PEC 4 - Agregación *(Clustering)*

Fernando Antonio Barbeiro Campos - [fbarbeiro@uoc.edu](mailto:fbarbeiro@uoc.edu)

25 de April, 2018

Table of Contents

# Loading the dataframe  
df <- read.csv2("seguros2.csv")  
head(df)

## Numero.incidente Sexo Edad Anyos.Coche Caballos Costos  
## 1 5126419 1 46 6 110 2436.86  
## 2 9915975 1 77 3 100 1322.65  
## 3 1250208 0 24 10 50 6871.83  
## 4 1332233 0 50 11 70 1652.94  
## 5 2558883 1 40 15 50 1495.18  
## 6 2862115 0 56 9 102 2018.52

summary(df)

## Numero.incidente Sexo Edad Anyos.Coche   
## Min. :1001809 Min. :0.0000 Min. :18.00 Min. : 0.000   
## 1st Qu.:3369286 1st Qu.:0.0000 1st Qu.:30.00 1st Qu.: 5.000   
## Median :5337270 Median :1.0000 Median :43.00 Median :10.000   
## Mean :5422967 Mean :0.5442 Mean :44.06 Mean : 9.896   
## 3rd Qu.:7412620 3rd Qu.:1.0000 3rd Qu.:57.00 3rd Qu.:15.000   
## Max. :9971262 Max. :1.0000 Max. :77.00 Max. :19.000   
##   
## Caballos Costos   
## Min. : 30.00 1003.45: 1   
## 1st Qu.: 49.00 1016.32: 1   
## Median : 58.00 1025.02: 1   
## Mean : 79.68 1029.29: 1   
## 3rd Qu.:100.00 1031.02: 1   
## Max. :354.00 1035.85: 1   
## (Other):492

nrow(df)

## [1] 498

ncol(df)

## [1] 6

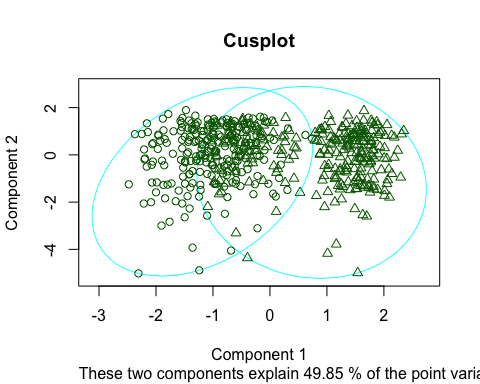
colnames(df)

## [1] "Numero.incidente" "Sexo" "Edad"   
## [4] "Anyos.Coche" "Caballos" "Costos"

res <- sapply(df, class)  
kable(data.frame(variables=names(res),clase=as.vector(res)))

|  |  |
| --- | --- |
| variables | clase |
| Numero.incidente | integer |
| Sexo | integer |
| Edad | integer |
| Anyos.Coche | integer |
| Caballos | integer |
| Costos | factor |

set.seed(20)  
k <- kmeans(df[, 2:3], 2, nstart = 20)  
k$cluster <- as.factor(k$cluster)  
  
library(cluster)  
clusplot(df, k$cluster, main = 'Cusplot')



