Cluster Viajeros

Alejandro Gómez de Miguel

10/11/2017

Importación de datos

```
setwd("~/Documents/CUNEF - Data Science/1º Semestre/Técnicas de Agrupación y Reducción de la Dimensión/
```

Librerias:

```
library(cluster)
```

```
## Warning: package 'cluster' was built under R version 3.2.5
```

```
library(dendextend)
```

```
## Warning: package 'dendextend' was built under R version 3.2.5
```

```
## Warning: replacing previous import by 'magrittr::%>%' when loading
## 'dendextend'
```

```
##
## ---------------------
## Welcome to dendextend version 1.5.2
## Type citation('dendextend') for how to cite the package.
##
## Type browseVignettes(package = 'dendextend') for the package vignette.
## The github page is: https://github.com/talgalili/dendextend/
##
## Suggestions and bug-reports can be submitted at: https://github.com/talgalili/dendextend/issues
## Or contact: <tal.galili@gmail.com>
##
##  To suppress this message use:  suppressPackageStartupMessages(library(dendextend))
## ---------------------
```

```
##
## Attaching package: 'dendextend'
```

```
## The following object is masked from 'package:stats':
##
##     cutree
```

```
library(fpc)
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.2.5
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
library(NbClust)
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.2.5
```

```
require(ggplot2)
```

```
viajeros <- read_csv("~/Documents/CUNEF - Data Science/1º Semestre/Técnicas de Agrupación y Reducción d
```

```
## Warning: Missing column names filled in: 'X1' [1]
```

```
## Parsed with column specification:
## cols(
##   .default = col_integer(),
##   PAIS_RESID_AGRUP = col_character(),
##   ALOJ_CATEG_1 = col_character(),
##   SEXO = col_character(),
##   OCUPACION = col_character(),
##   INGRESOS = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
viajeros_df <- data.frame(viajeros)
copia.viajeros_df <- viajeros_df
```

Primera columna como índice:

```
viajeros_df <- data.frame(viajeros_df[,-1], row.names = viajeros_df[,1])
```

Tratamiento de los datos

Número total de NA:

```
sum(is.na(viajeros_df))
```

```
## [1] 563771
```

Si de 50.000 obervaciones hay más de un 70% con valores ausentes, no resulta siginificativa según el criterio que se propone.

```
colSums(is.na(viajeros_df)) > 35000
```

```
##             PAIS_RESID_AGRUP                 ALOJ_CATEG_1
##                        FALSE                        FALSE
##                    IMPRESION              VALORACION_ALOJ
##                        FALSE                        FALSE
##        VALORACION_TRATO_ALOJ     VALORACION_GASTRONO_ALOJ
##                        FALSE                        FALSE
##             VALORACION_CLIMA       VALORACION_ZONAS_BANYO
##                        FALSE                        FALSE
##          VALORACION_PAISAJES    VALORACION_MEDIO_AMBIENTE
##                        FALSE                        FALSE
##      VALORACION_TRANQUILIDAD          VALORACION_LIMPIEZA
##                        FALSE                        FALSE
##   VALORACION_CALIDAD_RESTAUR  VALORACION_OFERTA_GASTR_LOC
##                        FALSE                        FALSE
##     VALORACION_TRATO_RESTAUR    VALORACION_PRECIO_RESTAUR
##                        FALSE                        FALSE
##           VALORACION_CULTURA          VALORACION_DEPORTES
##                        FALSE                        FALSE
##              VALORACION_GOLF      VALORACION_PARQUES_OCIO
```

```
##                          TRUE                          TRUE
## VALORACION_AMBIENTE_NOCTURNO       VALORACION_EXCURSIONES
##                         FALSE                         FALSE
##     VALORACION_RECREO_NINYOS             VALORACION_SALUD
##                          TRUE                          TRUE
##     VALORACION_SERVICIOS_BUS    VALORACION_SERVICIOS_TAXI
##                         FALSE                         FALSE
##         VALORACION_ALQ_VEHIC         VALORACION_SEGURIDAD
##                         FALSE                         FALSE
## VALORACION_ESTADO_CARRETERAS  VALORACION_CALIDAD_COMERCIO
##                         FALSE                         FALSE
##      VALORACION_HOSPITALIDAD                          SEXO
##                         FALSE                         FALSE
##                          EDAD                     OCUPACION
##                         FALSE                         FALSE
##                      INGRESOS
##                         FALSE
```

Eliminamos las columnas con más NAs. Estas columnas no aportan suficiente información como para ser relevantes en la clasificación de individuos y pueden distorsionar los resultados.

VALORACION_PARQUES_OCIO, VALORACION_SALUD, VALORACION_GOLF y VALORACION_RECREO_NINYOS están fuera del dataset.

```
viajeros_df <- viajeros_df[, -c(1:2)] # Residencia y alojamiento
viajeros_df <- viajeros_df[, -22] # Salud
viajeros_df <- viajeros_df[, -17] # Golf
viajeros_df <- viajeros_df[, -17] #Parques
viajeros_df <- viajeros_df[, -19] # Recreo
viajeros_df <- viajeros_df[, -28] # Ocupacion
viajeros_df <- viajeros_df[, -26] # Sexo
```

Nuevo dataframe eliminando las filas donde existen valores ausentes:

```
viajeros_NoNA <- na.omit(viajeros_df)
```

Bucle para cuantificar los ingresos y que no sea un rango. Se realiza la media entre el máximo y el mínimo.

```
for(i in 1:nrow(viajeros_NoNA)) {

  if (viajeros_NoNA[i, 27] == 'De 60001 a 72000') {
    viajeros_NoNA[i, 27] = (60001+72000)/2 }

  if (viajeros_NoNA[i, 27] == 'De 12000 a 24000') {
    viajeros_NoNA[i, 27] = 18000 }

  if (viajeros_NoNA[i, 27] == 'De 48001 a 60000') {
    viajeros_NoNA[i, 27] = (48001 + 60000)/2 }

  if (viajeros_NoNA[i, 27] == 'De 36001 a 48000') {
    viajeros_NoNA[i, 27] = (36001 + 48000)/2 }

  if (viajeros_NoNA[i, 27] == 'Más de 84000') {
    viajeros_NoNA[i, 27] = 84000 }

  if (viajeros_NoNA[i, 27] == 'De 24001 a 36000') {
    viajeros_NoNA[i, 27] = (24001 + 36000)/2 }
```

```
  if (viajeros_NoNA[i, 27] == 'De 72001 a 84000') {
    viajeros_NoNA[i, 27] = (72001 + 84000)/2 }

}
```

Reducción de la dimensión

Alojamiento

La media de tres variables que representarán una sola:

```
ALOJAMIENTO <- (viajeros_NoNA$VALORACION_ALOJ +
                  viajeros_NoNA$VALORACION_TRATO_ALOJ +
                  viajeros_NoNA$VALORACION_GASTRONO_ALOJ)/3
```

Un sólo decimal:

```
ALOJAMIENTO <- round(ALOJAMIENTO, 1)
```

Se une al dataset con el que trabajamos:

```
viajeros_NoNA <- cbind(viajeros_NoNA, ALOJAMIENTO)
```

