13 Analiza składowych głównych

13.1 Przykład

Przykład. Zbiór danych `USArrests` zawiera informacje dotyczące liczby morderstw, napadów, gwałtów przypadających na 100,000 osób w poszczególnych stanach USA w roku 1973 oraz procent ludności mieszkającej w miastach. Chcielibyśmy się dowiedzieć, czy stany są do siebie w pewien sposób zbliżone oraz spróbować zwizualizować je na płaszczyźnie.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
dim(USArrests)
```

```
## [1] 50  4
```

- przygotowanie danych do analizy składowych głównych

```
# sprawdzamy czy wariancje (,,zmienności'') zmiennych są bardzo zróżnicowane
var(USArrests)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder    18.970465  291.0624   4.386204  22.99141
## Assault  291.062367 6945.1657 312.275102 519.26906
## UrbanPop   4.386204  312.2751 209.518776  55.76808
## Rape      22.991412  519.2691  55.768082  87.72916
```

```
# tak są, więc centrujemy i skalujemy funkcją scale
USArrests_scale <- scale(USArrests)
var(USArrests_scale)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

- model analizy składowych głównych w R i procent wyjaśnianej wariancji zmiennych oryginalnych przez poszczególne składowe główne

```
pca <- prcomp(USArrests, scale = TRUE)
# lub
# pca <- prcomp(USArrests_scale)
pca
```

```
## Standard deviations (1, .., p=4):
## [1] 1.5748783 0.9948694 0.5971291 0.4164494
##
## Rotation (n x k) = (4 x 4):
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

```
# bez skalowania
prcomp(USArrests)
```

```
## Standard deviations (1, .., p=4):
## [1] 83.732400 14.212402  6.489426  2.482790
##
## Rotation (n x k) = (4 x 4):
##                 PC1         PC2         PC3         PC4
## Murder   0.04170432 -0.04482166  0.07989066 -0.99492173
## Assault  0.99522128 -0.05876003 -0.06756974  0.03893830
## UrbanPop 0.04633575  0.97685748 -0.20054629 -0.05816914
## Rape     0.07515550  0.20071807  0.97408059  0.07232502
```

```
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3     PC4
## Standard deviation     1.5749 0.9949 0.59713 0.41645
## Proportion of Variance 0.6201 0.2474 0.08914 0.04336
## Cumulative Proportion  0.6201 0.8675 0.95664 1.00000
```

- wyniki (ang. *scores*) - współrzędne obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne (to one najczęściej podlegają wizualizacji)

```
head(pca$x)
```

```
##                  PC1        PC2         PC3          PC4
## Alabama    -0.9756604  1.1220012 -0.43980366  0.154696581
## Alaska     -1.9305379  1.0624269  2.01950027 -0.434175454
## Arizona    -1.7454429 -0.7384595  0.05423025 -0.826264240
## Arkansas    0.1399989  1.1085423  0.11342217 -0.180973554
## California -2.4986128 -1.5274267  0.59254100 -0.338559240
## Colorado   -1.4993407 -0.9776297  1.08400162  0.001450164
```

- ładunki (ang. *loadings*) - współczynniki pokazujące wkład poszczególnych zmiennych bazowych w tworzenie składowych głównych (im wartość bezwzględna z ładunku jest większa, tym zmienna ma większy wkład w budowę składowej głównej)

```
pca$rotation
```

```
##                 PC1        PC2        PC3         PC4
## Murder   -0.5358995  0.4181809 -0.3412327  0.64922780
## Assault  -0.5831836  0.1879856 -0.2681484 -0.74340748
## UrbanPop -0.2781909 -0.8728062 -0.3780158  0.13387773
## Rape     -0.5434321 -0.1673186  0.8177779  0.08902432
```

- wykres osypiska (piargowy, ang. *scree plot*) - wykres wariancji poszczególnych składowych głównych

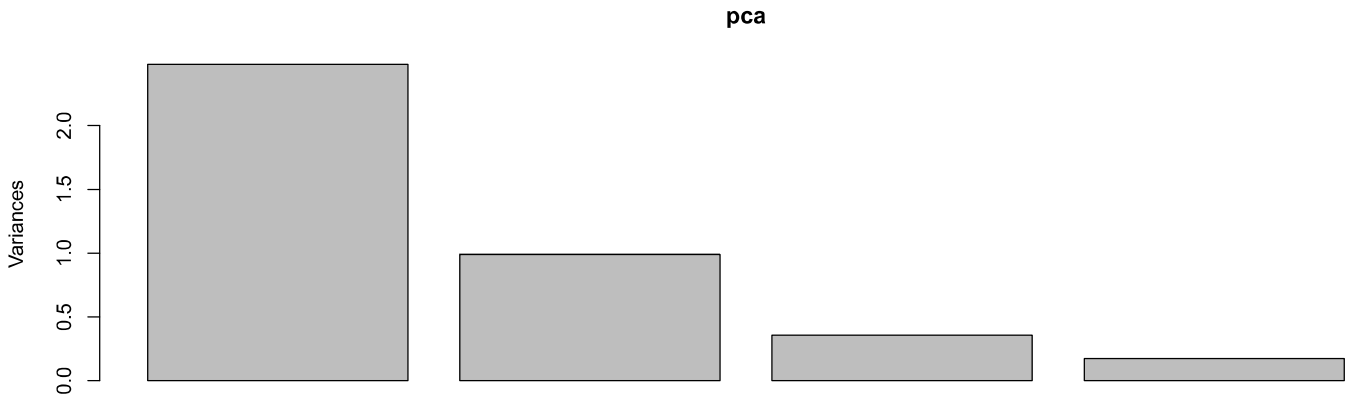
```
pca$sdev^2
```

