9.1 (PCA) Dane w pliku *Miasta.txt* zawierają wartości trzech atrybutów dla 46 miast na świecie:

- *Work* - ważona średnia liczba godzin pracy dla 12 zawodów,
- *Price* - indeks kosztów utrzymania na podstawie cen 112 towarów i usług (wartość indeksu dla Zurichu równa się 100),
- *Salary* - indeks płacy za godzinę w 12 zawodach po odjęciu podatku (wartość indeksu dla Zurichu równa się 100).

(a) Dokonaj standaryzacji zmiennych.
(b) Dla pary zmiennych: Work i Price wyznacz kierunki wzdłuż których występuje największa zmienność. Przedstaw wykres rozproszenia dla pary zmiennych Work i Price a następnie wrysuj linie wzdłuż których występuje największa zmienność.
(c) Wykonaj analizę składowych głównych dla wszystkich zmiennych (standaryzowanych).
(d) Jaki jest procent wariancji tłumaczony przez poszczególne składowe? Czy możemy dokonać redukcji wymiaru danych?
(e) Oblicz kierunki główne oraz składowe główne.
(f) Znajdź miasto o największej wartości pierwszej składowej głównej. O czym świadczy duża wartość pierwszej składowej głównej?

9.2 (PCR) Dane *meatspec (faraway)* dotyczą predykcji zawartości tłuszczu (zmienna *fat*) na podstawie wartości spektrum fali odbitej (100 zmiennych V1–V100 będących absorbancjami dla częstotliwości, od niskich do wysokich). Celem analizy jest redukcja wymiaru predyktorów poprzez zastąpienie ich pewną liczbą składowych głównych.

(a) Podziel dane w następujący sposób: pierwsze 172 rekordy stanowią zbiór treningowy na którym będzie przeprowadzone dopasowanie modelu; rekordy 173–215 stanowią zbiór testowy na którym testujemy model. Dopasuj model liniowy opisujący zależność zmiennej *fat* od pozostałych zmiennych na zbiorze treningowym. Oblicz wartość

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

dla predykcji zawartości tłuszczu dla zbioru testowego na podstawie zbioru treningowego.
(b) Stosując metodę krokowego wyboru zmiennych ("backward") z kryterium BIC wybierz zmienne na podstawie zbioru treningowego i oblicz RMSE dla wybranego modelu.
(c) Dokonaj standaryzacji zmiennych objaśniających. Wykonaj analizę składowych głównych dla standaryzowanych zmiennych objaśniających.
(d) Podaj interpretację pierwszej składowej głównej.
(e) Na podstawie wykresu odchyleń standardowych dla składowych głównych wybierz składowe główne. Na zbiorze treningowym dopasuj model (PCR) zależności zmiennej *fat* od wybranych składowych głównych. Oblicz RMSE.
(f) Podziel dane na 3 podzbiory: oberwacje 1-130 (zbiór treningowy), 131-172 (zbiór walidacyjny), 173-215 (zbiór testowy). Dokonajć wyboru składowych głównych minimalizując RMSE na zbiorze walidacyjnym. Oblicz RMSE dla wybranego modelu na zbiorze testowym.