Eliminamos las columnas una vez se ha reducido la dimensión.

```
viajeros_NoNA <- viajeros_NoNA[, -c(2:4)]
```

Restaurantes

La media de tres variables que representarán una sola:

```
RESTAURANTES <- (viajeros_NoNA$VALORACION_CALIDAD_RESTAUR +
                   viajeros_NoNA$VALORACION_TRATO_RESTAUR +
                   viajeros_NoNA$VALORACION_PRECIO_RESTAUR)/3
```

Un sólo decimal:

```
RESTAURANTES <- round(RESTAURANTES, 1)
```

Se une al dataset con el que trabajamos:

```
viajeros_NoNA <- cbind(viajeros_NoNA, RESTAURANTES)
```

Eliminamos las columnas una vez se ha reducido la dimensión.

```
viajeros_NoNA <- viajeros_NoNA[, -8]
viajeros_NoNA <- viajeros_NoNA[, -c(9:10)]
```

De las 36 variables iniciales, ahora disponemos de 23 con 4411 observaciones.

Metodo CLARA

```
viajeros_limpio <- viajeros_NoNA
```

```
sapply(viajeros_limpio, mode)
```

```
##                     IMPRESION              VALORACION_CLIMA
##                     "numeric"                     "numeric"
##       VALORACION_ZONAS_BANYO           VALORACION_PAISAJES
##                     "numeric"                     "numeric"
##    VALORACION_MEDIO_AMBIENTE       VALORACION_TRANQUILIDAD
##                     "numeric"                     "numeric"
##          VALORACION_LIMPIEZA  VALORACION_OFERTA_GASTR_LOC
##                     "numeric"                     "numeric"
##           VALORACION_CULTURA           VALORACION_DEPORTES
##                     "numeric"                     "numeric"
## VALORACION_AMBIENTE_NOCTURNO       VALORACION_EXCURSIONES
##                     "numeric"                     "numeric"
##     VALORACION_SERVICIOS_BUS    VALORACION_SERVICIOS_TAXI
##                     "numeric"                     "numeric"
##         VALORACION_ALQ_VEHIC         VALORACION_SEGURIDAD
##                     "numeric"                     "numeric"
## VALORACION_ESTADO_CARRETERAS  VALORACION_CALIDAD_COMERCIO
##                     "numeric"                     "numeric"
##      VALORACION_HOSPITALIDAD                          EDAD
##                     "numeric"                     "numeric"
##                      INGRESOS                  ALOJAMIENTO
##                   "character"                     "numeric"
##                  RESTAURANTES
##                     "numeric"
```

Ingresos es character y no vamos a poder tipificar, es necesario cambiarlo.

```
viajeros_limpio <- transform(viajeros_limpio, INGRESOS  = as.numeric(INGRESOS))
```

```
class(viajeros_limpio$INGRESOS) # comprobación
```

```
## [1] "numeric"
```

Se salva el dataset escalado.

```
viajeros.tip <- scale(viajeros_limpio)
```

```
summary(viajeros.tip)
```

```
##    IMPRESION       VALORACION_CLIMA  VALORACION_ZONAS_BANYO
##  Min.   :-3.7748   Min.   :-4.6192   Min.   :-3.87055
##  1st Qu.:-0.3105   1st Qu.:-0.3406   1st Qu.:-0.59495
##  Median :-0.3105   Median : 0.2706   Median :-0.04901
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.8443   3rd Qu.: 0.8818   3rd Qu.: 1.04286
##  Max.   : 0.8443   Max.   : 0.8818   Max.   : 1.04286
##  VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD
##  Min.   :-4.0179     Min.   :-4.02608          Min.   :-3.81451
##  1st Qu.:-0.6964     1st Qu.:-0.62092          1st Qu.:-0.59570
##  Median : 0.4108     Median :-0.05339          Median :-0.05923
##  Mean   : 0.0000     Mean   : 0.00000          Mean   : 0.00000
##  3rd Qu.: 0.9643     3rd Qu.: 0.51413          3rd Qu.: 1.01371
##  Max.   : 0.9643     Max.   : 1.08166          Max.   : 1.01371
##  VALORACION_LIMPIEZA VALORACION_OFERTA_GASTR_LOC VALORACION_CULTURA
##  Min.   :-3.708785   Min.   :-3.2484             Min.   :-2.99751
##  1st Qu.:-0.521465   1st Qu.:-0.7045             1st Qu.:-0.54835
##  Median : 0.009755   Median : 0.3130             Median :-0.05852
```

```
##  Mean   : 0.000000   Mean   : 0.0000             Mean   : 0.00000
##  3rd Qu.: 0.540975   3rd Qu.: 0.8218             3rd Qu.: 0.43131
##  Max.   : 1.072195   Max.   : 1.3306             Max.   : 1.41097
##  VALORACION_DEPORTES VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES
##  Min.   :-3.3575     Min.   :-2.8861              Min.   :-3.0628
##  1st Qu.:-0.7485     1st Qu.:-0.5566              1st Qu.:-0.6605
##  Median : 0.2950     Median : 0.3752              Median : 0.3004
##  Mean   : 0.0000     Mean   : 0.0000              Mean   : 0.0000
##  3rd Qu.: 0.8168     3rd Qu.: 0.8410              3rd Qu.: 0.7809
##  Max.   : 1.3386     Max.   : 1.3069              Max.   : 1.2613
##  VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI VALORACION_ALQ_VEHIC
##  Min.   :-3.1245          Min.   :-3.712298         Min.   :-3.3483
##  1st Qu.:-0.6966          1st Qu.:-0.529504         1st Qu.:-0.3550
##  Median : 0.2745          Median : 0.000962         Median : 0.1439
##  Mean   : 0.0000          Mean   : 0.000000         Mean   : 0.0000
##  3rd Qu.: 0.7601          3rd Qu.: 0.531428         3rd Qu.: 0.6427
##  Max.   : 1.2457          Max.   : 1.061893         Max.   : 1.1416
##  VALORACION_SEGURIDAD VALORACION_ESTADO_CARRETERAS
##  Min.   :-3.882943    Min.   :-3.2635
##  1st Qu.:-0.549740    1st Qu.:-0.3020
##  Median : 0.005793    Median : 0.1916
##  Mean   : 0.000000    Mean   : 0.0000
##  3rd Qu.: 0.561327    3rd Qu.: 0.6852
##  Max.   : 1.116861    Max.   : 1.1787
##  VALORACION_CALIDAD_COMERCIO VALORACION_HOSPITALIDAD      EDAD
##  Min.   :-3.2977             Min.   :-4.3367         Min.   :-1.7119
##  1st Qu.:-0.7362             1st Qu.:-0.1894         1st Qu.:-0.8106
##  Median : 0.2884             Median : 0.4031         Median :-0.1551
##  Mean   : 0.0000             Mean   : 0.0000         Mean   : 0.0000
##  3rd Qu.: 0.8007             3rd Qu.: 0.9956         3rd Qu.: 0.7463
##  Max.   : 1.3130             Max.   : 0.9956         Max.   : 5.0891
##     INGRESOS         ALOJAMIENTO        RESTAURANTES
##  Min.   :-1.13000   Min.   :-3.75778   Min.   :-4.3809
##  1st Qu.:-1.13000   1st Qu.:-0.47414   1st Qu.:-0.4724
##  Median :-0.08345   Median : 0.07313   Median : 0.1791
##  Mean   : 0.00000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.96309   3rd Qu.: 0.78458   3rd Qu.: 0.6351
##  Max.   : 1.74797   Max.   : 1.16767   Max.   : 1.4819
```

