Hierarchical clustering Verkhovod

# Download the data

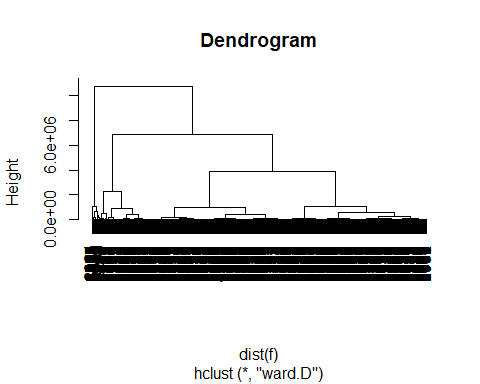
set.seed(123)  
setwd('D:/EP-61/R/')  
f <- read.csv2('clients.csv', header = TRUE, encoding = 'UNICOD')  
f <- f[, c('CLIENT\_TOGETHER.INCOME', 'LOAN\_OUTSTANDINGLOANSCOUNT')]  
head(f)

## 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.

# Hierarchical clustering

model\_hc <- hclust(dist(f), method = "ward.D" )  
plot(model\_hc, main = paste('Dendrogram'))



Висновок: на основі навчальної вибірки побудовано дендрограму з використанням методу Ward.D.

# Fitting HC to the dataset

y\_hc <- cutree(model\_hc, k = 3)  
#cluster cores  
aggregate(f,by=list(y\_hc),FUN=mean)

## Group.1 CLIENT\_TOGETHER.INCOME LOAN\_OUTSTANDINGLOANSCOUNT  
## 1 1 6411.53 2.780847  
## 2 2 19006.01 4.017331  
## 3 3 86163.67 4.934783

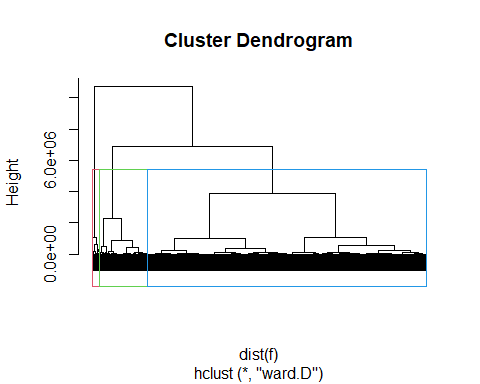
#Cluster stat  
f$hc <- y\_hc  
table(f$hc)

##   
## 1 2 3   
## 3331 577 92

Висновок: на основі аналізу дендрограми виявлено три кластера: 1 – “кредит для тих у кого малий дохід та мало виданих кредитів”, 2 – “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.

# Plotting the dendrogram

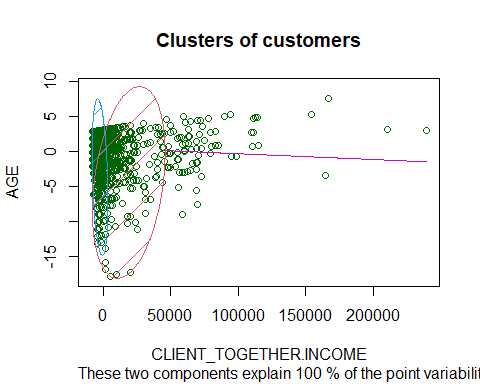
plot(model\_hc, cex = 0.7, labels = FALSE)  
rect.hclust(model\_hc, k = 3, border = 2:5)



Висновок: проведено візуалізацію кластерів на дендрограмі.

# Visualising the clusters

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



Висновок: проведено візуалізацію кластерів на датасеті.