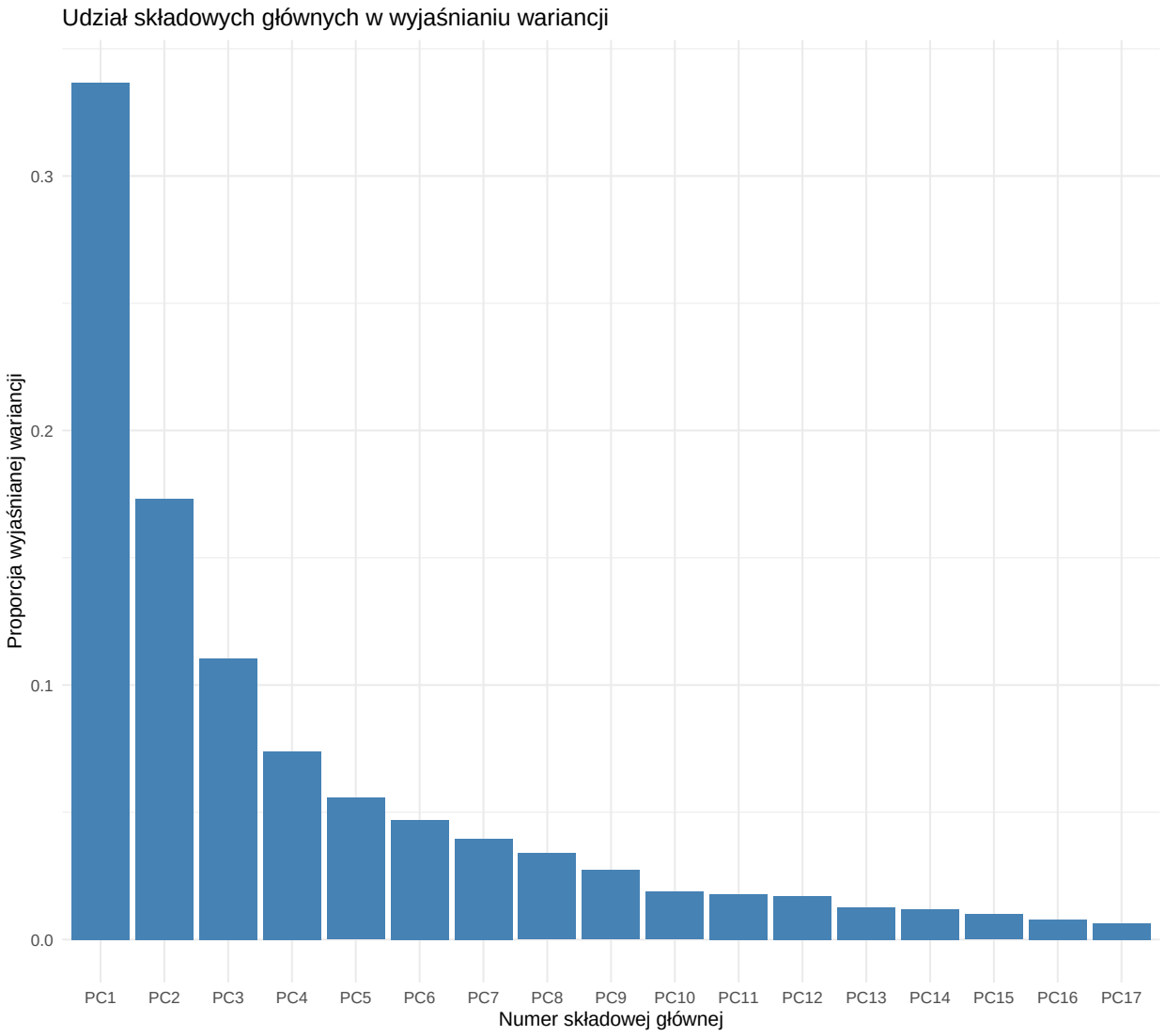
2.3 Wyznaczanie składowych głównych.

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.