Se toma la muestra para estimar el numero de clusters.

```
set.seed(123)
viajeros.sample = viajeros.tip[sample(1:nrow(viajeros.tip), 1000, replace=FALSE),]
```

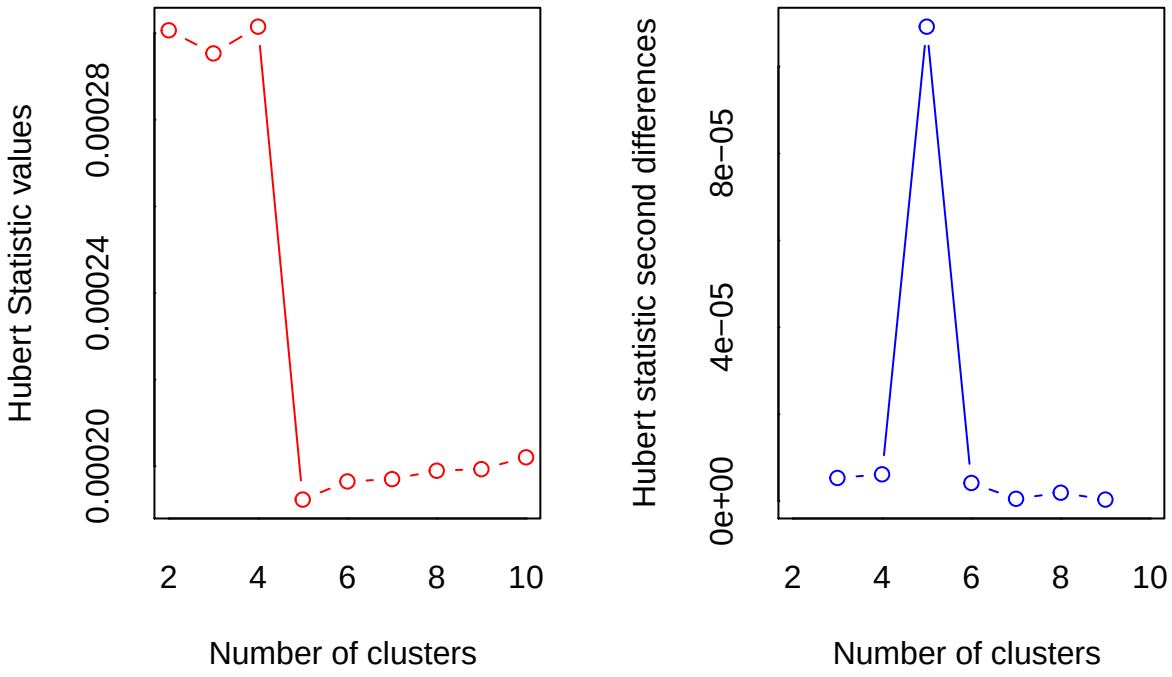
```
summary(viajeros.sample)
```

```
##    IMPRESION        VALORACION_CLIMA   VALORACION_ZONAS_BANYO
##  Min.   :-3.77479   Min.   :-4.61916   Min.   :-3.87055
##  1st Qu.:-0.31049   1st Qu.:-0.34060   1st Qu.:-0.59495
##  Median :-0.31049   Median : 0.27062   Median :-0.04901
##  Mean   :-0.02988   Mean   : 0.01635   Mean   : 0.03288
##  3rd Qu.: 0.84428   3rd Qu.: 0.88184   3rd Qu.: 1.04286
##  Max.   : 0.84428   Max.   : 0.88184   Max.   : 1.04286
##  VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD
##  Min.   :-4.01790    Min.   :-4.02608          Min.   :-3.8145
```