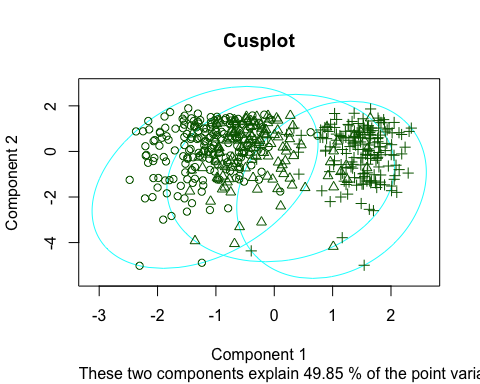
k

## K-means clustering with 2 clusters of sizes 238, 260  
##   
## Cluster means:  
## Sexo Edad  
## 1 0.5336134 58.84034  
## 2 0.5538462 30.52308  
##   
## Clustering vector:  
## [1] 1 1 2 1 2 1 2 1 1 2 2 2 2 1 1 1 1 1 2 2 2 1 2 1 1 1 1 2 1 1 2 2 2 2 2  
## [36] 2 1 2 1 2 1 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 2 1 2 2 1 2 1 1 1 1 1 1 1 1  
## [71] 2 1 2 2 2 2 2 2 2 1 2 2 1 2 1 1 1 2 2 2 2 2 2 1 1 2 2 2 1 1 2 2 2 1 1  
## [106] 1 1 1 2 1 1 2 2 2 1 1 2 2 1 2 1 2 1 2 2 2 2 1 1 1 2 1 1 1 2 1 1 1 2 2  
## [141] 2 2 2 2 1 2 1 1 1 2 1 1 1 2 1 1 2 1 1 1 2 2 1 2 1 2 2 2 1 1 2 2 1 2 2  
## [176] 1 2 2 2 2 1 1 2 2 1 1 2 2 1 2 2 2 1 2 2 2 1 2 1 2 2 1 2 2 2 2 1 1 1 1  
## [211] 2 2 2 1 2 2 2 2 1 2 2 1 2 2 2 1 1 1 2 2 2 2 2 2 2 1 1 2 1 1 2 2 1 1 1  
## [246] 1 1 2 1 2 2 2 1 1 1 1 1 2 1 1 2 2 1 1 2 2 2 2 2 2 1 1 2 2 2 1 1 2 2 2  
## [281] 1 2 2 2 1 1 2 2 1 2 2 2 2 1 1 1 1 2 1 1 2 1 2 2 1 2 2 1 2 2 1 2 1 2 2  
## [316] 2 2 1 1 1 2 2 2 1 1 2 2 1 1 2 2 1 2 2 1 2 2 1 2 2 1 2 2 1 1 1 1 1 1 2  
## [351] 2 1 1 2 1 1 1 1 1 1 2 2 1 2 2 2 1 1 2 2 1 1 1 2 2 1 1 2 1 1 1 1 2 1 2  
## [386] 2 1 1 2 2 1 1 2 1 1 1 2 1 1 1 2 1 1 1 1 2 1 1 1 1 2 2 2 2 1 1 1 1 1 2  
## [421] 1 1 1 1 2 2 2 2 2 1 2 2 2 1 2 1 1 1 1 2 2 2 2 2 1 2 1 1 1 1 1 1 2 2 2  
## [456] 1 2 2 2 2 2 1 2 2 1 2 2 2 2 2 1 2 2 2 1 1 1 1 2 2 1 1 2 1 1 1 1 2 2 1  
## [491] 2 2 1 2 2 1 1 2  
## Levels: 1 2  
##   
## Within cluster sum of squares by cluster:  
## [1] 15701.16 14489.11  
## (between\_SS / total\_SS = 76.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

set.seed(10)  
k2 <- kmeans(df[, 2:4], 3, nstart = 1)  
k2$cluster <- as.factor(k2$cluster)  
k2

## K-means clustering with 3 clusters of sizes 183, 139, 176  
##   
## Cluster means:  
## Sexo Edad Anyos.Coche  
## 1 0.5191257 62.00546 9.879781  
## 2 0.5539568 42.87050 10.158273  
## 3 0.5625000 26.32955 9.704545  
##   
## Clustering vector:  
## [1] 2 1 3 2 2 1 3 1 1 3 2 3 3 1 1 1 1 2 3 3 3 1 3 2 1 2 2 3 1 1 3 2 3 3 3  
## [36] 3 1 2 1 3 2 3 3 3 2 2 2 3 3 1 3 1 1 2 1 2 3 2 3 2 1 3 1 1 1 1 2 2 2 1  
## [71] 3 1 2 3 3 3 3 2 2 1 2 3 1 3 2 1 1 3 2 3 3 3 3 2 2 3 2 3 1 1 3 3 3 1 1  
## [106] 1 1 2 3 1 1 3 3 3 2 1 2 2 1 3 1 3 1 3 2 2 3 1 1 1 3 2 2 1 3 2 1 2 2 3  
## [141] 3 3 2 2 1 3 1 2 1 3 1 1 1 3 1 1 2 1 1 1 3 3 1 3 1 3 3 2 2 1 3 2 1 3 3  
## [176] 1 3 3 2 3 2 1 3 3 2 1 2 3 1 3 2 2 1 3 3 3 1 2 2 3 3 1 3 2 3 2 1 1 1 1  
## [211] 3 2 3 1 3 2 2 3 1 3 3 1 3 3 2 1 1 1 3 3 2 3 3 3 3 2 1 2 1 1 2 2 2 1 2  
## [246] 1 1 3 1 3 2 3 1 1 1 1 1 3 1 1 3 3 1 1 2 2 2 3 3 3 1 1 2 3 2 1 1 3 2 3  
## [281] 1 3 2 3 1 1 3 2 1 3 3 2 3 1 2 1 1 3 1 2 3 1 2 3 1 3 3 2 3 3 1 3 1 3 3  
## [316] 2 3 1 2 2 2 3 3 1 1 3 3 1 1 2 3 2 3 3 1 3 3 1 3 3 2 2 3 1 2 1 2 1 2 2  
## [351] 3 1 1 2 1 1 1 1 1 1 2 3 1 2 2 3 1 1 3 3 1 1 2 3 3 1 1 3 1 1 1 1 2 1 3  
## [386] 3 1 1 3 3 1 1 3 1 1 1 2 1 1 1 2 1 1 1 1 2 1 2 1 1 2 2 3 3 1 2 2 1 1 3  
## [421] 1 1 1 1 3 2 3 3 2 2 3 3 3 1 3 2 2 1 1 2 2 3 2 3 1 3 1 2 1 2 1 1 3 2 2  
## [456] 2 3 3 3 2 3 1 3 2 1 3 3 2 3 2 1 3 3 3 1 1 2 1 2 2 2 2 2 1 2 2 1 3 3 2  
## [491] 2 3 1 2 2 1 1 2  
## Levels: 1 2 3  
##   
## Within cluster sum of squares by cluster:  
## [1] 13802.033 7902.532 9284.835  
## (between\_SS / total\_SS = 78.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

clusplot(df, k2$cluster, main = 'Cusplot')



dfi <- read.csv("iris.csv")  
head(dfi)