```
## [1] 2.4802416 0.9897652 0.3565632 0.1734301
```

```
apply(pca$x, 2, var)
```

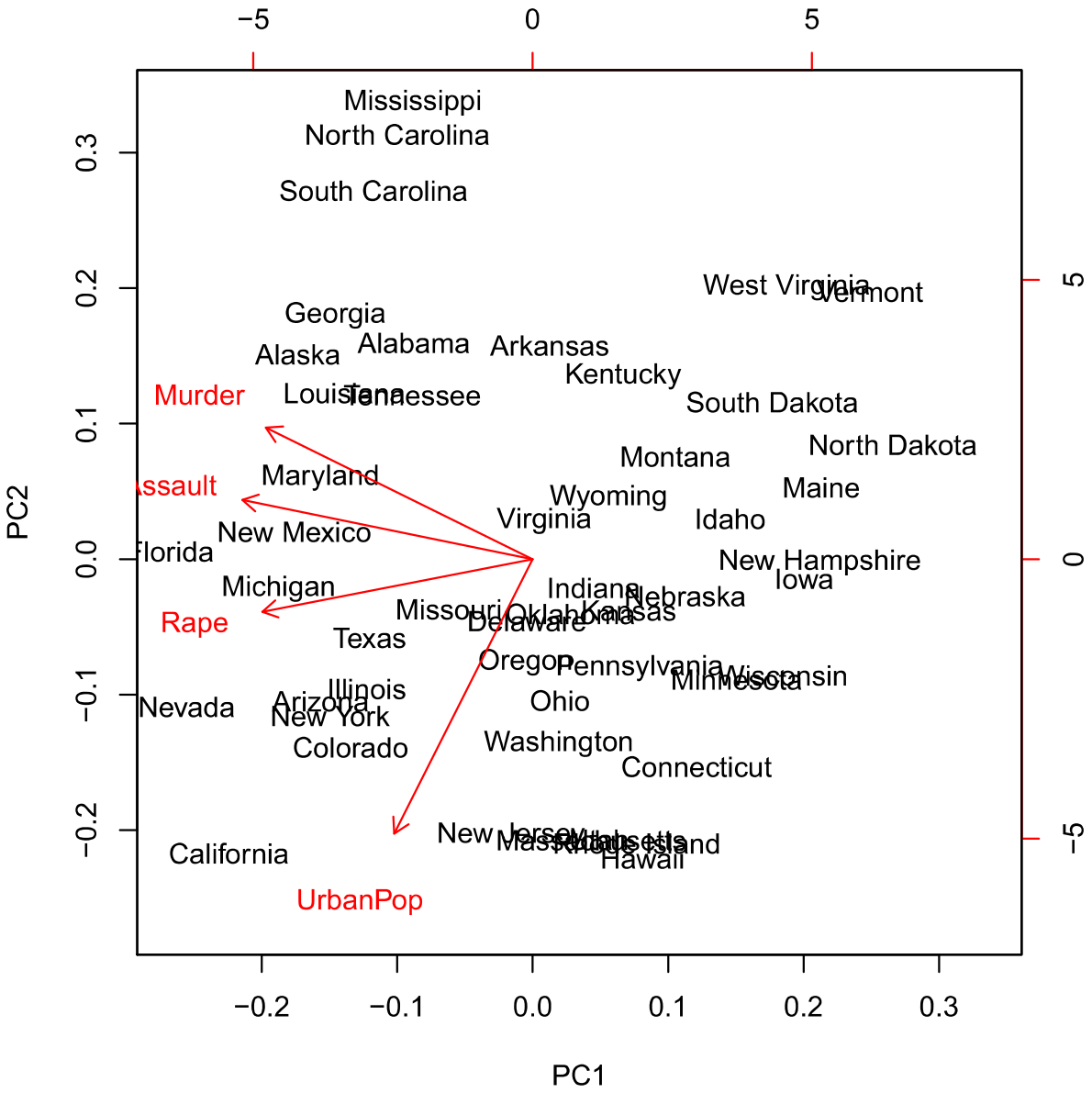
```
##       PC1       PC2       PC3       PC4
## 2.4802416 0.9897652 0.3565632 0.1734301
```

```
plot(pca)
```



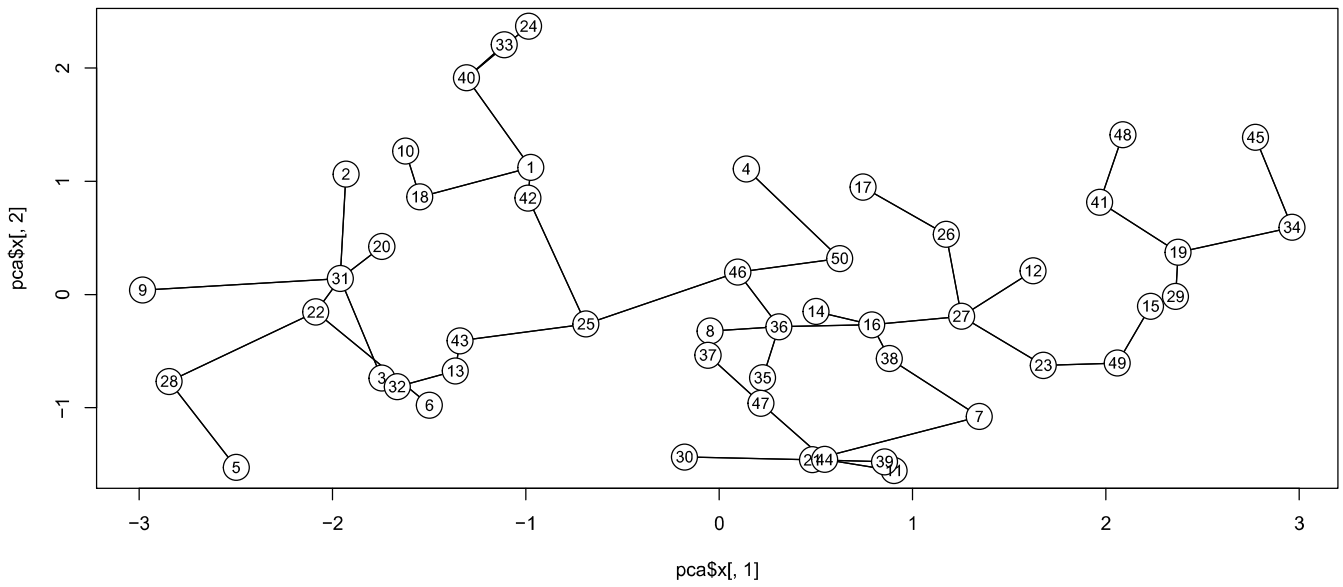
- biplot (ang. *biplot*) - wykres, na którym punkty przedstawiają poszczególne obserwacje w nowym układzie dwóch pierwszych składowych głównych, a strzałki oznaczają zmienne. Kierunek strzałek pokazuje wpływ tych zmiennych odpowiednio na pierwszą i drugą składową główną. Kąt przecięcia strzałek jest proporcjonalny do zależności pomiędzy zmiennymi, a ich długość odzwierciedla odchylenie standardowe.

```
biplot(pca)
```



- Żeby stwierdzić, czy taki wykres jest adekwatnym odzwierciedleniem położenia oryginalnych punktów, można na niego nanieść minimalne drzewo rozpinające (MST). MST to graf, którego wierzchołkami są obserwacje, dwa punkty połączone są dokładnie jedną ścieżką, a suma krawędzi jest minimalna. Punkty połączone krawędziami powinny być blisko siebie na wykresie.

```
library(ape)
plot(mst(dist(USArrests_scale)), x1 = pca$x[, 1], x2 = pca$x[, 2])
```




```
# odczytywanie nazw obserwacji
row.names(USArrests_scale[c(24, 33),])
```

```
## [1] "Mississippi"    "North Carolina"
```

13.2 Zadania

Zadanie 1. W powyższym przykładzie do analizy składowych głównych zostały wykorzystane wszystkie zmienne. Jednak jedna z nich jest bardzo słabo skorelowana z pozostałymi. Ustal tę zmienną, a następnie wykonaj poniższe polecenia bez jej uzwględnienia:

1. Dokonaj analizy składowych głównych.

```
## Standard deviations (1, .., p=3):
## [1] 1.5357670 0.6767949 0.4282154
##
## Rotation (n x k) = (3 x 3):
##                PC1        PC2        PC3
## Murder  -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape    -0.5393836 -0.8179779 -0.1999436
```

2. Jaki procent wariancji tłumaczony jest przez poszczególne składowe?