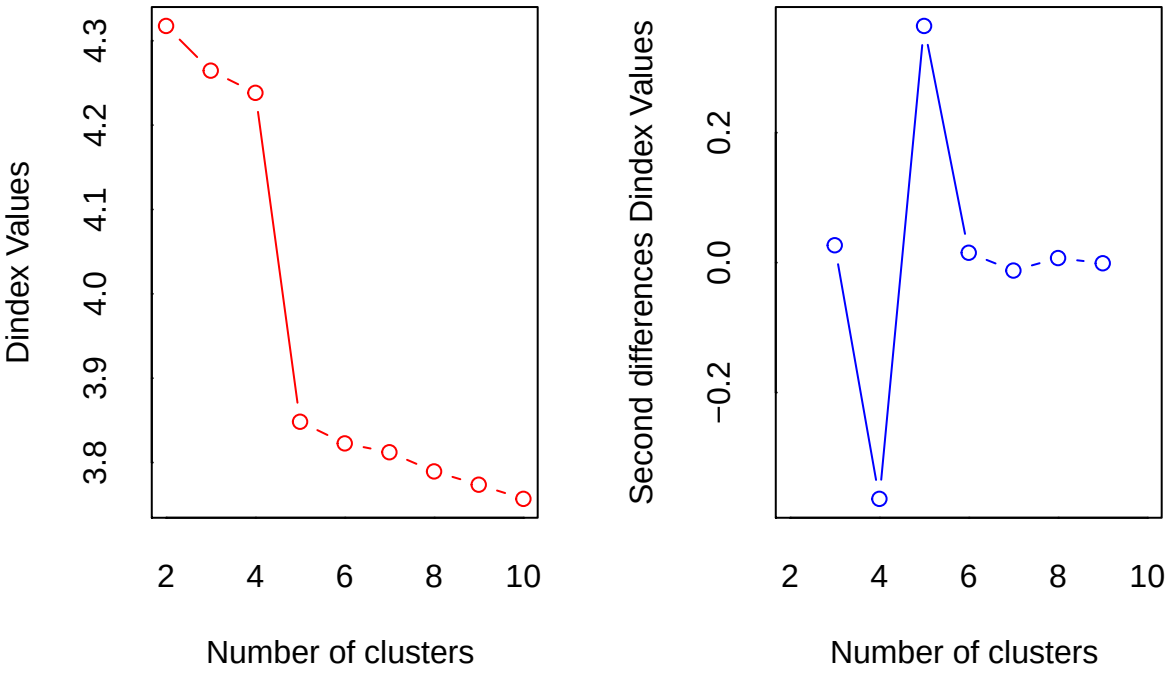
```
##  1st Qu.:-0.69640    1st Qu.:-0.62092          1st Qu.:-0.5957
##  Median : 0.41076    Median :-0.05339          Median : 0.4772
##  Mean   : 0.02713    Mean   : 0.01868          Mean   : 0.0604
##  3rd Qu.: 0.96435    3rd Qu.: 0.65601          3rd Qu.: 1.0137
##  Max.   : 0.96435    Max.   : 1.08166          Max.   : 1.0137
##  VALORACION_LIMPIEZA VALORACION_OFERTA_GASTR_LOC VALORACION_CULTURA
##  Min.   :-3.708785   Min.   :-3.2484             Min.   :-2.99751
##  1st Qu.:-0.521465   1st Qu.:-0.7045             1st Qu.:-0.54835
##  Median : 0.009755   Median : 0.3130             Median :-0.05852
##  Mean   : 0.032066   Mean   : 0.0556             Mean   :-0.01444
##  3rd Qu.: 0.540975   3rd Qu.: 0.8218             3rd Qu.: 0.43131
##  Max.   : 1.072195   Max.   : 1.3306             Max.   : 1.41097
##  VALORACION_DEPORTES VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES
##  Min.   :-3.357465   Min.   :-2.88606             Min.   :-3.06285
##  1st Qu.:-0.748548   1st Qu.:-0.55662             1st Qu.:-0.66052
##  Median : 0.295019   Median : 0.37516             Median : 0.30041
##  Mean   :-0.008137   Mean   : 0.02015             Mean   : 0.02655
##  3rd Qu.: 0.816802   3rd Qu.: 0.84105             3rd Qu.: 0.78088
##  Max.   : 1.338585   Max.   : 1.30694             Max.   : 1.26134
##  VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI VALORACION_ALQ_VEHIC
##  Min.   :-3.124458        Min.   :-3.712298         Min.   :-3.34831
##  1st Qu.:-0.696599        1st Qu.:-0.529504         1st Qu.:-0.35502
##  Median : 0.274545        Median : 0.000962         Median : 0.14386
##  Mean   :-0.003203        Mean   :-0.002221         Mean   : 0.01166
##  3rd Qu.: 0.760116        3rd Qu.: 0.531428         3rd Qu.: 0.64274
##  Max.   : 1.245688        Max.   : 1.061893         Max.   : 1.14162
##  VALORACION_SEGURIDAD VALORACION_ESTADO_CARRETERAS
##  Min.   :-3.882943    Min.   :-3.26352
##  1st Qu.:-0.549740    1st Qu.:-0.30201
##  Median : 0.005793    Median : 0.19157
##  Mean   : 0.028570    Mean   : 0.02869
##  3rd Qu.: 0.561327    3rd Qu.: 0.68515
##  Max.   : 1.116861    Max.   : 1.17874
##  VALORACION_CALIDAD_COMERCIO VALORACION_HOSPITALIDAD      EDAD
##  Min.   :-3.29770            Min.   :-4.3367         Min.   :-1.71193
##  1st Qu.:-0.73622            1st Qu.:-0.1894         1st Qu.:-0.81059
##  Median : 0.28838            Median : 0.4031         Median :-0.23701
##  Mean   : 0.04401            Mean   : 0.0399         Mean   :-0.04224
##  3rd Qu.: 0.80068            3rd Qu.: 0.9956         3rd Qu.: 0.66432
##  Max.   : 1.31297            Max.   : 0.9956         Max.   : 5.08906
##     INGRESOS         ALOJAMIENTO        RESTAURANTES
##  Min.   :-1.13000   Min.   :-3.75778   Min.   :-4.38087
##  1st Qu.:-1.13000   1st Qu.:-0.47414   1st Qu.:-0.47236
##  Median :-0.08345   Median : 0.07313   Median : 0.17906
##  Mean   :-0.01674   Mean   : 0.02256   Mean   : 0.03132
##  3rd Qu.: 0.96309   3rd Qu.: 0.78458   3rd Qu.: 0.83048
##  Max.   : 1.74797   Max.   : 1.16767   Max.   : 1.48190
```

Se estima el cluster con distancia euclídea, donde el numero mínimo de clusters será 2 y el máximo 10. Mediante esta función obtendremos el numero de clusters más apropiado.

```
Nb.viajeros = NbClust(viajeros.sample, distance = "euclidean", min.nc = 2,
                      max.nc = 10, method = "complete", index ="all")
```




```
## *** : The Hubert index is a graphical method of determining the number of clusters.
##                 In the plot of Hubert index, we seek a significant knee that corresponds to a
##                 significant increase of the value of the measure i.e the significant peak in Hubert
##                 index second differences plot.
##
```



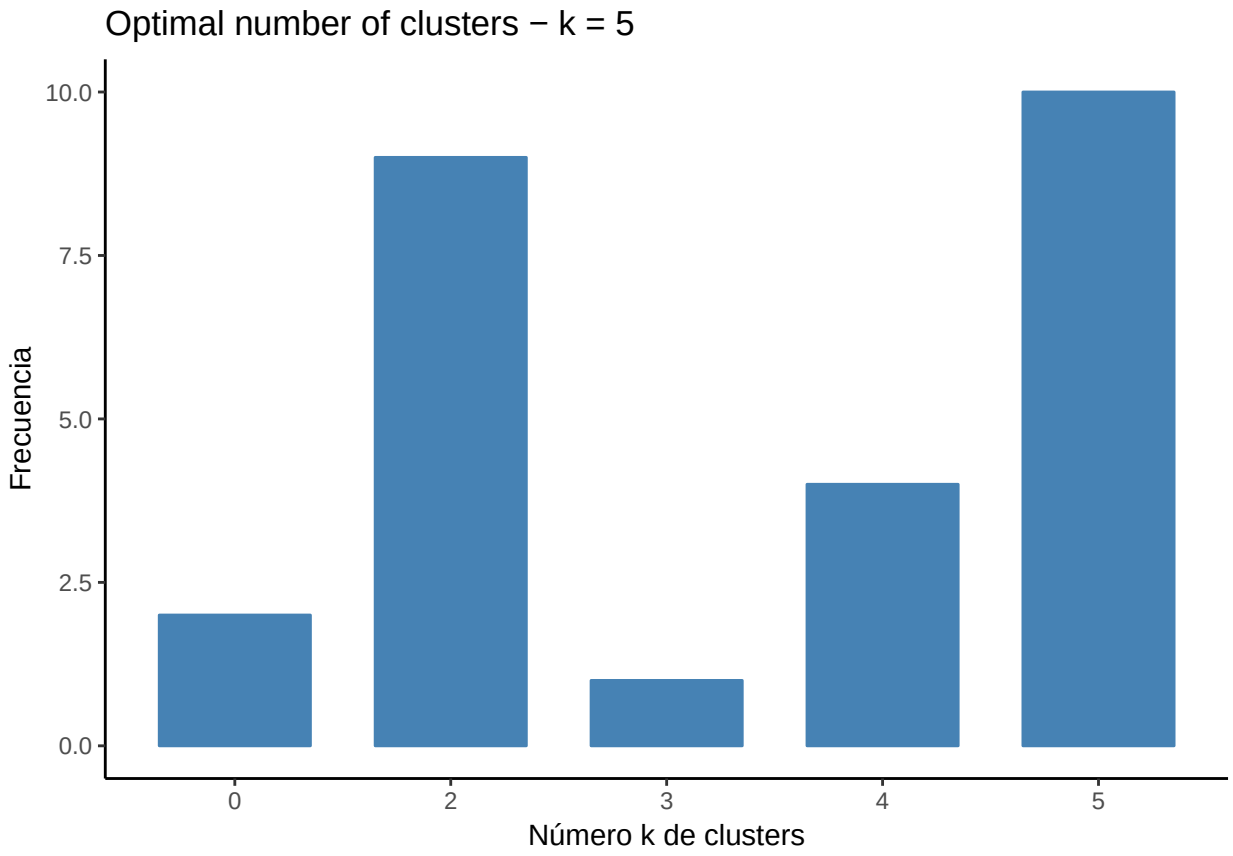

```
## *** : The D index is a graphical method of determining the number of clusters.
##                 In the plot of D index, we seek a significant knee (the significant peak in Dindex
##                 second differences plot) that corresponds to a significant increase of the value of
##                 the measure.
##
```

```
## *****************************************************************
## * Among all indices:
## * 9 proposed 2 as the best number of clusters
## * 1 proposed 3 as the best number of clusters
## * 4 proposed 4 as the best number of clusters
## * 10 proposed 5 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  5
##
##
## *****************************************************************
```

