K-Means Verkhovod

# Download the data

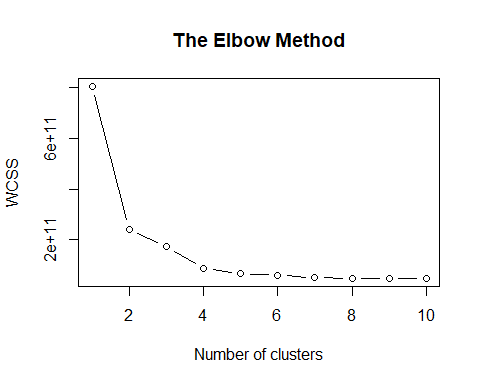
set.seed(123)  
setwd('D:/EP-61/R/')  
f <- read.csv2('clients.csv', header = TRUE, encoding = 'UNICOD')  
f <- f[, c('CLIENT\_TOGETHER.INCOME', 'LOAN\_OUTSTANDINGLOANSCOUNT')]  
head(f)

## CLIENT\_TOGETHER.INCOME LOAN\_OUTSTANDINGLOANSCOUNT  
## 1 3750 2  
## 2 4450 5  
## 3 8000 0  
## 4 2150 1  
## 5 6000 2  
## 6 3500 1

Висновок: для побудови моделі кластеризації використано дані про наявність прострочених платежів по кредиту. Залишимо для моделювання лише кількісні змінні. Датасет не потребує розподілу на навчальну та тестову вибірки, не потребує шкалювання.

# Elbow method to find optimal number of clusters

library(stats)   
wcss = vector()   
for (i in 1:10) wcss[i] = sum(kmeans(f, i)$withinss)   
plot(1:10,   
 wcss,   
 type = 'b',   
 main = paste('The Elbow Method'),   
 xlab = 'Number of clusters',   
 ylab = 'WCSS')



Висновок: з використанням метода ліктя зроблено висновок про доцільність виявлення трьох кластерів.

# Fitting K-Means to the dataset

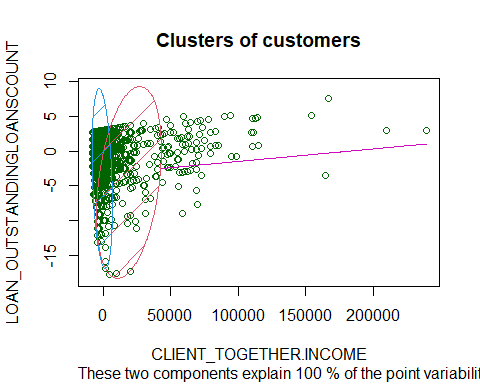
set.seed(29)  
model\_km = kmeans(f, 3)  
#cluster cores  
y\_km = model\_km$cluster  
aggregate(f,by=list(y\_km),FUN=mean)

## Group.1 CLIENT\_TOGETHER.INCOME LOAN\_OUTSTANDINGLOANSCOUNT  
## 1 1 21870.307 4.173442  
## 2 2 6759.006 2.833805  
## 3 3 83912.930 4.909091

Висновок: на основі аналізу дендрограми виявлено три кластера: 2 – “кредит для тих у кого малий дохід та мало виданих кредитів”, 1 – “кредит для тих у кого середній дохід та середня кількість виданих кредитів”, 3 – “кредит для тих у кого високий рінень доходу та багато виданих кредитів”. Розраховано характеристики типового об’єкту кластерів.

# Visualising the clusters

library(cluster)  
clusplot(f[,c('CLIENT\_TOGETHER.INCOME', 'LOAN\_OUTSTANDINGLOANSCOUNT')],  
 y\_km,  
 lines = 0,  
 shade = TRUE,  
 color = TRUE,  
 labels= 0,  
 plotchar = FALSE,  
 span = TRUE,  
 main = paste('Clusters of customers'),  
 xlab = 'CLIENT\_TOGETHER.INCOME',  
 ylab = 'LOAN\_OUTSTANDINGLOANSCOUNT')



Висновок: проведено візуалізацію кластерів на датасеті.