```
## Importance of components:
##                            PC1    PC2     PC3
## Standard deviation     1.5358 0.6768 0.42822
## Proportion of Variance 0.7862 0.1527 0.06112
## Cumulative Proportion  0.7862 0.9389 1.00000
```

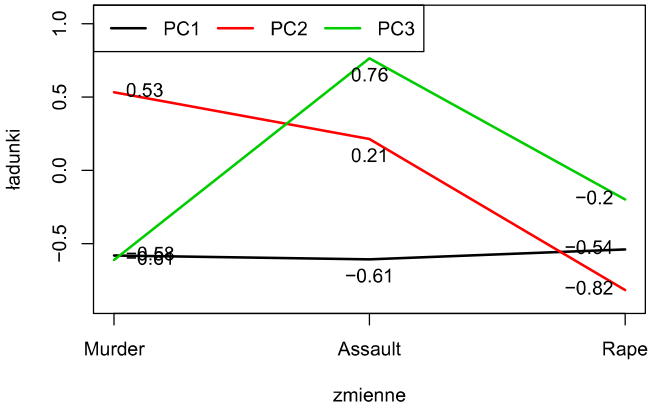
3. Wyznacz współrzędne obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne.

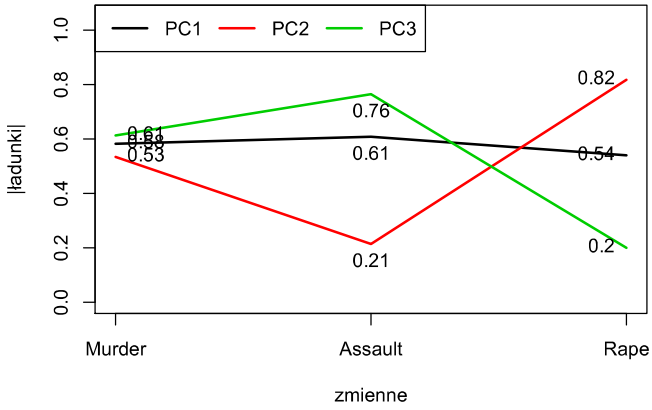
```
##                   PC1        PC2         PC3
## Alabama    -1.1980278  0.8338118 -0.16217848
## Alaska     -2.3087473 -1.5239622  0.03833574
## Arizona    -1.5033307 -0.4983038  0.87822311
## Arkansas   -0.1759894  0.3247326  0.07111174
```

```
## California -2.0452358 -1.2725770  0.38153933
## Colorado   -1.2634133 -1.4264063 -0.08369314
## ...
```

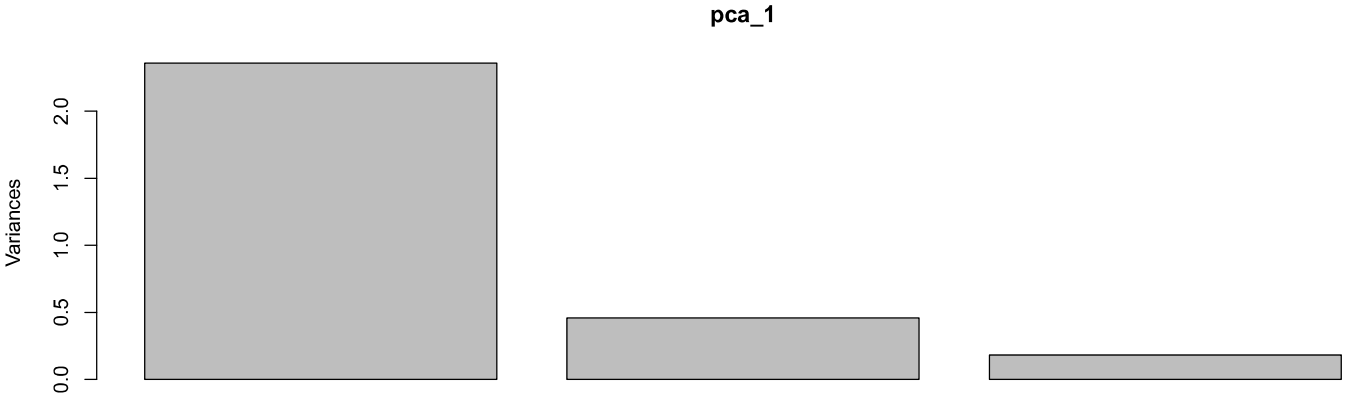
4. Dokonaj interpretacji ładunków i zilustruj je na wykresie.

```
##                PC1        PC2        PC3
## Murder  -0.5826006  0.5339532 -0.6127565
## Assault -0.6079818  0.2140236  0.7645600
## Rape    -0.5393836 -0.8179779 -0.1999436
```



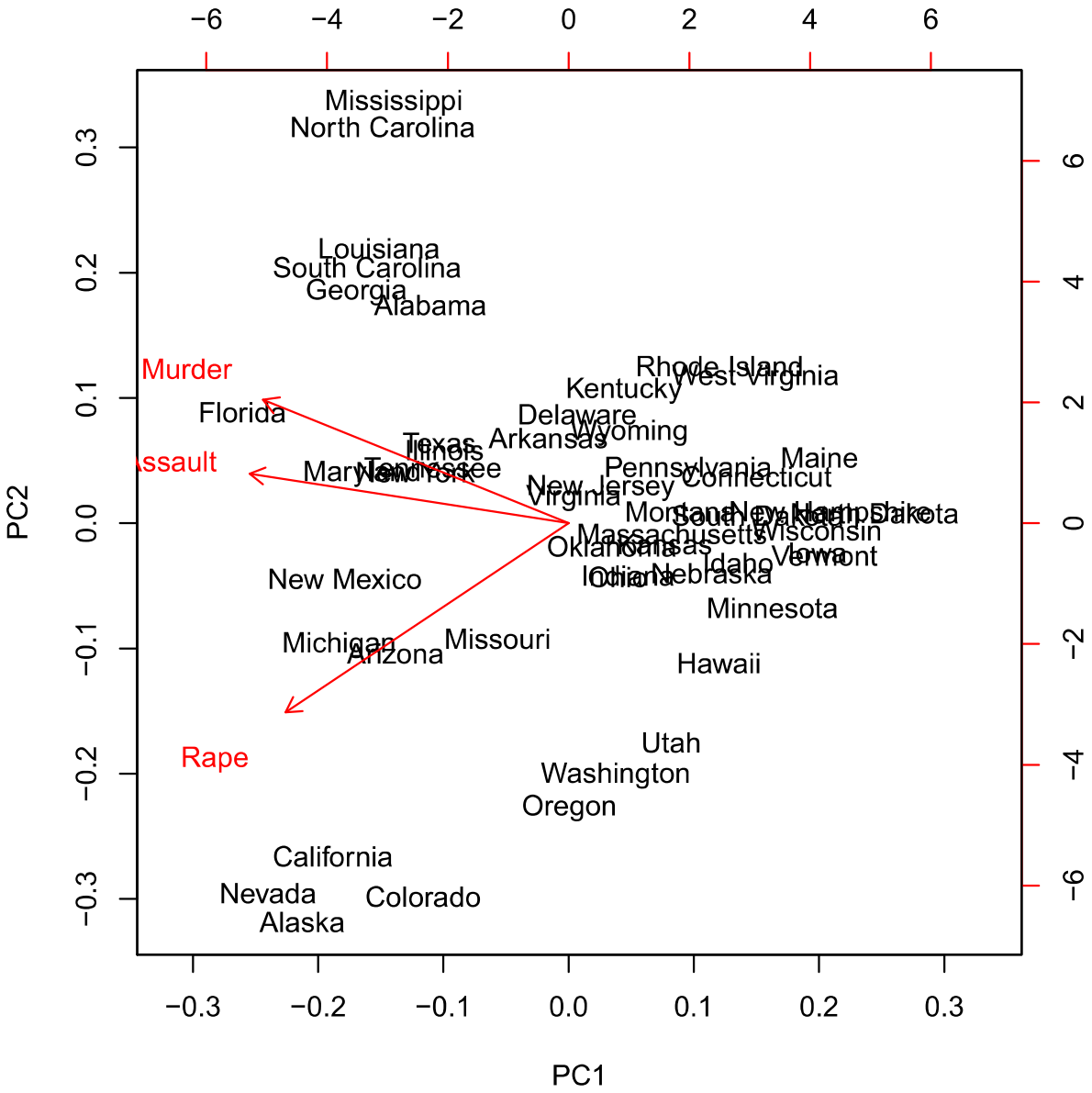


5. Narysuj wykres osypiska i zaproponuj optymalną liczbę składowych głównych w oparciu o trzy kryteria.

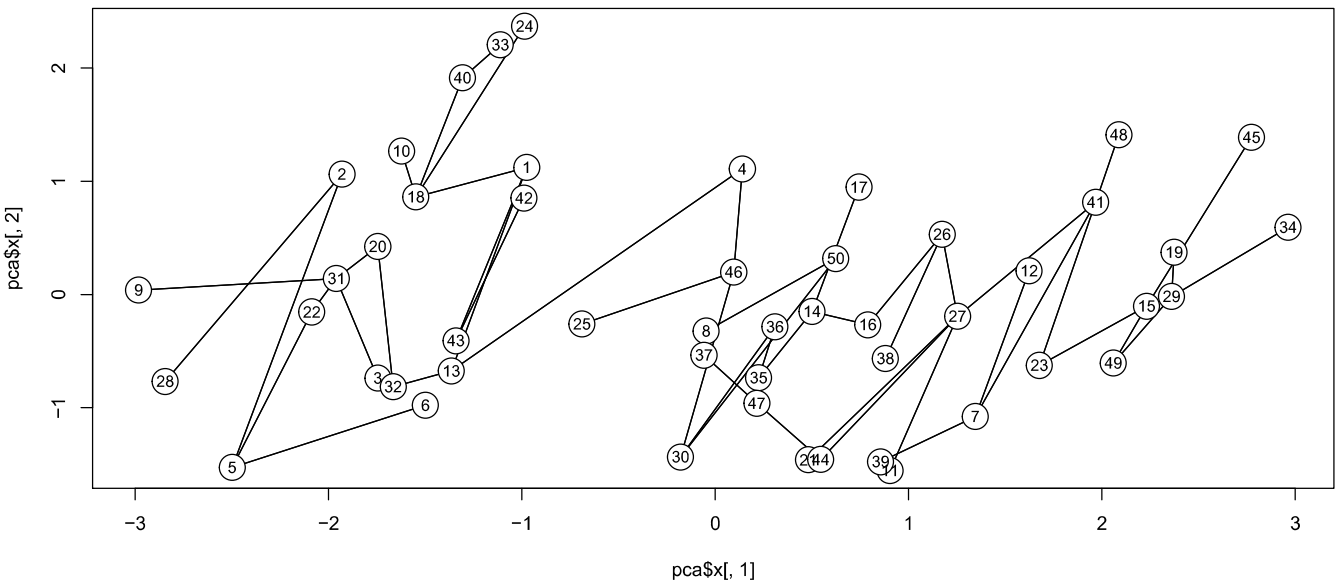