Gráficamente:

```
fviz_nbclust(Nb.viajeros) + theme_classic() + labs(x = "Número k de clusters",
                                                   y = "Frecuencia")
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 9 proposed  2 as the best number of clusters
## * 1 proposed  3 as the best number of clusters
## * 4 proposed  4 as the best number of clusters
## * 10 proposed  5 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  5 .
```



Recomendacion de 5 clusters, aplicación sobre todas las observaciones:

```
viajeros.clara <- clara(viajeros_limpio, 5, samples = 200)
```

Representación del cluster en 2 dimensiones:

```
fviz_cluster(viajeros.clara, stand = TRUE, geom = "point", pointsize = 1)
```

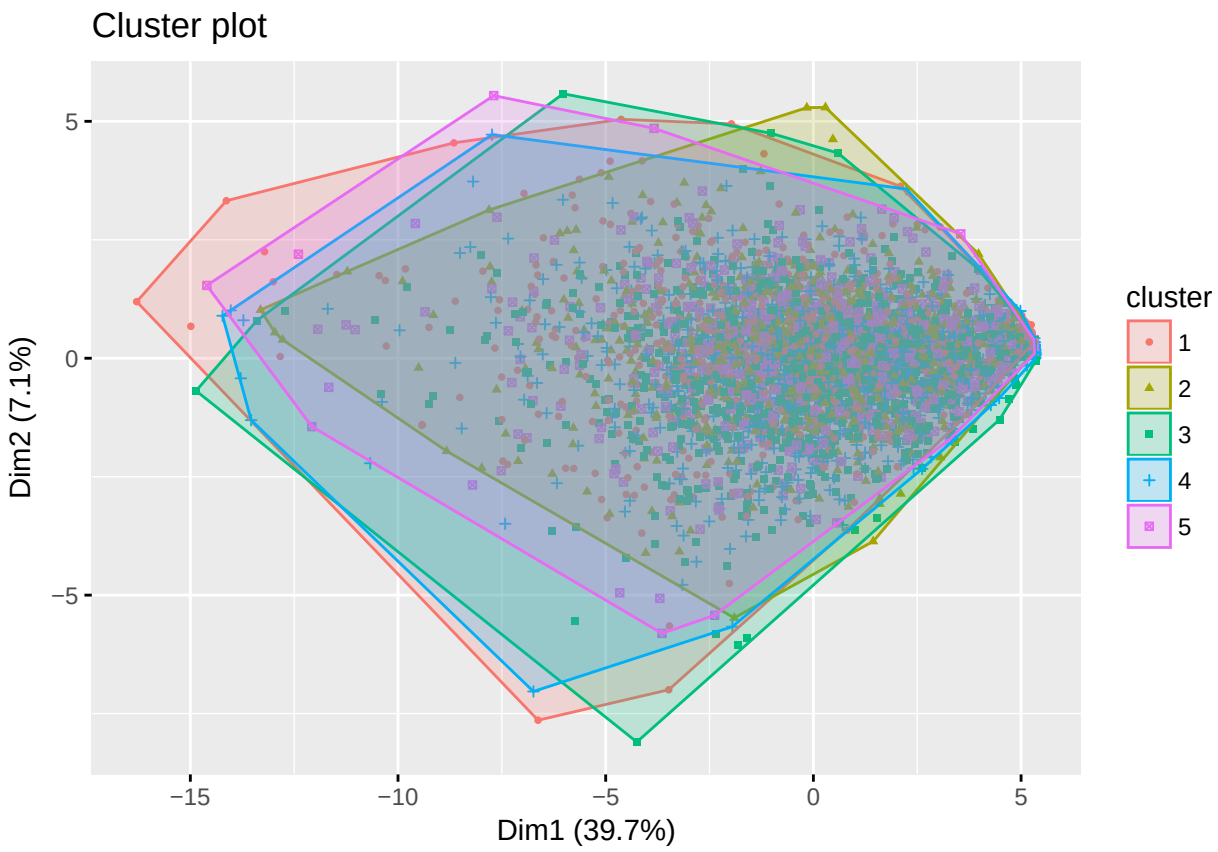
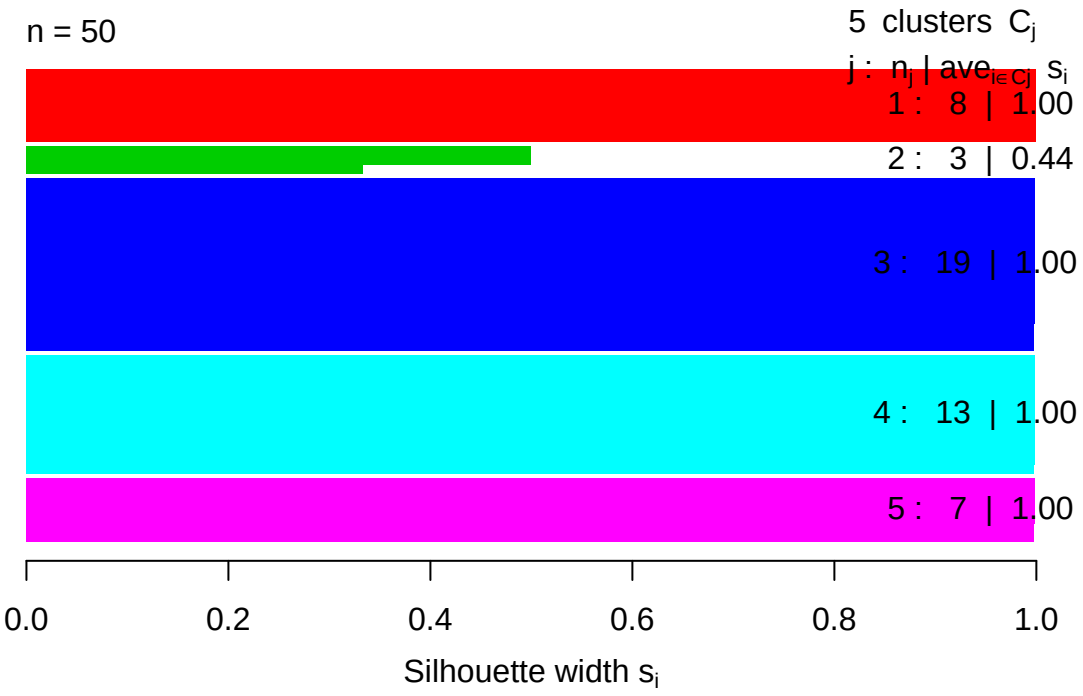


