Verkhovod clustering

# Download the data

library (psych)  
set.seed(123)  
setwd('D:')  
f <- read.csv2('clients.csv', header = TRUE, encoding = 'UNICOD')  
f <- f[,c('LOAN\_AMOUNT','CLIENT\_TOTALEXPERIENCE', 'CLIENT\_TOGETHER.INCOME', 'LOAN\_OUTSTANDINGLOANSCOUNT', 'LOAN.TERM', 'AGE')]  
head(f)

## LOAN\_AMOUNT CLIENT\_TOTALEXPERIENCE CLIENT\_TOGETHER.INCOME  
## 1 6224.07 31 3750  
## 2 16816.14 3 4450  
## 3 15560.17 20 8000  
## 4 5605.38 29 2150  
## 5 16816.14 30 6000  
## 6 10775.86 30 3500  
## LOAN\_OUTSTANDINGLOANSCOUNT LOAN.TERM AGE  
## 1 2 364 45  
## 2 5 1095 22  
## 3 0 364 41  
## 4 1 1095 45  
## 5 2 1095 54  
## 6 1 729 49

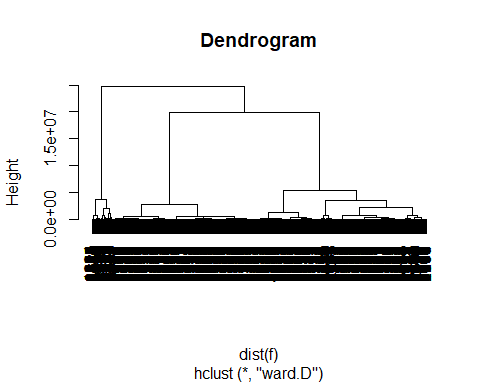
describe(f)

## vars n mean sd median trimmed  
## LOAN\_AMOUNT 1 4000 21470.59 20885.55 16816.14 17493.79  
## CLIENT\_TOTALEXPERIENCE 2 4000 18.93 12.30 17.00 18.24  
## CLIENT\_TOGETHER.INCOME 3 4000 10062.58 14185.66 7000.00 7404.52  
## LOAN\_OUTSTANDINGLOANSCOUNT 4 4000 3.01 2.48 2.00 2.67  
## LOAN.TERM 5 4000 794.45 341.65 729.00 803.05  
## AGE 6 4000 40.78 13.27 40.00 40.25  
## mad min max range skew kurtosis  
## LOAN\_AMOUNT 14958.96 1000.0 116822.4 115822.4 2.75 8.55  
## CLIENT\_TOTALEXPERIENCE 14.83 0.1 52.0 51.9 0.38 -0.97  
## CLIENT\_TOGETHER.INCOME 3203.57 1800.0 249464.5 247664.5 7.04 71.88  
## LOAN\_OUTSTANDINGLOANSCOUNT 1.48 0.0 21.0 21.0 1.85 5.89  
## LOAN.TERM 541.15 60.0 2007.0 1947.0 -0.16 -1.12  
## AGE 16.31 21.0 69.0 48.0 0.25 -1.06  
## se  
## LOAN\_AMOUNT 330.23  
## CLIENT\_TOTALEXPERIENCE 0.19  
## CLIENT\_TOGETHER.INCOME 224.29  
## LOAN\_OUTSTANDINGLOANSCOUNT 0.04  
## LOAN.TERM 5.40  
## AGE 0.21

Висновок: для побудови моделі кластеризації використано дані про наявність прострочених платежів по кредиту.кількість спостережень – 4000, кількість змінних – 6. Роботу з масивом будемо проводити без шкалювання та поділу на тестову та навчальну вибірки. Пропущених значень масив не має.