```
## 1 lub 2
```

6. Przedstaw stany w układzie dwóch pierwszych składowych głównych (dokładniej narysuj biplot i dokonaj jego interpretacji).



7. Przedstaw stany za pomocą minimalnego drzewa rozpinającego.



Zadanie 2. Zbiór danych `mtcars` zawiera informacje na temat 32 samochodów z roku 1974.

1. Dokonaj analizy składowych głównych biorąc pod uwagę cechy: `mpg`, `disp`, `hp`, `drat`, `wt`, `qsec`.

```
## Standard deviations (1, .., p=6):
## [1] 2.0463129 1.0714999 0.5773705 0.3928874 0.3532648 0.2279872
##
## Rotation (n x k) = (6 x 6):
##              PC1         PC2         PC3        PC4         PC5         PC6
## mpg  -0.4586835  0.05867609 -0.19479235 0.78205878 -0.1111533 -0.35249327
## disp  0.4660354 -0.06065296  0.09688406 0.60001871  0.2946297  0.56825752
## hp    0.4258534  0.36147576  0.14613554 0.12301873 -0.8057408 -0.04771555
## drat -0.3670963  0.43652537  0.80049152 0.02259258  0.1437714  0.11277675
## wt    0.4386179 -0.29953457  0.41776208 0.10438337  0.2301541 -0.69246040
## qsec -0.2528320 -0.76284877  0.34059066 0.04268124 -0.4218755  0.24152663
```

2. Jaki procent wariancji tłumaczony jest przez poszczególne składowe?

```
## Importance of components:
##                           PC1    PC2     PC3     PC4    PC5     PC6
## Standard deviation     2.0463 1.0715 0.57737 0.39289 0.3533 0.22799
## Proportion of Variance 0.6979 0.1913 0.05556 0.02573 0.0208 0.00866
## Cumulative Proportion  0.6979 0.8892 0.94481 0.97054 0.9913 1.00000
```

3. Wyznacz współrzędne obserwacji w nowym układzie współrzędnych utworzonym przez składowe główne.

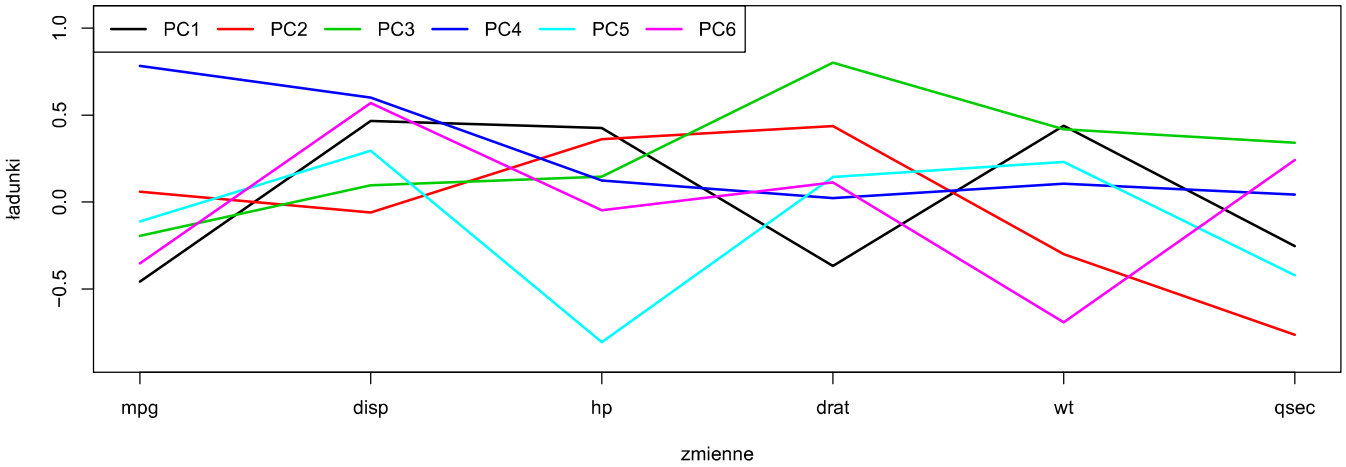
```
##                           PC1          PC2        PC3        PC4          PC5
## Mazda RX4         -0.8425806  0.873469391 -0.2282783 -0.3742725  0.51522641
## Mazda RX4 Wag     -0.8075041  0.556341552 -0.0126678 -0.3336931  0.44299870
## Datsun 710        -1.6850448 -0.040006569 -0.1564937 -0.4057157 -0.03340433
## Hornet 4 Drive    -0.0964443 -1.294377904 -0.5702297  0.2520788 -0.04326023
## Hornet Sportabout  1.2915096 -0.006516693 -0.5250741  0.4813192  0.12822104
## Valiant            0.2187309 -2.005957905 -0.7258399 -0.3136170 -0.21465335
##                           PC6
## Mazda RX4         -0.05293884
## Mazda RX4 Wag     -0.15771326
## Datsun 710         0.10756126
```

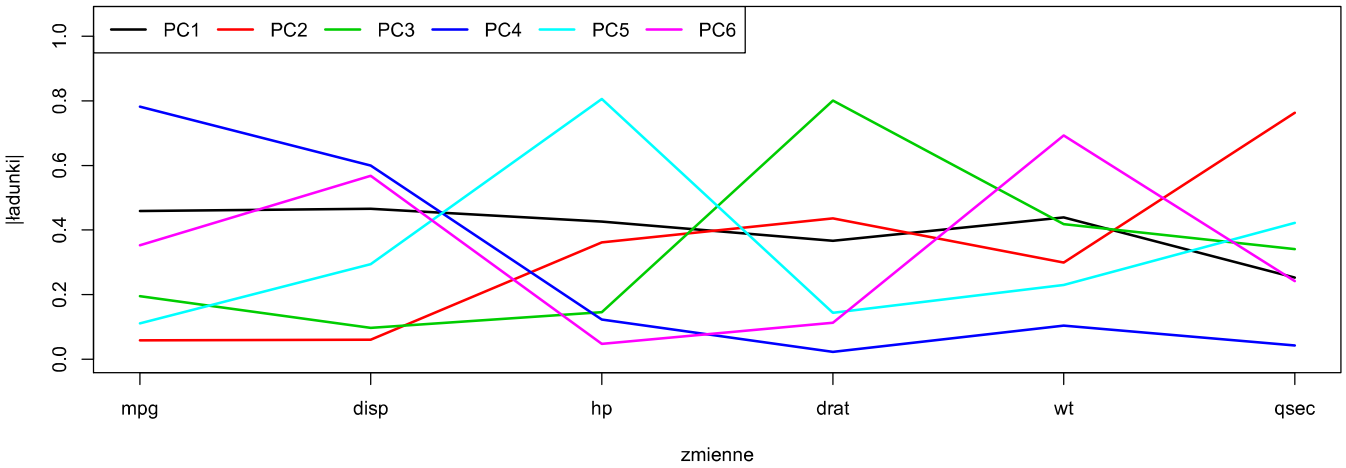
```
## Hornet 4 Drive     0.18173489
## Hornet Sportabout  0.29051949
## Valiant            0.09145688
```

