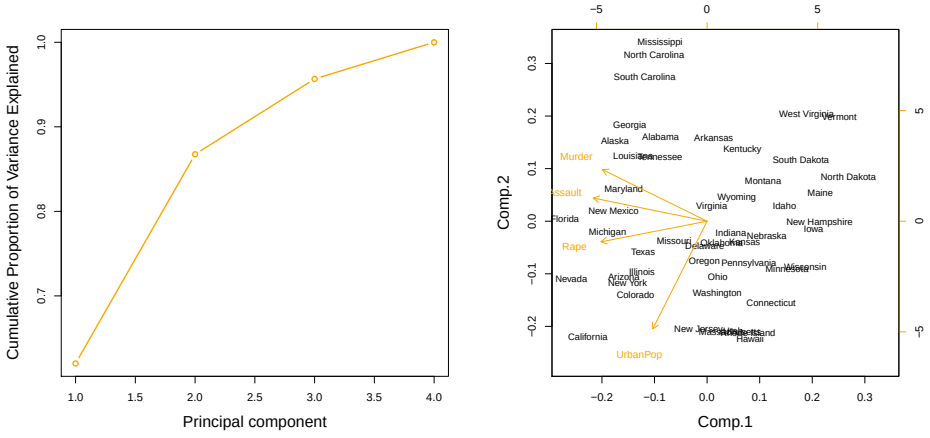
Analiza składowych głównych (Principal Component Analysis)

Zadania:

1. Własna implementacja metody PCA w R.
   - Wczytaj dane `USArrest`.
   - Dokonaj standaryzacji i centrowania danych (funkcja `scale`).
   - Wyznacz macierz kowariancji (funkcja `cov`).
   - Wyznacz rozkład spektralny macierzy kowariancji (funkcja `eigen`).
   - Wyznacz wektory i wartości własne macierzy kowariancji.
   - Oblicz składowe główne.
   - Porównaj wyznaczone wektory ładunków i składowe główne z otrzymanymi przy użyciu funkcji `princomp`, na danych `USArrest`.
2. Analiza danych `USArrest`.
   - Wykonaj analizę składowych głównych na danych `USArrest`.
   - Sprawdź jaki procent zmienności danych tłumaczą poszczególne składowe; wykonaj wykres pokazujący skumulowane proporcje wariancji dla poszczególnych składowych głównych, Rysunek 1 (a).
   - Dokonaj interpretacji dwóch pierwszych składowych głównych, wykonaj wykres rozproszenia dla dwóch pierwszych składowych (funkcja `biplot`), Rysunek 1 (b).



Rysunek 1: (a) Wykres skumulowanych proporcji wariancji. (b) Wykres rozproszenia dla dwóch pierwszych składowych, dla danych `USArrests`.

3. Analiza danych `Hitters` (z pakietu `ISLR`). Dane dotyczą problemu regresji w którym zmienna *Salary* jest zmienną odpowiedzi, a pozostałe zmienne traktujemy jako objaśniające.

   - Usuń wiersze w których znajdują się braki danych (funkcja `na.omit`).
   - Wykonaj analizę składowych głównych na danych `Hitters`.
   - Dopasuj model liniowy w którym *Salary* jest zmienną odpowiedzi, a składowe główne są zmiennymi objaśniającymi. Dokonaj oceny istotności zmiennych objaśniających.
   - Dopasuj zagnieżdżoną rodzinę modeli liniowych, zbudowanych na pierwszych $k = 1, 2, \ldots$ składowych głównych. Wykonaj wykres który pokazuje jak R^2 (współczynnik determinacji) zależy od liczby składowych głównych.
   - Dopasuj zagnieżdżoną rodzinę modeli liniowych, zbudowanych na $k = 1, 2, \ldots$ najbardziej istotnych zmiennych. Istotność zmiennej oceń na podstawie p-wartości statystyki t obliczonej dla modelu dopasowanego na wszystkich zmiennych. Wykonaj wykres który pokazuje jak R^2 zależy od liczby najbardziej istotnych zmiennych.
   - Porównaj wykresy otrzymane w powyższych punktach (celem eksperymentu jest pokazanie że model zbudowany na niewielkiej liczbie składowych głównych jest lepiej dopasowany niż model zbudowany na takiej samej liczbie oryginalnych zmiennych).