# 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 LOAN\_AMOUNT CLIENT\_TOTALEXPERIENCE CLIENT\_TOGETHER.INCOME  
## 1 1 7702.351 20.61943 5595.487  
## 2 2 25172.966 17.31228 9200.908  
## 3 3 87911.761 21.08675 49938.737  
## LOAN\_OUTSTANDINGLOANSCOUNT LOAN.TERM AGE  
## 1 2.931872 637.8424 43.28969  
## 2 2.977382 896.9293 38.64341  
## 3 3.841880 1014.1880 41.61538

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

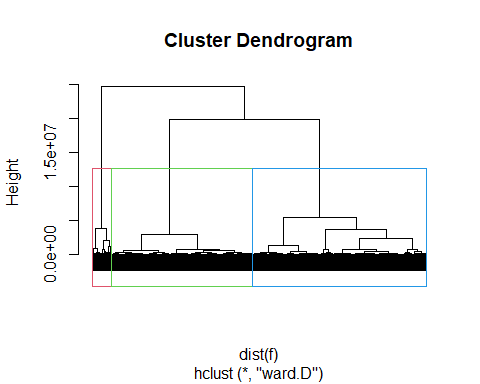
##   
## 1 2 3   
## 1688 2078 234

f= subset(f, select = -c(hc))

Висновок: на основі аналізу дендрограми виявлено три кластера. Розраховано характеристики типового об’єкту кластерів. 1 кластер - 1688 об’єктів: кредит на невелику суму клієнтів з малим доходом, віку більше 43 років на короткий термін. 2 кластер - 2078 об’єктів: кредит на середнього розміру для клієнтів з середнім доходом, віку менше 40 років на середній термін. 3 кластер - 234 об’єкти: кредит на велику суму клієнтів з великим доходом, віку близько 40 років на довгий термін.

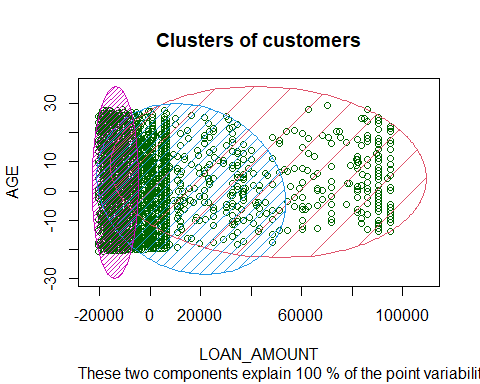
# Plotting the dendrogram

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('LOAN\_AMOUNT', 'AGE')],  
 y\_hc,  
 lines = 0,  
 shade = TRUE,  
 color = TRUE,  
 labels= 0,  
 plotchar = FALSE,  
 span = TRUE,  
 main = paste('Clusters of customers'),  
 xlab = 'LOAN\_AMOUNT',  
 ylab = 'AGE')

 Висновок: проведено візуалізацію кластерів на датасеті.

# Historgram for each attribute (K-Means)

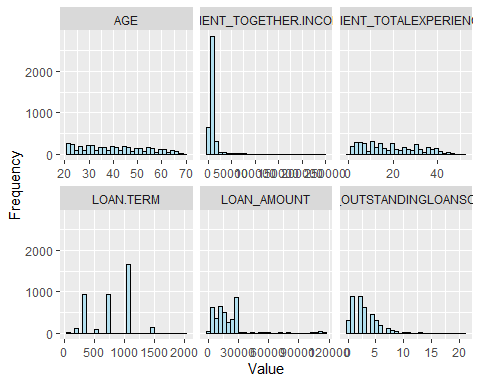
library(tidyr)  
library(ggplot2)

##   
## Attaching package: 'ggplot2'

## The following objects are masked from 'package:psych':  
##   
## %+%, alpha

f %>%   
 gather(Attributes, value, 1:6) %>%   
 ggplot(aes(x=value)) +  
 geom\_histogram(fill = "lightblue2", color = "black") +   
 facet\_wrap(~Attributes, scales = "free\_x") +  
 labs(x = "Value", y = "Frequency")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