Gráfico de perfil:

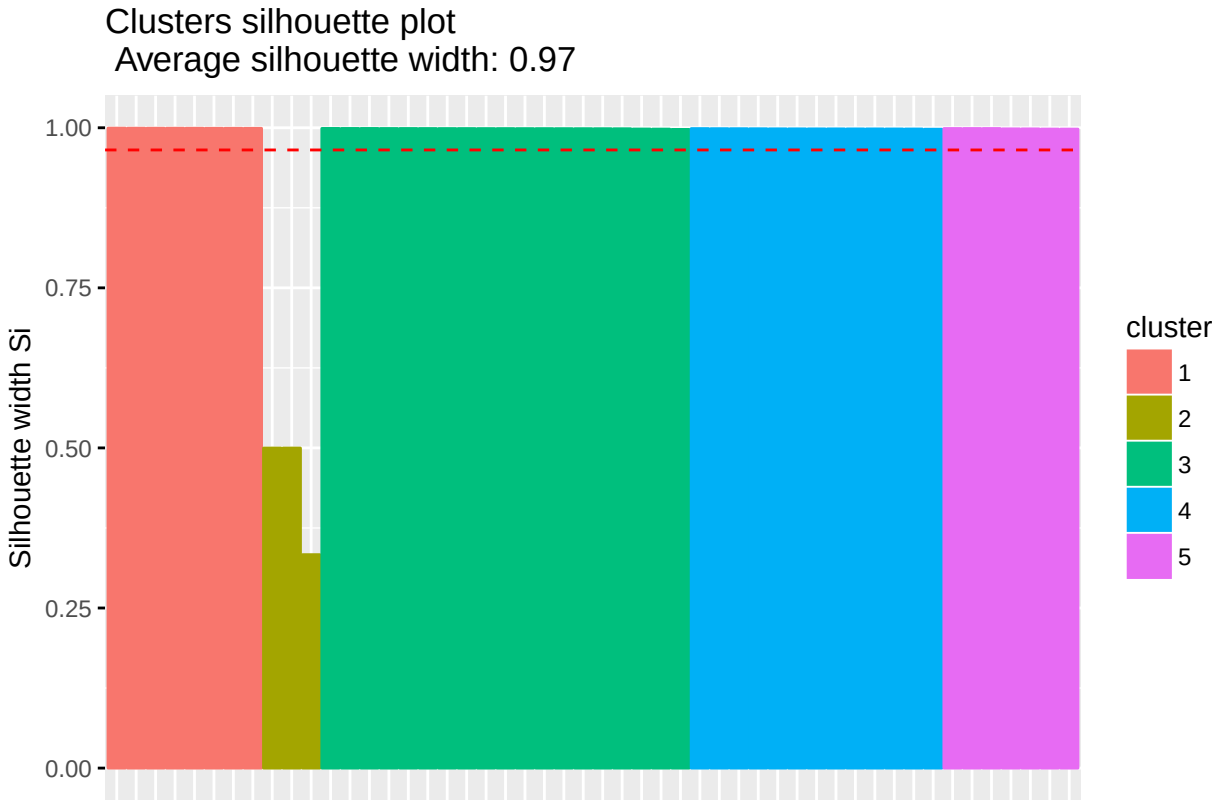
```
plot(silhouette(viajeros.clara), col = 2:6, main = "Gráfico de perfil")
```



El gráfico de perfil señala una media de silueta de 0.97.

```
fviz_silhouette(viajeros.clara)
```

```
##   cluster size ave.sil.width
## 1       1    8          1.00
## 2       2    3          0.44
## 3       3   19          1.00
## 4       4   13          1.00
## 5       5    7          1.00
```



Aparentemente, todos los individuos se encuentran bien clasificados en sus cluster a excepción del cluster número 2, donde la media de la silueta es 0.44 y debería ser igual o superior a 0.97, como en el caso de los otros 4 clusters.

Interpretación

Unimos los resultados del cluster a las variables categóricas:

```
copia.idex.viajeros_df <- data.frame(copia.viajeros_df[,-1], row.names = copia.viajeros_df[,1])
```