```
## ...
```

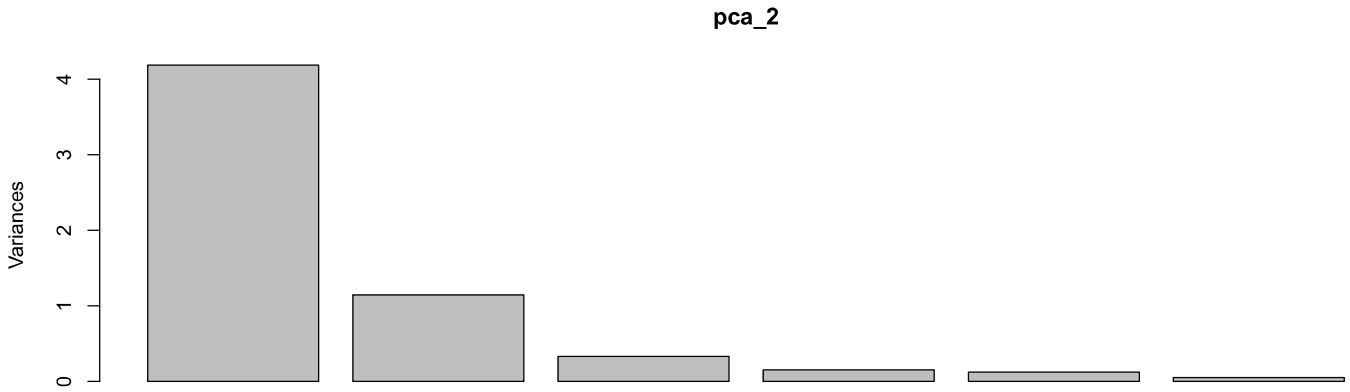
4. Dokonaj interpretacji ładunków i zilustruj je na wykresie.