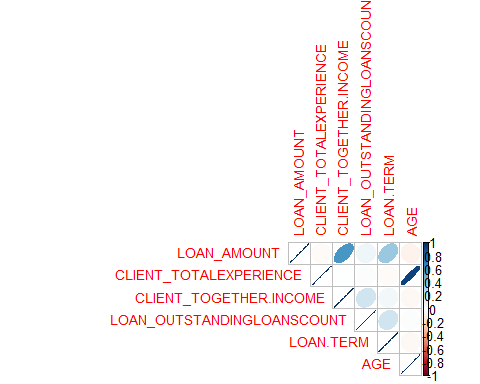
## Correlation

library(corrplot)

## Warning: package 'corrplot' was built under R version 4.0.3

## corrplot 0.84 loaded

corrplot(cor(f), type = "upper", method = "ellipse", tl.cex = 0.9)



Висновок: кореляція між показниками на середнбому та малому рівні. Кореляція між CLIENT\_TOTALEXPERIENCE та AGE дуже висока.

# Elbow method

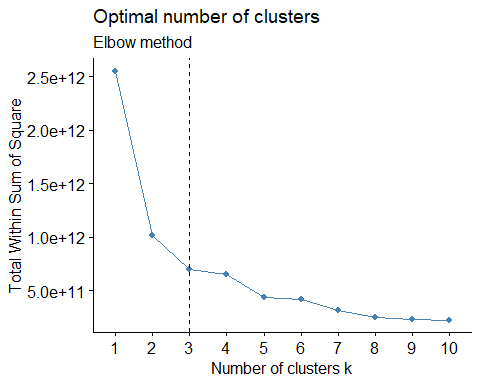
library(factoextra)

## Warning: package 'factoextra' was built under R version 4.0.3

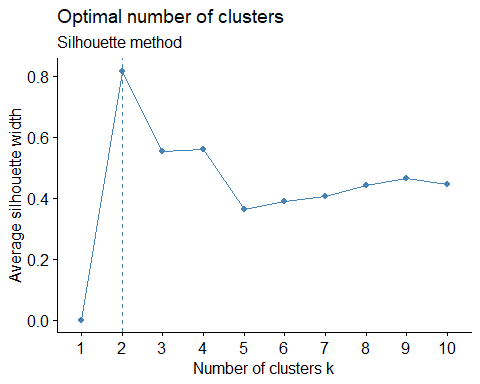
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

fviz\_nbclust(f, kmeans, method = "wss") +  
 geom\_vline(xintercept = 3, linetype = 2)+  
 labs(subtitle = "Elbow method")

## Warning: did not converge in 10 iterations

 # Silhouette method

fviz\_nbclust(f, kmeans, method = "silhouette")+  
 labs(subtitle = "Silhouette method")

 Висновок: з використанням метода ліктя зроблено висновок про доцільність виявлення трьох кластерів. Метод силуету - 2.

# Clustree

library(clustree)

## Warning: package 'clustree' was built under R version 4.0.3

## Loading required package: ggraph

## Warning: package 'ggraph' was built under R version 4.0.3

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

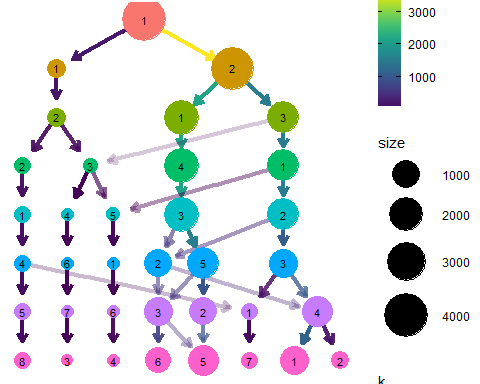
tmp <- NULL  
for (k in 1:8){  
 tmp[k] <- kmeans(f, k, nstart = 30)  
}

## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length

## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length  
  
## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length  
  
## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length  
  
## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length  
  
## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length  
  
## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length  
  
## Warning in tmp[k] <- kmeans(f, k, nstart = 30): number of items to replace is  
## not a multiple of replacement length

df <- data.frame(tmp)  
# add a prefix to the column names  
colnames(df) <- seq(1:8  
 )  
colnames(df) <- paste0("k",colnames(df))  
# get individual PCA  
df.pca <- prcomp(df, center = TRUE, scale. = FALSE)  
ind.coord <- df.pca$x  
ind.coord <- ind.coord[,1:2]  
df <- bind\_cols(as.data.frame(df), as.data.frame(ind.coord))  
clustree(df, prefix = "k")

## Warning: The `add` argument of `group\_by()` is deprecated as of dplyr 1.0.0.  
## Please use the `.add` argument instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_warnings()` to see where this warning was generated.