Se accede a las filas de los individuos por índice ('nombre de las columnas'):

```
head(copia.idex.viajeros_df[as.character(rownames(viajeros_limpio)),])
```

```
##        PAIS_RESID_AGRUP                            ALOJ_CATEG_1 IMPRESION
## 242037      Reino Unido Hoteles - apartahoteles de 4 estrellas         1
## 161764      Reino Unido                          Extrahoteleros         5
## 228332            Otros Hoteles - apartahoteles de 4 estrellas         4
## 146449           España Hoteles - apartahoteles de 4 estrellas         4
## 219486      Reino Unido                          Extrahoteleros         4
## 254647           España Hoteles - apartahoteles de 4 estrellas         4
##        VALORACION_ALOJ VALORACION_TRATO_ALOJ VALORACION_GASTRONO_ALOJ
## 242037               7                     7                        1
## 161764              10                     7                       10
## 228332               7                     9                        6
## 146449               9                     9                        9
## 219486               8                     7                        9
## 254647              10                    10                        9
##        VALORACION_CLIMA VALORACION_ZONAS_BANYO VALORACION_PAISAJES
```

```
## 242037                8                     8                 10
## 161764                9                    10                 10
## 228332               10                     8                  9
## 146449               10                     9                 10
## 219486               10                     9                  9
## 254647               10                     9                  8
##        VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD
## 242037                        10                      9
## 161764                        10                     10
## 228332                         8                      6
## 146449                         9                      8
## 219486                         9                      5
## 254647                         9                     10
##        VALORACION_LIMPIEZA VALORACION_CALIDAD_RESTAUR
## 242037                   7                        10
## 161764                   5                        10
## 228332                   6                         8
## 146449                   9                         6
## 219486                   9                        10
## 254647                  10                         9
##        VALORACION_OFERTA_GASTR_LOC VALORACION_TRATO_RESTAUR
## 242037                           7                      10
## 161764                          10                      10
## 228332                           8                       9
## 146449                           8                       4
## 219486                           8                       9
## 254647                           8                       9
##        VALORACION_PRECIO_RESTAUR VALORACION_CULTURA VALORACION_DEPORTES
## 242037                         8                 7                   7
## 161764                        10                 1                   1
## 228332                        10                 1                  10
## 146449                         5                 7                   6
## 219486                         9                 5                   5
## 254647                         9                 8                   8
##        VALORACION_GOLF VALORACION_PARQUES_OCIO
## 242037               3                      10
## 161764               1                      10
## 228332              10                       8
## 146449               7                       6
## 219486              10                       5
## 254647               9                       9
##        VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES
## 242037                           10                      8
## 161764                           10                     10
## 228332                            8                      6
## 146449                            8                      7
## 219486                            8                      5
## 254647                            8                      9
##        VALORACION_RECREO_NINYOS VALORACION_SALUD VALORACION_SERVICIOS_BUS
## 242037                        8                8                       3
## 161764                        1                1                      10
## 228332                        9                4                       8
## 146449                        6                7                       8
## 219486                        9                5                       8
```

```
## 254647                        9                9                       10
##        VALORACION_SERVICIOS_TAXI VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
## 242037                         5                   5                    7
## 161764                        10                  10                   10
## 228332                         9                   6                    7
## 146449                         7                   6                    8
## 219486                         8                  10                    8
## 254647                         9                   9                    9
##        VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
## 242037                            5                          7
## 161764                           10                         10
## 228332                            4                          8
## 146449                            5                          8
## 219486                           10                         10
## 254647                            7                          9
##        VALORACION_HOSPITALIDAD   SEXO EDAD                    OCUPACION
## 242037                       7 Hombre   28 Autónomo - profesión liberal
## 161764                      10 Hombre   25    Asalariado alta dirección
## 228332                      10 Hombre   38                   Estudiante
## 146449                       9 Hombre   25 Otros trabajadores y obreros
## 219486                      10  Mujer   18 Otros trabajadores y obreros
## 254647                       9 Hombre   30 Autónomo - profesión liberal
##                INGRESOS
## 242037     Más de 84000
## 161764 De 60001 a 72000
## 228332     Más de 84000
## 146449 De 12000 a 24000
## 219486 De 12000 a 24000
## 254647 De 24001 a 36000
```

```
viajeros_index <- copia.idex.viajeros_df[as.character(rownames(viajeros_limpio)),]
```

Unimos ambos dataframes, de este modo referenciamos cada individuo con el cluster al que pertenece para poder comparar cada cluster con sus variables categóricas:

```
viajeros_postcluster <- cbind(viajeros_index, viajeros.clara$clustering)
```

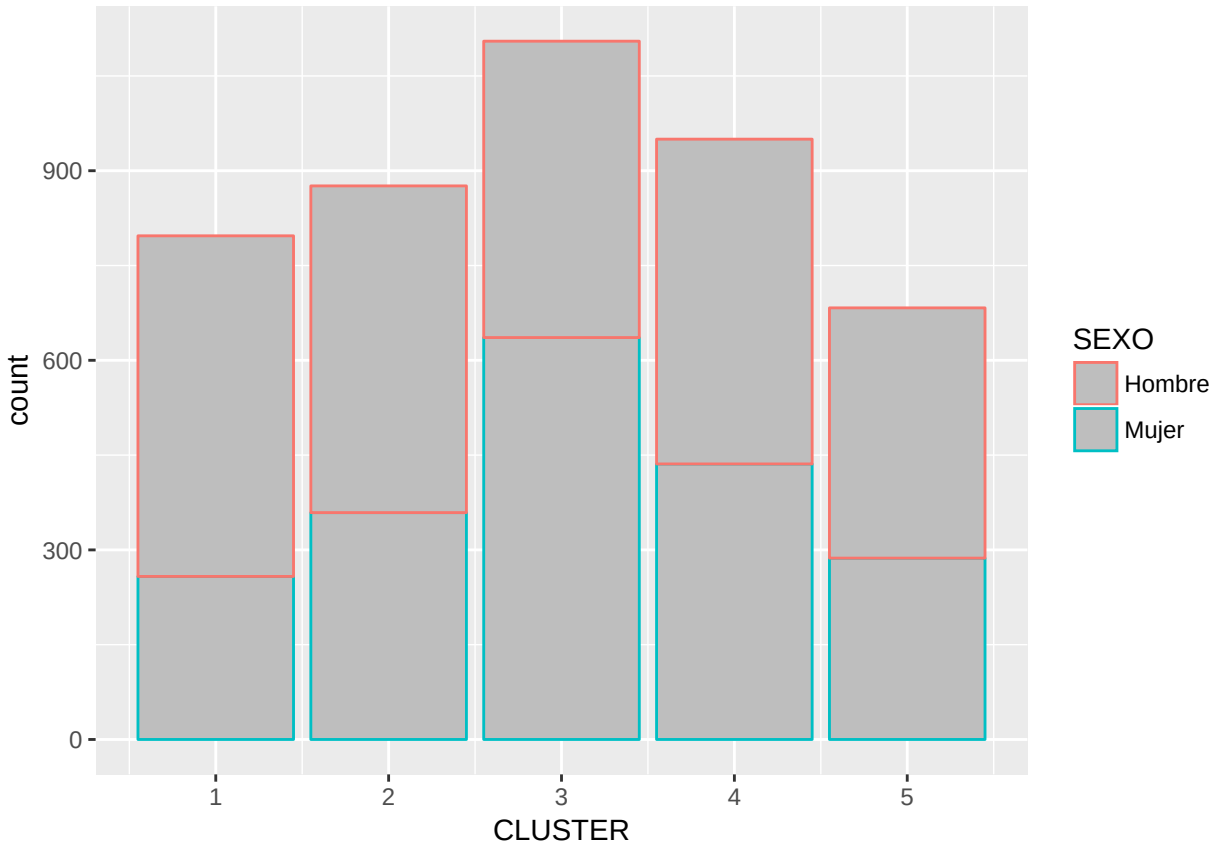
Vamos a quedarnos con las variables categóricas:

```
viajeros_postcluster <- viajeros_postcluster[, -c(3:31)]
```

```
names(viajeros_postcluster)[7]<-paste("CLUSTER") # Modificación del nombre de la columna
```