```
##              PC1         PC2         PC3        PC4         PC5         PC6
## mpg  -0.4586835  0.05867609 -0.19479235 0.78205878 -0.1111533 -0.35249327
## disp  0.4660354 -0.06065296  0.09688406 0.60001871  0.2946297  0.56825752
## hp    0.4258534  0.36147576  0.14613554 0.12301873 -0.8057408 -0.04771555
## drat -0.3670963  0.43652537  0.80049152 0.02259258  0.1437714  0.11277675
## wt    0.4386179 -0.29953457  0.41776208 0.10438337  0.2301541 -0.69246040
## qsec -0.2528320 -0.76284877  0.34059066 0.04268124 -0.4218755  0.24152663
```



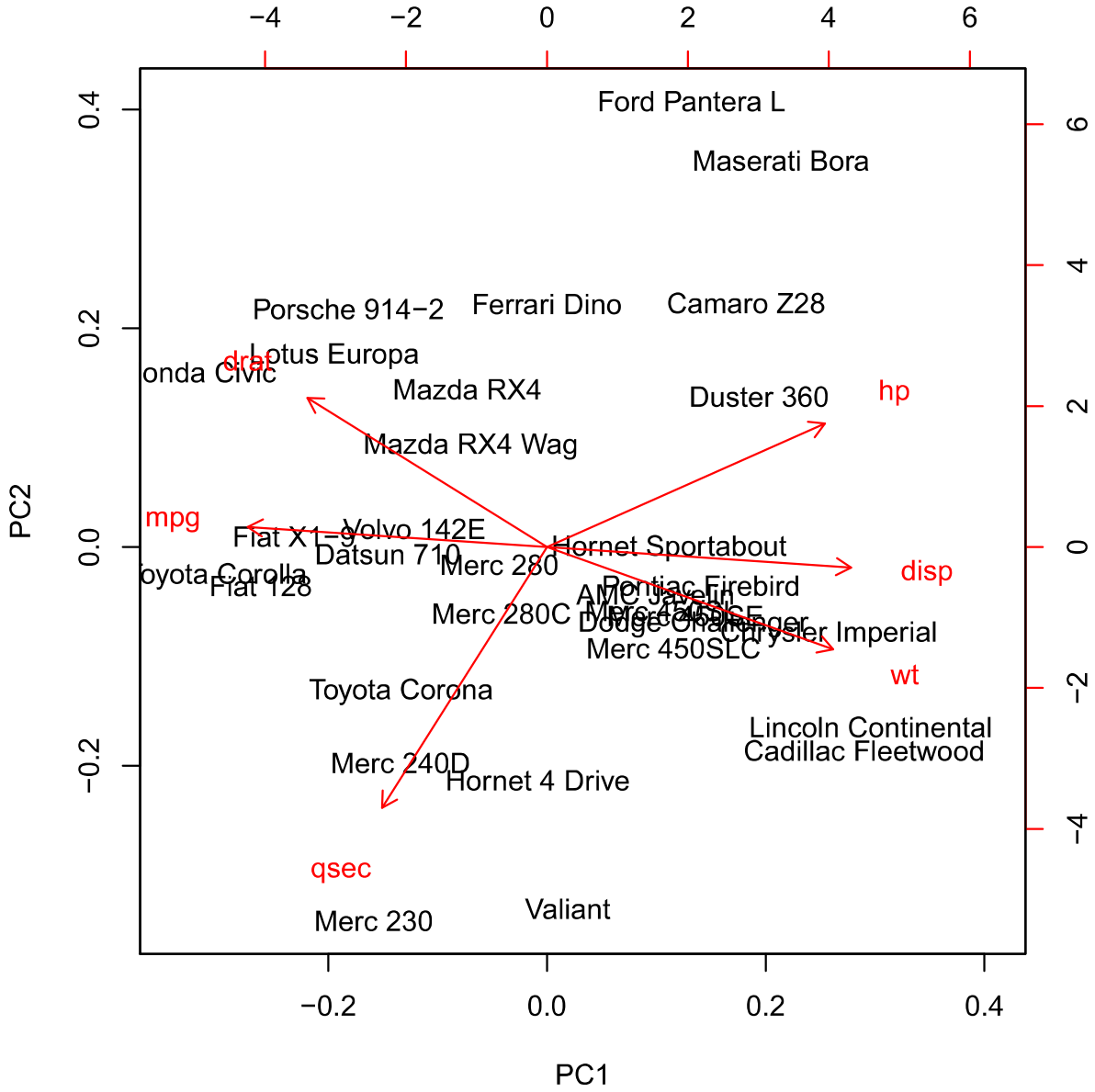


5. Narysuj wykres osypiska i zaproponuj optymalną liczbę składowych głównych w oparciu o trzy kryteria.

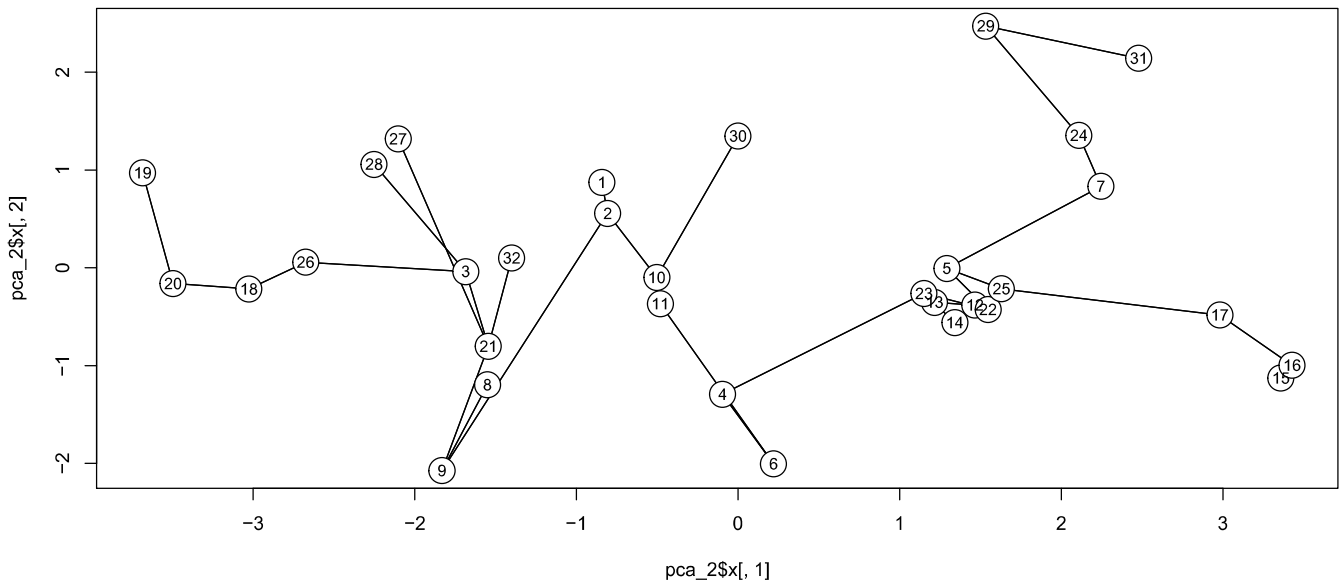



```
## 2 lub 3
```

6. Przedstaw samochody w układzie dwóch pierwszych składowych głównych (dokładniej narysuj biplot i dokonaj jego interpretacji).



7. Przedstaw samochody za pomocą minimalnego drzewa rozpinającego.



8. Jak bardzo będą różniły się wyniki, jeśli nie wykonamy skalowania danych?

Ad. 1

```
## Standard deviations (1, .., p=6):
## [1] 310.0207637  40.8471739  15.7168252   2.1068823   0.3894500   0.2969505
##
## Rotation (n x k) = (6 x 6):
##              PC1          PC2         PC3          PC4          PC5
## mpg  -0.05193468  0.121255352 -0.82446804  0.540735371 -0.064362234
## disp -0.85253108 -0.522102198 -0.00915689  0.022137483  0.001587345
## hp   -0.51734213  0.841835388  0.15361995 -0.004990023 -0.006795464
## drat -0.01010286  0.021298587 -0.10869056 -0.033506518  0.982931599
## wt   -0.01067910  0.001369032 -0.04162846 -0.192177061  0.129755288
## qsec -0.05132793  0.059700171 -0.53199901 -0.817945952 -0.113215907
##                PC6
## mpg   0.0794678281
## disp -0.0048593900
## hp   -0.0003699391
## drat -0.1426655136
## wt    0.9717935462
## qsec -0.1700734209
```

Ad. 2.