# Fitting K-Means to the dataset

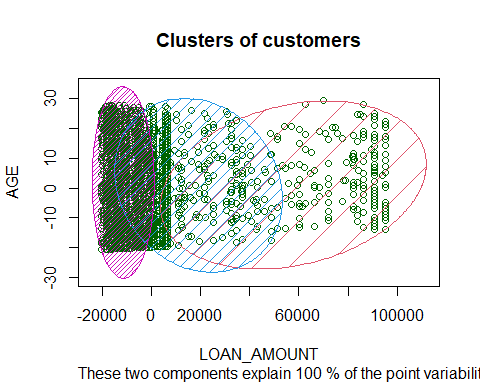
set.seed(29)  
model\_km = kmeans(f, 3)  
#cluster cores  
y\_km = model\_km$cluster  
aggregate(f,by=list(y\_km),FUN=mean)

## Group.1 LOAN\_AMOUNT CLIENT\_TOTALEXPERIENCE CLIENT\_TOGETHER.INCOME  
## 1 1 27805.412 17.31858 10603.239  
## 2 2 9695.563 19.97307 5971.038  
## 3 3 94473.334 20.28645 47836.897  
## LOAN\_OUTSTANDINGLOANSCOUNT LOAN.TERM AGE  
## 1 2.963727 917.3765 38.65916  
## 2 2.968907 680.8441 42.31321  
## 3 3.752336 1037.0374 40.92056

Висновок: на основі методу k-середніх описано три кластера:Розраховано характеристики типового об’єкту кластерів.

# Visualising the clusters

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

 Висновок: проведено візуалізацію кластерів на датасеті.

# Comparing to HC

library(clusteval)

## Warning: package 'clusteval' was built under R version 4.0.3

cluster\_similarity(y\_hc,y\_km)

## [1] 0.6245229

Висновок: стійкість кластерів 62,5%.

# Kohonen maps

# Features Scaling

mLOAN\_AMOUNT <- mean(f$LOAN\_AMOUNT)  
sLOAN\_AMOUNT <- sd(f$LOAN\_AMOUNT)  
f$LOAN\_AMOUNT <- (f$LOAN\_AMOUNT -mLOAN\_AMOUNT)/sLOAN\_AMOUNT  
  
mCLIENT\_TOTALEXPERIENCE <- mean(f$CLIENT\_TOTALEXPERIENCE)  
sCLIENT\_TOTALEXPERIENCE <- sd(f$CLIENT\_TOTALEXPERIENCE)  
f$CLIENT\_TOTALEXPERIENCE <- (f$CLIENT\_TOTALEXPERIENCE -mCLIENT\_TOTALEXPERIENCE)/sCLIENT\_TOTALEXPERIENCE  
  
mCLIENT\_TOGETHER.INCOME <- mean(f$CLIENT\_TOGETHER.INCOME)  
sCLIENT\_TOGETHER.INCOME <- sd(f$CLIENT\_TOGETHER.INCOME)  
f$CLIENT\_TOGETHER.INCOME <- (f$CLIENT\_TOGETHER.INCOME -mCLIENT\_TOGETHER.INCOME)/sCLIENT\_TOGETHER.INCOME  
  