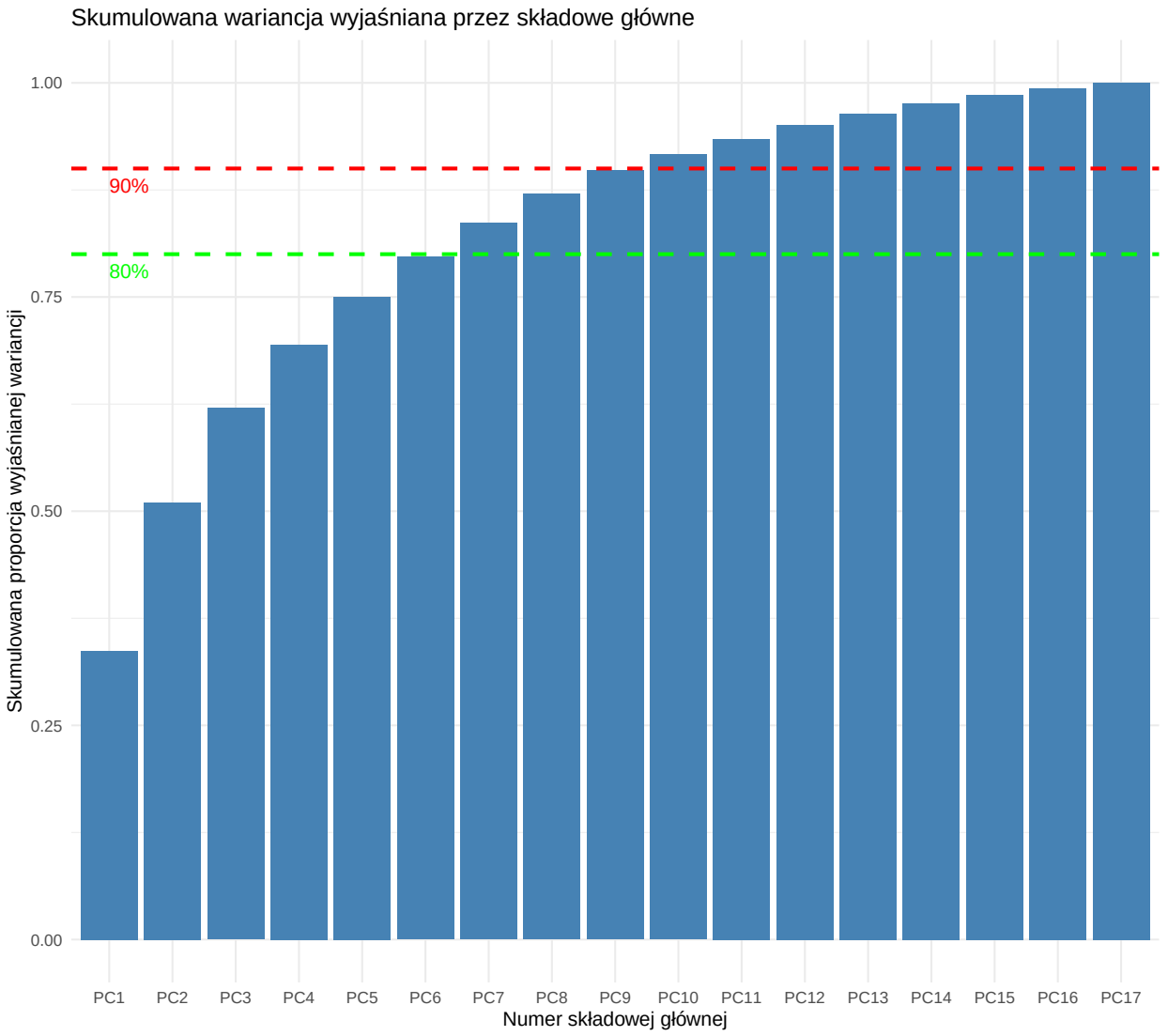


Rysunek 5: Wykresy pudełkowe dla składowych głównych

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 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 znacząco odbiegają od pozostałych pod względem analizowanych cech.

Rysunek 8: Wizualizacja pierwszych dwóch składowych głównych

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 metropolie 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.