Rozdział 5 - Analiza składowych głównych PCA - ćwiczenia

## Analiza składowych głównych PCA danych *USArrests*

1. Wczytanie i transformacja danych *USArrests*.

rm(list=ls())  
library(dplyr)

##   
 ## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
 ##   
 ## filter, lag

## The following objects are masked from 'package:base':  
 ##   
 ## intersect, setdiff, setequal, union

library(tidyr)  
Dane<-USArrests  
colnames(Dane)<-c("Morderstwa","Zabójstwa","Populacja miejska[%]","Gwałty")  
Dane<-Dane[,c(1,2,4,3)]  
Dane %>% head()

## Morderstwa Zabójstwa Gwałty Populacja miejska[%]  
 ## Alabama 13.2 236 21.2 58  
 ## Alaska 10.0 263 44.5 48  
 ## Arizona 8.1 294 31.0 80  
 ## Arkansas 8.8 190 19.5 50  
 ## California 9.0 276 40.6 91  
 ## Colorado 7.9 204 38.7 78

1. Wstępna analiza

library(knitr)  
kable(apply(Dane,2,function(x) c(summary(x),Sd=sd(x),Var=var(x))),digits = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Morderstwa | Zabójstwa | Gwałty | Populacja miejska[%] |
| Min. | 0.80 | 45.00 | 7.30 | 32.00 |
| 1st Qu. | 4.08 | 109.00 | 15.08 | 54.50 |
| Median | 7.25 | 159.00 | 20.10 | 66.00 |
| Mean | 7.79 | 170.76 | 21.23 | 65.54 |
| 3rd Qu. | 11.25 | 249.00 | 26.17 | 77.75 |
| Max. | 17.40 | 337.00 | 46.00 | 91.00 |
| Sd | 4.36 | 83.34 | 9.37 | 14.47 |
| Var | 18.97 | 6945.17 | 87.73 | 209.52 |

kable(cor(Dane),digits = 2)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Morderstwa | Zabójstwa | Gwałty | Populacja miejska[%] |
| Morderstwa | 1.00 | 0.80 | 0.56 | 0.07 |
| Zabójstwa | 0.80 | 1.00 | 0.67 | 0.26 |
| Gwałty | 0.56 | 0.67 | 1.00 | 0.41 |
| Populacja miejska[%] | 0.07 | 0.26 | 0.41 | 1.00 |

Zmienna ***Zabójstwa*** ma najwieksza wariancję ale nie jest to powód, dla którego należy wykonać normalizację. Powodem tym jest zmienna ***Populacja miejska[%]*** wyskalowana w innych jednostkach niż pozostałe zmienne, ktore określają ilość osób aresztowanych na 100tyś. mieszkańców. Normalizacja może być przeprowadzona podczas wywołania funkcji do analizy czynników głównych.

1. Analiza PCA

library(devtools)

## Loading required package: usethis

library(tibble)  
install\_github("vqv/ggbiplot")

## Skipping install of 'ggbiplot' from a github remote, the SHA1 (7325e880) has not changed since last install.  
 ## Use `force = TRUE` to force installation

PCA\_model <- prcomp(Dane,scale. = T,center = T)  
summary(PCA\_model)

## 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

screeplot(PCA\_model,type="lines")  
PCA\_model

## 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  
 ## Morderstwa -0.5358995 0.4181809 -0.3412327 0.64922780  
 ## Zabójstwa -0.5831836 0.1879856 -0.2681484 -0.74340748  
 ## Gwałty -0.5434321 -0.1673186 0.8177779 0.08902432  
 ## Populacja miejska[%] -0.2781909 -0.8728062 -0.3780158 0.13387773

cat("Średnie dla zmiennych przed normalizacją",PCA\_model$center)

## Średnie dla zmiennych przed normalizacją 7.788 170.76 21.232 65.54

cat("Odchylania std dla zmiennych przed normalizacją",PCA\_model$scale)

## Odchylania std dla zmiennych przed normalizacją 4.35551 83.33766 9.366385 14.47476

PCA\_model$x %>% data.frame() %>% rownames\_to\_column() %>% head()

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

library(ggbiplot)

## Loading required package: ggplot2

## Loading required package: plyr

## -------------------------------------------------------------------------

## You have loaded plyr after dplyr - this is likely to cause problems.  
 ## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:  
 ## library(plyr); library(dplyr)

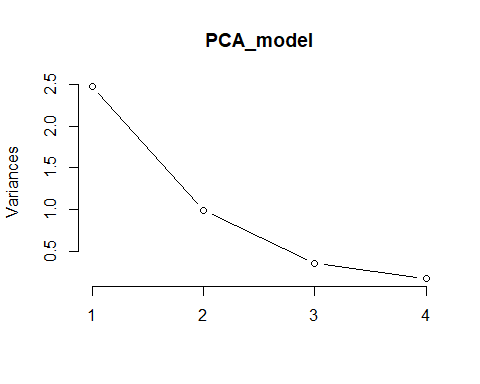
## -------------------------------------------------------------------------

##   
 ## Attaching package: 'plyr'

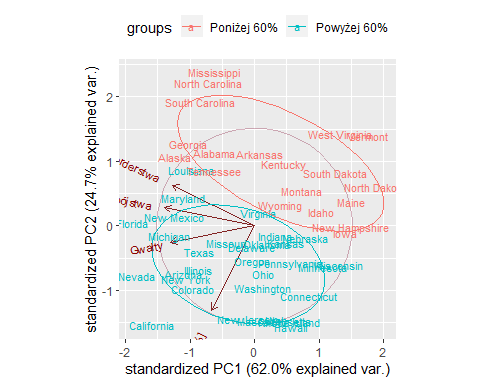
## The following objects are masked from 'package:dplyr':  
 ##   
 ## arrange, count, desc, failwith, id, mutate, rename, summarise,  
 ## summarize

## Loading required package: scales

## Loading required package: grid



Dane.pop<-as.factor(ifelse(Dane$`Populacja miejska[%]`>60,"Powyżej 60%","Poniżej 60%"))  
ggbiplot(PCA\_model,ellipse = TRUE, circle = TRUE,labels=rownames(Dane),groups = Dane.pop) + theme(legend.position = "top")



Z wykresu osypiska wynika, że zmienne mozną rzutować na dwie płaszczyzny - składową PC1 i PC2. Pierwsza z nich jest najbardziej skorelowana z wszystkimi rodzajami przestępstw (odwrotnie proporcjonalna), a druga opisuje stopień urbanizacji stanu. Zależy od procent populacji stanu zamiaszkałej w obszarach miejskich (odwrotnie proporcjonalna).

## Ćwiczenia

1. Porównać dla prezentowanych danych wyniki PCA bez normalizacji,
2. Wykonać PCA dla danych ***mtcars***