```
## Importance of components:
##                             PC1      PC2      PC3     PC4    PC5   PC6
## Standard deviation     310.0208 40.84717 15.71683 2.10688 0.3895 0.297
## Proportion of Variance   0.9804  0.01702  0.00252 0.00005 0.0000 0.000
## Cumulative Proportion    0.9804  0.99743  0.99995 1.00000 1.0000 1.000
```

Ad. 3.

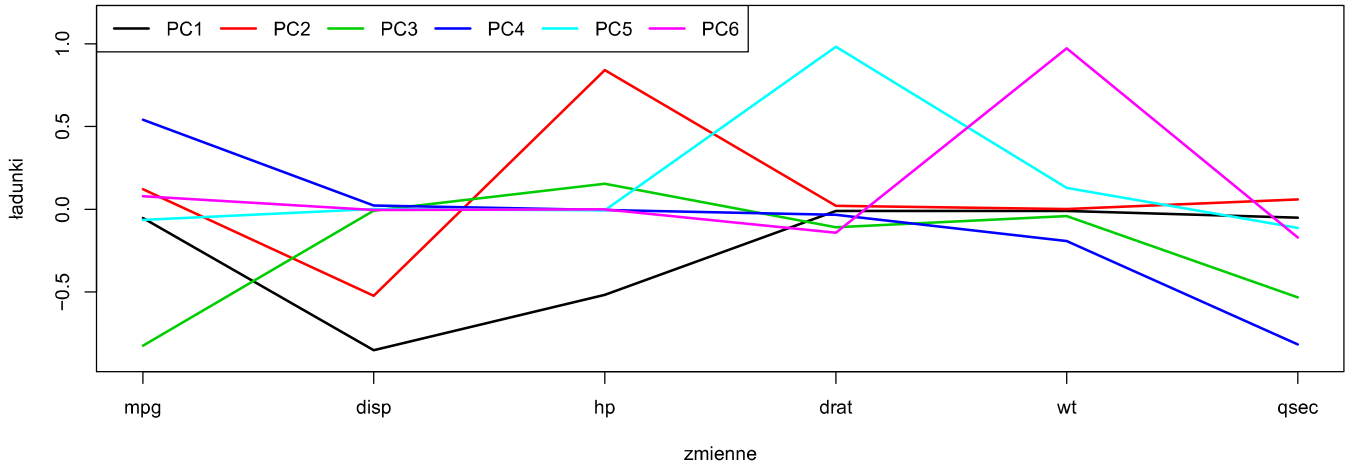
```
##                         PC1       PC2        PC3        PC4         PC5
## Mazda RX4         -195.3155  12.68122 -11.170400  0.2509678  0.46472555
## Mazda RX4 Wag     -195.3469  12.71500 -11.478935 -0.2560871  0.43441224
## Datsun 710        -142.3892  25.86447 -15.915699 -1.5412826  0.05036709
```

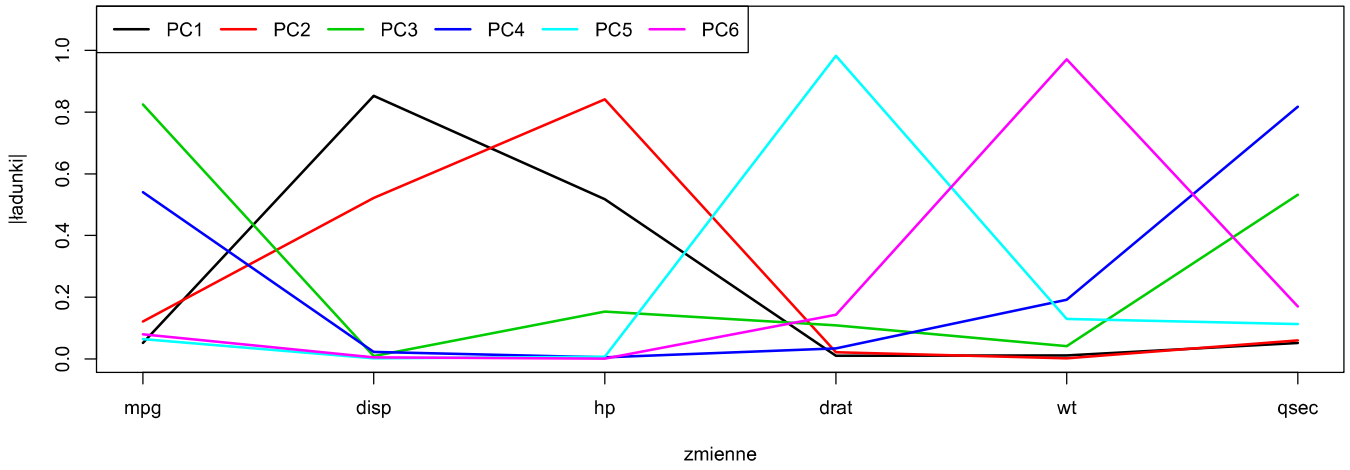
```
## Hornet 4 Drive    -279.0353 -38.27504 -13.918563  0.1123864 -0.47164240
## Hornet Sportabout -399.3594 -37.28023  -1.370742  2.5199166 -0.20567766
## Valiant           -248.1831 -25.61490 -12.054118 -3.0519863 -0.64870858
##                          PC6
## Mazda RX4           0.04092377
## Mazda RX4 Wag       0.19349001
## Datsun 710         -0.20711953
## Hornet 4 Drive     -0.21512523
## Hornet Sportabout -0.32914754
## Valiant            -0.16407438
```

```
## ...
```

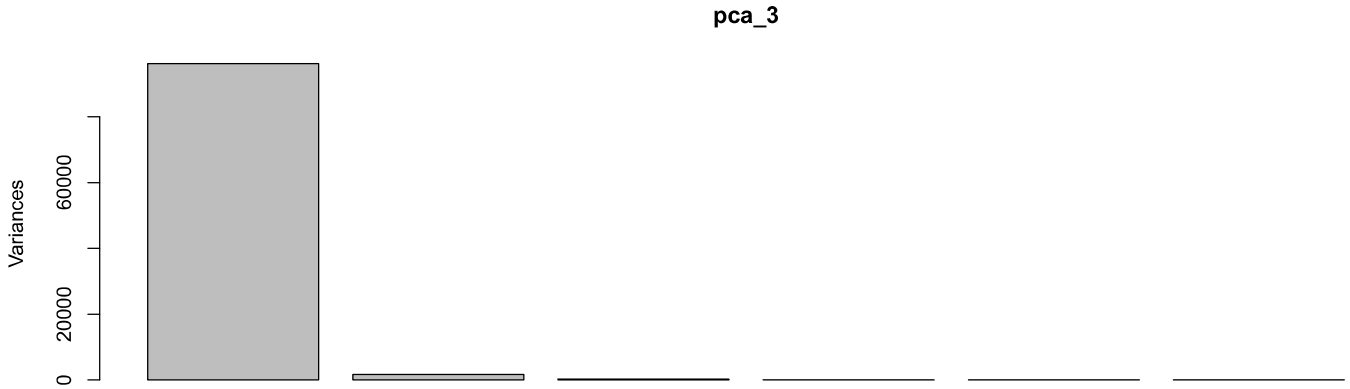
Ad. 4.