mLOAN\_OUTSTANDINGLOANSCOUNT <- mean(f$LOAN\_OUTSTANDINGLOANSCOUNT)  
sLOAN\_OUTSTANDINGLOANSCOUNT <- sd(f$LOAN\_OUTSTANDINGLOANSCOUNT)  
f$LOAN\_OUTSTANDINGLOANSCOUNT <- (f$LOAN\_OUTSTANDINGLOANSCOUNT -mLOAN\_OUTSTANDINGLOANSCOUNT)/sLOAN\_OUTSTANDINGLOANSCOUNT  
  
mLOAN.TERM <- mean(f$LOAN.TERM)  
sLOAN.TERM <- sd(f$LOAN.TERM)  
f$LOAN.TERM <- (f$LOAN.TERM -mLOAN.TERM)/sLOAN.TERM  
  
mAGE <- mean(f$AGE)  
sAGE <- sd(f$AGE)  
f$AGE <- (f$AGE -mAGE)/sAGE  
  
  
f\_matrix <- as.matrix(f)

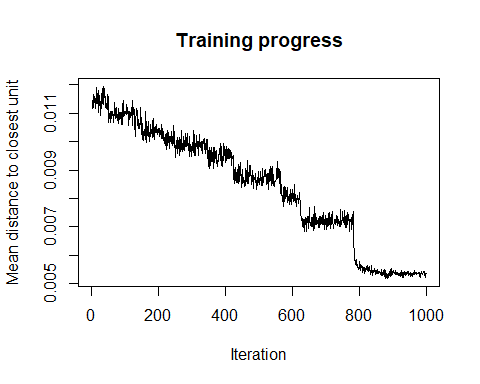
Висновок: модель Кохонена потребуює шкалювання, виконано шкалювання кількісних змінних.

# Fitting the NN

set.seed(123)  
library(kohonen)

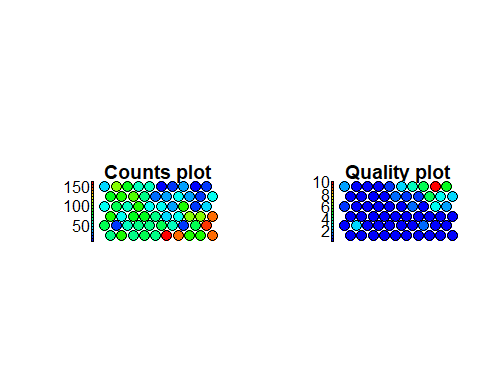
## Warning: package 'kohonen' was built under R version 4.0.3

som\_grid <- somgrid(xdim = 10, ydim = 6, topo = "hexagonal")   
som\_model <- som(f\_matrix, grid = som\_grid, rlen = 1000,  
 alpha = c(0.05,0.01), keep.data = TRUE)  
plot(som\_model, type = "changes")

 Висновок: графік зміни помилки нейронної мережі Кохонена свідчить про успішне навчання моделі.

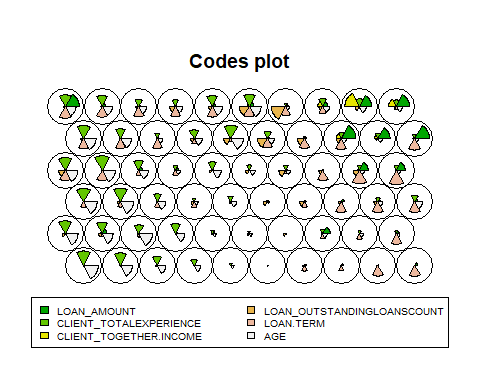
# Visualization

#Palette  
coolBlueHotRed <- function(n, alpha = 1) {  
 rainbow(n, end = 4/6, alpha = alpha)[n:1]   
}  
par(mfrow = c(1, 2))  
#Number of objects at sells  
plot(som\_model, type = "counts", palette.name = coolBlueHotRed)  
#Distance to core  
plot(som\_model, type = "quality", palette.name = coolBlueHotRed)

 Висновок: в кожну клітинку мережі потрапило від 50 до 200 спостережень, мережа досить повна.