## Sepal.Length Sepal.Width Petal.Length Petal.Width Species  
## 1 5.1 3.5 1.4 0.2 Iris-setosa  
## 2 4.9 3.0 1.4 0.2 Iris-setosa  
## 3 4.7 3.2 1.3 0.2 Iris-setosa  
## 4 4.6 3.1 1.5 0.2 Iris-setosa  
## 5 5.0 3.6 1.4 0.2 Iris-setosa  
## 6 5.4 3.9 1.7 0.4 Iris-setosa

summary(dfi)

## Sepal.Length Sepal.Width Petal.Length Petal.Width   
## Min. :4.300 Min. :2.000 Min. :1.000 Min. :0.100   
## 1st Qu.:5.100 1st Qu.:2.800 1st Qu.:1.600 1st Qu.:0.300   
## Median :5.800 Median :3.000 Median :4.350 Median :1.300   
## Mean :5.843 Mean :3.054 Mean :3.759 Mean :1.199   
## 3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800   
## Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500   
## Species   
## Iris-setosa :50   
## Iris-versicolor:50   
## Iris-virginica :50   
##   
##   
##

nrow(dfi)

## [1] 150

ncol(dfi)

## [1] 5

colnames(dfi)

## [1] "Sepal.Length" "Sepal.Width" "Petal.Length" "Petal.Width"   
## [5] "Species"

res <- sapply(dfi, class)  
kable(data.frame(variables=names(res),clase=as.vector(res)))

|  |  |
| --- | --- |
| variables | clase |
| Sepal.Length | numeric |
| Sepal.Width | numeric |
| Petal.Length | numeric |
| Petal.Width | numeric |
| Species | factor |

set.seed(20)  
irisCluster <- kmeans(dfi[, 3:4], 3, nstart = 20)  
irisCluster

## K-means clustering with 3 clusters of sizes 50, 52, 48  
##   
## Cluster means:  
## Petal.Length Petal.Width  
## 1 1.464000 0.244000  
## 2 4.269231 1.342308  
## 3 5.595833 2.037500  
##   
## Clustering vector:  
## [1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1  
## [36] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [71] 2 2 2 2 2 2 2 3 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 3 3 3 3 3  
## [106] 3 2 3 3 3 3 3 3 3 3 3 3 3 3 2 3 3 3 3 3 3 2 3 3 3 3 3 3 3 3 3 3 3 2 3  
## [141] 3 3 3 3 3 3 3 3 3 3  
##   
## Within cluster sum of squares by cluster:  
## [1] 2.03840 13.05769 16.29167  
## (between\_SS / total\_SS = 94.3 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

table(irisCluster$cluster, dfi$Species)

##   
## Iris-setosa Iris-versicolor Iris-virginica  
## 1 50 0 0  
## 2 0 48 4  
## 3 0 2 46

a b c

#Sepal.Lenght and Sepal.Width en vez de Petal (Pasamos un grupo a más que existe a ver que ocurre)  
irisCluster2 <- kmeans(dfi[, 1:2], 4, nstart = 30)  
irisCluster2

## K-means clustering with 4 clusters of sizes 28, 41, 28, 53  
##   
## Cluster means:  
## Sepal.Length Sepal.Width  
## 1 5.232143 3.667857  
## 2 6.880488 3.097561  
## 3 4.782143 2.950000  
## 4 5.924528 2.750943  
##   
## Clustering vector:  
## [1] 1 3 3 3 1 1 3 1 3 3 1 3 3 3 1 1 1 1 1 1 1 1 1 1 3 3 1 1 1 3 3 1 1 1 3  
## [36] 3 1 3 3 1 1 3 3 1 1 3 1 3 1 3 2 2 2 4 2 4 2 3 2 3 3 4 4 4 4 2 4 4 4 4  
## [71] 4 4 4 4 4 2 2 2 4 4 4 4 4 4 4 4 2 4 4 4 4 4 4 3 4 4 4 4 3 4 2 4 2 4 2  
## [106] 2 3 2 2 2 2 4 2 4 4 2 2 2 2 4 2 4 2 4 2 2 4 4 4 2 2 2 4 4 4 2 2 2 4 2  
## [141] 2 2 4 2 2 2 4 2 4 4  
##   
## Within cluster sum of squares by cluster:  
## [1] 3.902143 10.634146 5.151071 8.250566  
## (between\_SS / total\_SS = 78.5 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss"   
## [5] "tot.withinss" "betweenss" "size" "iter"   
## [9] "ifault"

table(irisCluster2$cluster, dfi$Species)

##   
## Iris-setosa Iris-versicolor Iris-virginica  
## 1 28 0 0  
## 2 0 11 30  
## 3 22 5 1  
## 4 0 34 19