```
##               PC1          PC2         PC3          PC4          PC5
## mpg  -0.05193468  0.121255352 -0.82446804  0.540735371 -0.064362234
## disp -0.85253108 -0.522102198 -0.00915689  0.022137483  0.001587345
## hp   -0.51734213  0.841835388  0.15361995 -0.004990023 -0.006795464
## drat -0.01010286  0.021298587 -0.10869056 -0.033506518  0.982931599
## wt   -0.01067910  0.001369032 -0.04162846 -0.192177061  0.129755288
## qsec -0.05132793  0.059700171 -0.53199901 -0.817945952 -0.113215907
##                 PC6
## mpg   0.0794678281
## disp -0.0048593900
## hp   -0.0003699391
## drat -0.1426655136
## wt    0.9717935462
## qsec -0.1700734209
```


PC1
PC2
PC3
PC4
PC5
PC6
ładunki
1.0
0.5
0.0
−0.5
mpg
disp
hp
drat
wt
qsec
zmienne

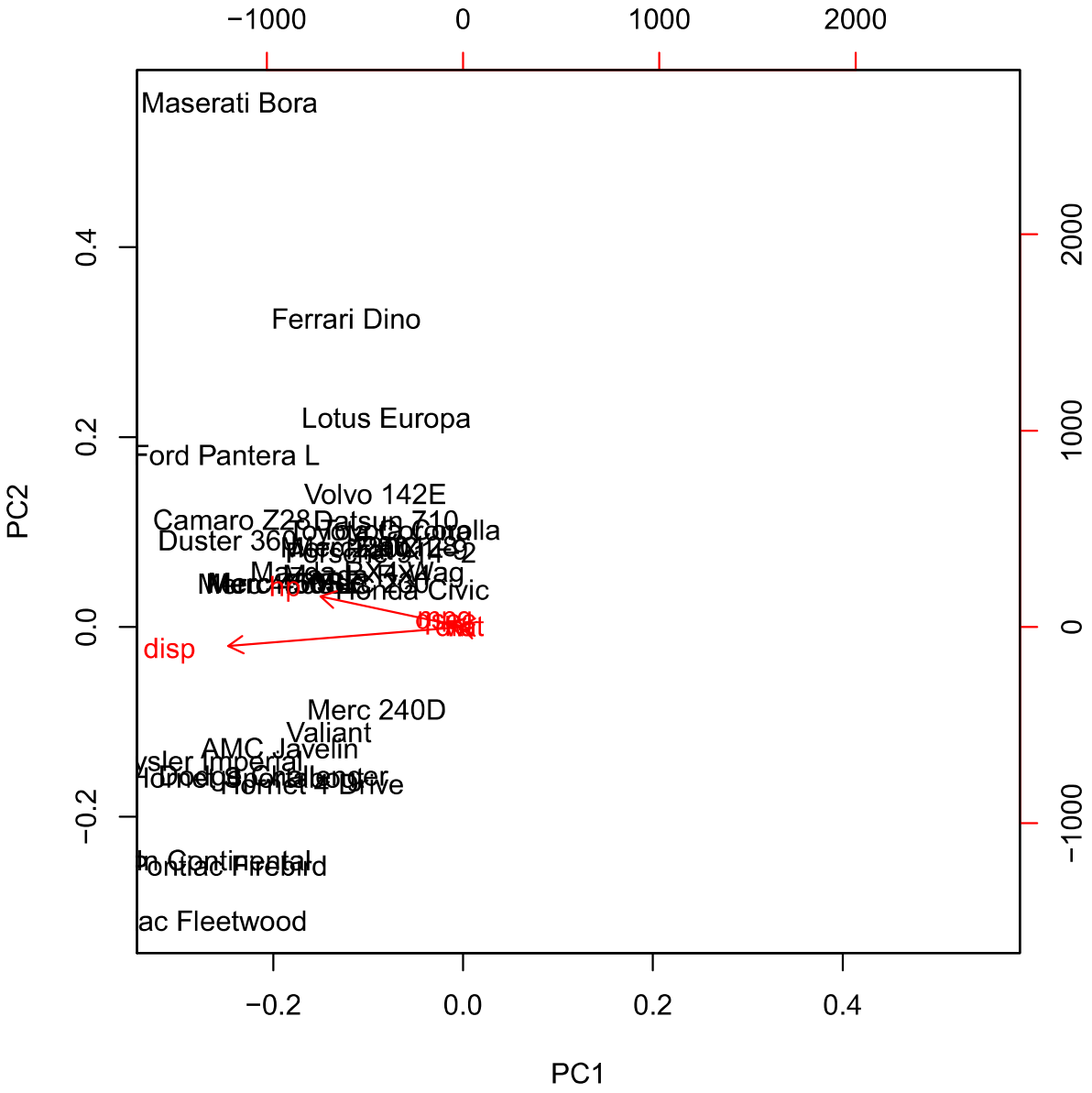

PC1
PC2
PC3
PC4
PC5
PC6
|ładunki|
1.0
0.8
0.6
0.4
0.2
0.0
mpg
disp
hp
drat
wt
qsec
zmienne

Ad. 5.

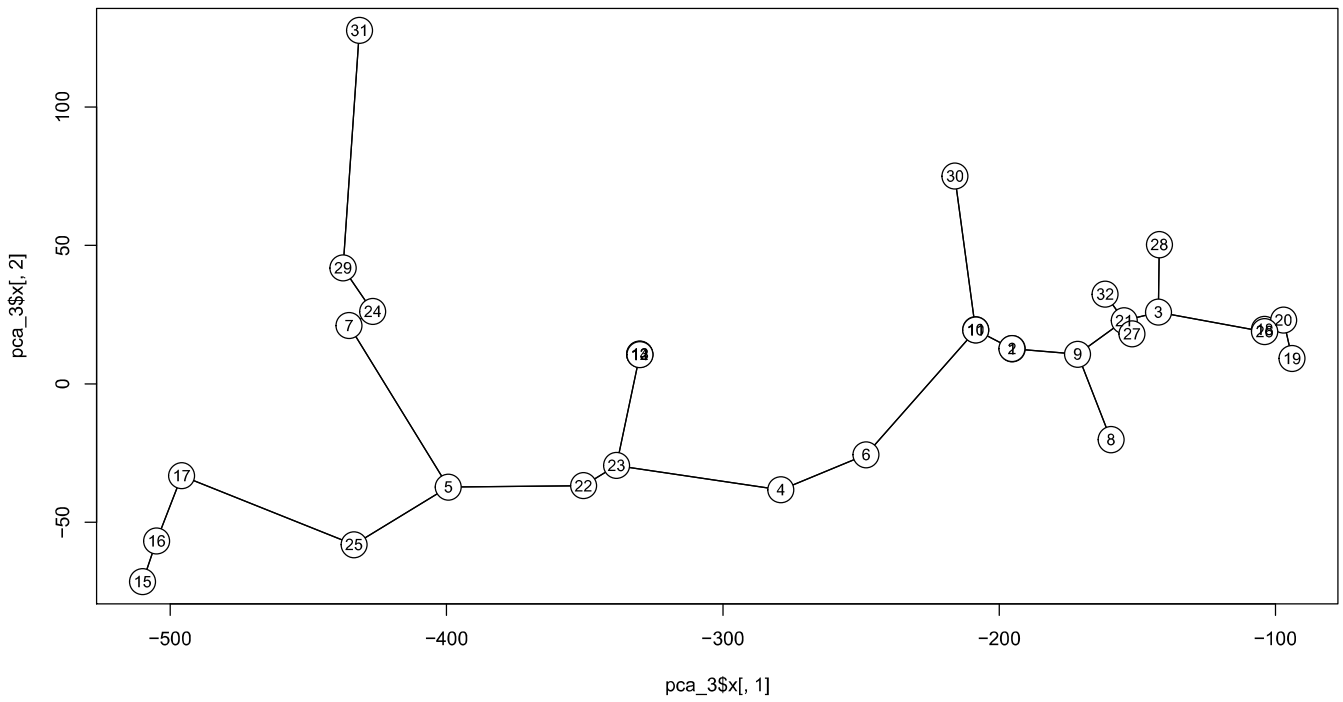

pca_3
Variances
0
20000
60000

```
## 1
```

Ad. 6.


−1000
0
1000
2000
Maserati Bora
Ferrari Dino
Lotus Europa
Volvo 142E
Camaro Z28
Datsun 710
Duster 360
Merc 240D
Valiant
AMC Javelin
Pontiac Firebird
ac Fleetwood
disp
0.4
0.2
0.0
−0.2
PC2
2000
1000
0
−1000
−0.2
0.0
0.2
0.4
PC1

Ad. 7.



14 Analiza skupień

14.1 Przykład

Przykład. Zbiór danych `USArrests` zawiera informacje dotyczące liczby morderstw, napadów, gwałtów przypadających na 100,000 osób w poszczególnych stanach USA w roku 1973 oraz procent ludności mieszkającej w miastach. Chcielibyśmy się dowiedzieć, czy i które stany są do siebie w pewien sposób zbliżone.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
dim(USArrests)
```

```
## [1] 50  4
```

1. metoda hierarchiczna

```
(skupienia_1 <- hclust(dist(USArrests)))
```

```
##
## Call:
## hclust(d = dist(USArrests))
##
## Cluster method   : complete
## Distance         : euclidean
```