L6

Sokolova V

customers<- read.csv(file = "C:\\Users\\User\\Desktop\\3курс\\mmoi\\6\\marketing\_campaign11.csv", header = TRUE, sep = ";")  
  
products <- subset(customers , select = c(Wines, MeatProducts, FishProducts, SweetProducts, GoldProds))  
  
str(products)

## 'data.frame': 1582 obs. of 5 variables:  
## $ Wines : int 635 11 426 11 173 235 76 6 194 233 ...  
## $ MeatProducts : int 546 6 127 20 118 164 56 11 480 53 ...  
## $ FishProducts : int 172 2 111 10 46 50 3 11 225 3 ...  
## $ SweetProducts: int 88 1 21 3 27 49 1 1 112 5 ...  
## $ GoldProds : int 88 6 42 5 15 27 23 16 30 14 ...

summary(products)

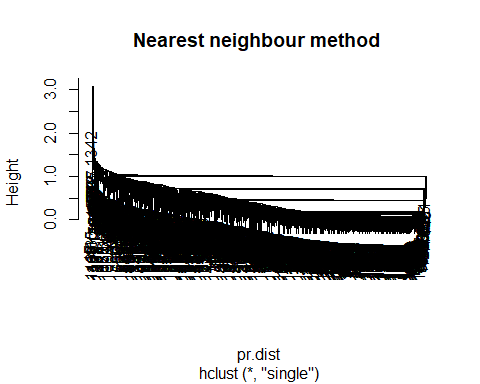
## Wines MeatProducts FishProducts SweetProducts   
## Min. : 1.0 Min. : 1.0 Min. : 1.0 Min. : 1.00   
## 1st Qu.: 33.0 1st Qu.: 25.0 1st Qu.: 7.0 1st Qu.: 5.00   
## Median : 222.5 Median : 103.0 Median : 23.0 Median : 15.50   
## Mean : 326.6 Mean : 197.8 Mean : 48.8 Mean : 35.80   
## 3rd Qu.: 522.8 3rd Qu.: 285.0 3rd Qu.: 70.5 3rd Qu.: 48.75   
## Max. :1493.0 Max. :1725.0 Max. :259.0 Max. :262.00   
## GoldProds   
## Min. : 1.00   
## 1st Qu.: 14.00   
## Median : 32.00   
## Mean : 50.63   
## 3rd Qu.: 68.00   
## Max. :321.00

# Для обраного датасету провести кластерний аналіз за кількома ознаками, використовуючи ієрархічну кластеризацію та метод к-середніх (або будь-які інші методики). Якщо потрібно – трансформувати дані. Відстань між спостереженнями для кількісних даних брати евклідову  
pr.scale <- scale(products)  
  
head(pr.scale)

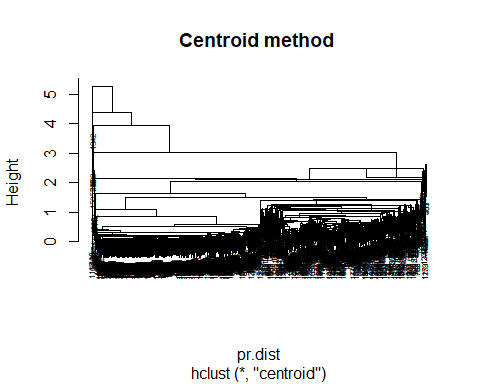
## Wines MeatProducts FishProducts SweetProducts GoldProds  
## 1 0.9260233 1.4764999 2.10755874 1.1686366 0.7085802  
## 2 -0.9476226 -0.8131484 -0.80067919 -0.7791477 -0.8461160  
## 3 0.2984721 -0.3000975 1.06401454 -0.3313812 -0.1635665  
## 4 -0.9476226 -0.7537871 -0.66382093 -0.7343711 -0.8650758  
## 5 -0.4611953 -0.3382584 -0.04795878 -0.1970512 -0.6754786  
## 6 -0.2750317 -0.1432142 0.02047034 0.2954919 -0.4479621

pr.dist <- dist(pr.scale, method = "euclidean")

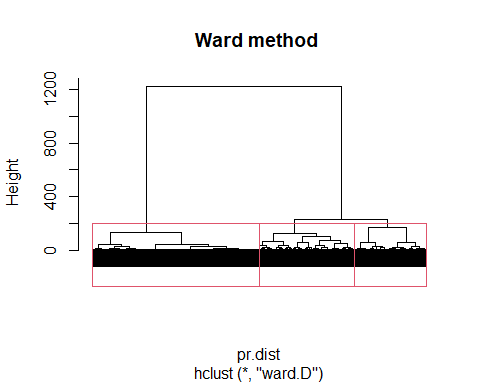
#метод найближчого сусіда   
nearest.neighbour = hclust(pr.dist, method = "single")  
plot(nearest.neighbour,main = "Nearest neighbour method ")



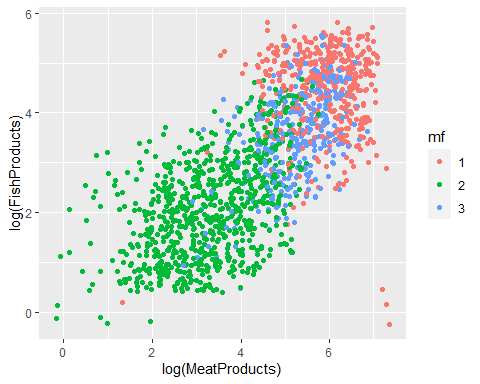
#Метод центроїдів  
centr.method <- hclust(pr.dist, method = "centroid")  
plot(centr.method, main = "Centroid method ", cex= 0.5)