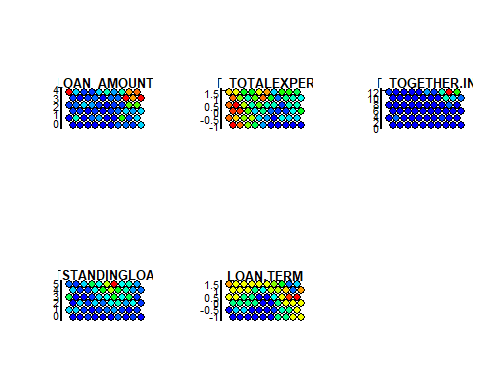
# Maps of the factors

plot(som\_model, type = "codes")



Висновок: ця візуалізація дозволяє аналізувати всі фактори на одній карті.

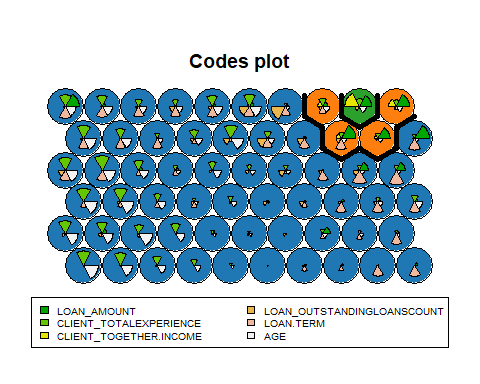
par(mfrow = c(2, 3))  
plot(som\_model, type = "property",   
 property = som\_model$codes[[1]][,1],   
 main = "LOAN\_AMOUNT",  
 palette.name = coolBlueHotRed)  
plot(som\_model, type = "property",   
 property = som\_model$codes[[1]][,2],  
 main = "CLIENT\_TOTALEXPERIENCE",  
 palette.name = coolBlueHotRed)  
plot(som\_model, type = "property",   
 property = som\_model$codes[[1]][,3],   
 main = "CLIENT\_TOGETHER.INCOME",  
 palette.name = coolBlueHotRed)  
plot(som\_model, type = "property",   
 property = som\_model$codes[[1]][,4],   
 main = "LOAN\_OUTSTANDINGLOANSCOUNT",  
 palette.name = coolBlueHotRed)  
plot(som\_model, type = "property",   
 property = som\_model$codes[[1]][,5],   
 main = "LOAN.TERM",  
 palette.name = coolBlueHotRed)



Висновок: ця візуалізація дозволяє аналізувати всі фактори на окремих картах та зробити висновок про наявність 3 кластерів.

# Clusters description

mydata <- as.matrix(som\_model$codes[[1]])  
#Use hierarchical clustering, k=3  
som\_cluster <- cutree(hclust(dist(mydata)), 3)  
#Palette  
pretty\_palette <- c("#1f77b4", '#ff7f0e', '#2ca02c',  
 '#d62728', '#9467bd', '#8c564b', '#e377c2')  
#Colored clusters  
plot(som\_model, type = "codes",   
 bgcol = pretty\_palette[som\_cluster])  
add.cluster.boundaries(som\_model, som\_cluster)



Висновок: ця візуалізація дозволяє аналізувати парамерти трьох виявлених кластерів на одній карті.

aggregate(mydata,by=list(som\_cluster),FUN=mean)

## Group.1 LOAN\_AMOUNT CLIENT\_TOTALEXPERIENCE CLIENT\_TOGETHER.INCOME  
## 1 1 0.03809592 0.12173535 -0.03081846  
## 2 2 2.95471918 0.03556275 4.54662164  
## 3 3 3.25540077 0.32537974 13.24174172  
## LOAN\_OUTSTANDINGLOANSCOUNT LOAN.TERM AGE  
## 1 0.3095990 -0.03709898 0.12529133  
## 2 0.5318332 -0.11903322 -0.07636121  
## 3 1.0052842 -1.05482060 0.14531125

Висновок: на основі нейроних мереж Кохонена виявлено три кластери. Розраховано характеристики типових об’єктів кластерів.