#Метод Ворда  
ward.method <- hclust(pr.dist, method = "ward.D")  
plot(ward.method,main = "Ward method ", labels = FALSE)  
rect.hclust(ward.method , k = 3)



products$mf <- factor(cutree(ward.method, k = 3))  
  
library(ggplot2)  
ggplot(products, aes(x = log(MeatProducts),y = log(FishProducts),color = mf)) +  
 geom\_point( position = position\_jitter(h = 0.3, w = 0.3))



#Обрати оптимальну кількість кластерів. Для цього вивести необхідні зображення, які оптимізують різні показники для різної кількості кластерів  
library(fpc)

## Warning: package 'fpc' was built under R version 4.1.2

clusterboot1 <- clusterboot(pr.scale, clustermethod = hclustCBI, method = "ward.D", k = 3,count = FALSE)  
hcl\_cboot\_groups <- clusterboot1$result$partition

clusterboot1$bootmean

## [1] 0.6409466 0.8633873 0.4817525

clusterboot1$bootbrd

## [1] 9 0 54

clustering\_ch <- kmeansruns(pr.scale, krange = 1:10, criterion = "ch")  
clustering\_ch$bestk

## [1] 2

clustering\_asw <- kmeansruns(pr.scale, krange = 1:10, criterion = "asw")  
clustering\_asw$bestk

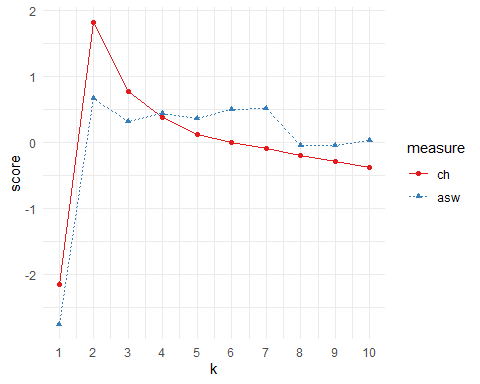
## [1] 2

library(reshape2)

## Warning: package 'reshape2' was built under R version 4.1.2

criteria <- data.frame(k = 1:10, ch = scale(clustering\_ch$crit),   
 asw = scale(clustering\_asw$crit))  
criteria <- melt(criteria, id.vars = c("k"),   
 variable.name = "measure",   
 value.name = "score")

ggplot(criteria, aes(x = k, y = score, col = measure)) +   
 geom\_point(aes(shape = measure)) + geom\_line(aes(linetype = measure)) +   
 scale\_x\_continuous(breaks = 1:10, labels = 1:10) +   
 scale\_color\_brewer(palette = "Set1") +   
 theme\_minimal()

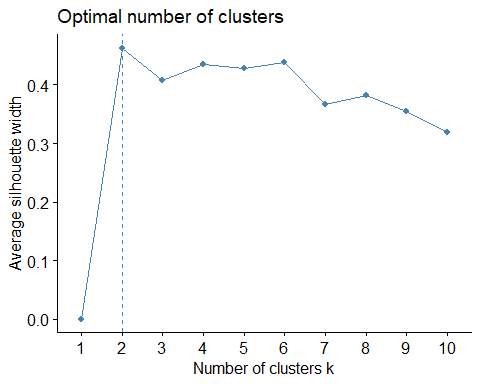


library(factoextra)

## Warning: package 'factoextra' was built under R version 4.1.2

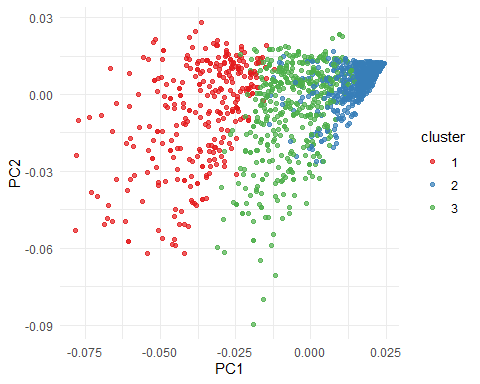
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa

fviz\_nbclust(pr.scale , kmeans , method = "silhouette")



#Для двох-трьох найкращих, на Вашу думку, варіантів кластеризації, зобразити результати кластеризації  
  
principal.component<- prcomp(pr.scale)  
visual <- predict(principal.component, newdata = pr.scale)[, 1:2]  
visual1 <- cbind(as.data.frame(visual),cluster = as.factor(kmeans\_cboot\_groups))

ggplot(visual1, aes(x = PC1, y = PC2)) +   
 geom\_point(aes(col = cluster), alpha = 0.7) +theme\_minimal() +scale\_color\_brewer(palette = "Set1")



#Порівняти обрані варіанти кластеризації, використовуючи індекс Ренда.   
groups5 <- cutree(ward.method, k = 3)  
m <- dist(pr.scale)  
hc2 <- hclust(m, method = "average")   
groups5\_2 <- cutree(hc2, k = 3) #Between-groups linkage  
library(fossil)

## Warning: package 'fossil' was built under R version 4.1.2

## Loading required package: sp

## Warning: package 'sp' was built under R version 4.1.2

## Loading required package: maps

## Warning: package 'maps' was built under R version 4.1.2

## Loading required package: shapefiles

## Loading required package: foreign

##   
## Attaching package: 'shapefiles'

## The following objects are masked from 'package:foreign':  
##   
## read.dbf, write.dbf

rand.index(groups5, groups5\_2)

## [1] 0.3798033