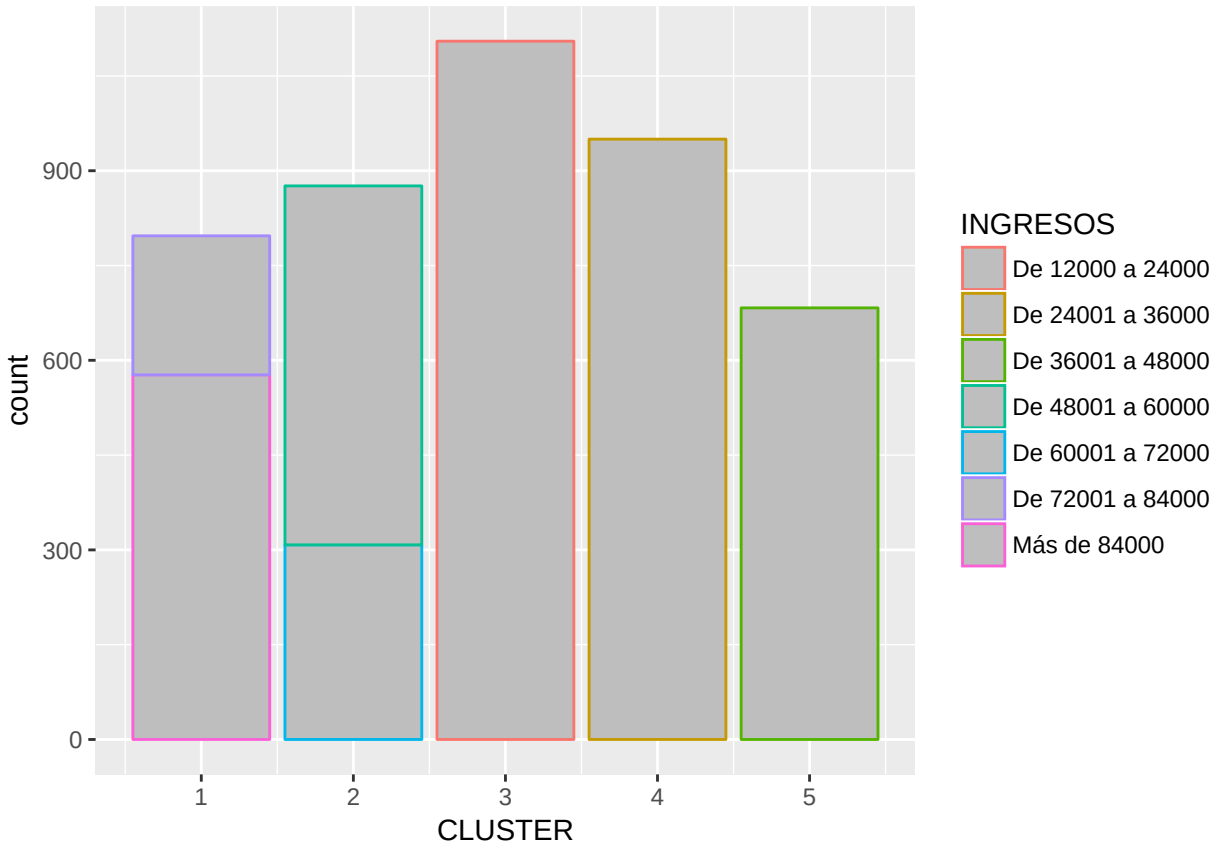
Gráficos

```
ggplot(viajeros_postcluster, mapping = aes(CLUSTER, col = SEXO)) + geom_bar(fill = 8)
```



El tercer cluster es el que más individuos agrupa. Existe una proporción de hombres y muejeres aparentemente similar en cada cluster.

```
ggplot(viajeros_postcluster, aes(x = CLUSTER, color = INGRESOS)) +
  geom_bar(fill = 8)
```



Cada cluster parece estar identificado con un intervalo de ingresos en particular, incluyendo:

El cluster número 1 los individuos con rentas mas altas (mas de 72.000)

El cluster número 2 los indiviudos con rentas medias-altas (de 48.001 a 72.000)

El cluster número 3 los indiviudos con rentas bajas (entre 12.000 y 24.000)

El cluster número 4 los individuos con rentas medias-bajas (de 24.001 a 36.000)

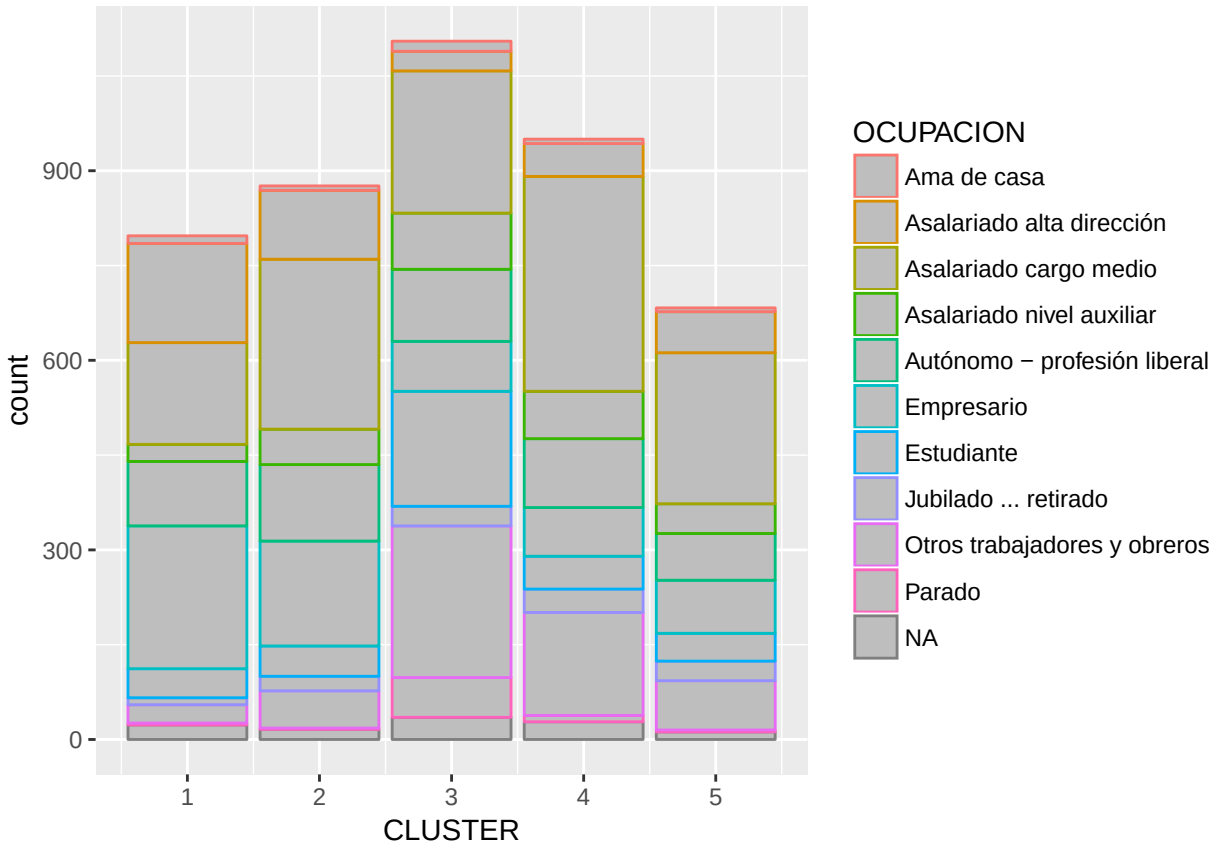
El cluster número 5 los individuos con rentas medias (de 36.001 a 48.000)

En este apartado influye que la variable ingresos fue cuantificada y utilizada para realizar el análisis.

```
ggplot(viajeros_postcluster, aes(x = CLUSTER, color = OCUPACION)) +
  geom_bar(fill = 8)
```

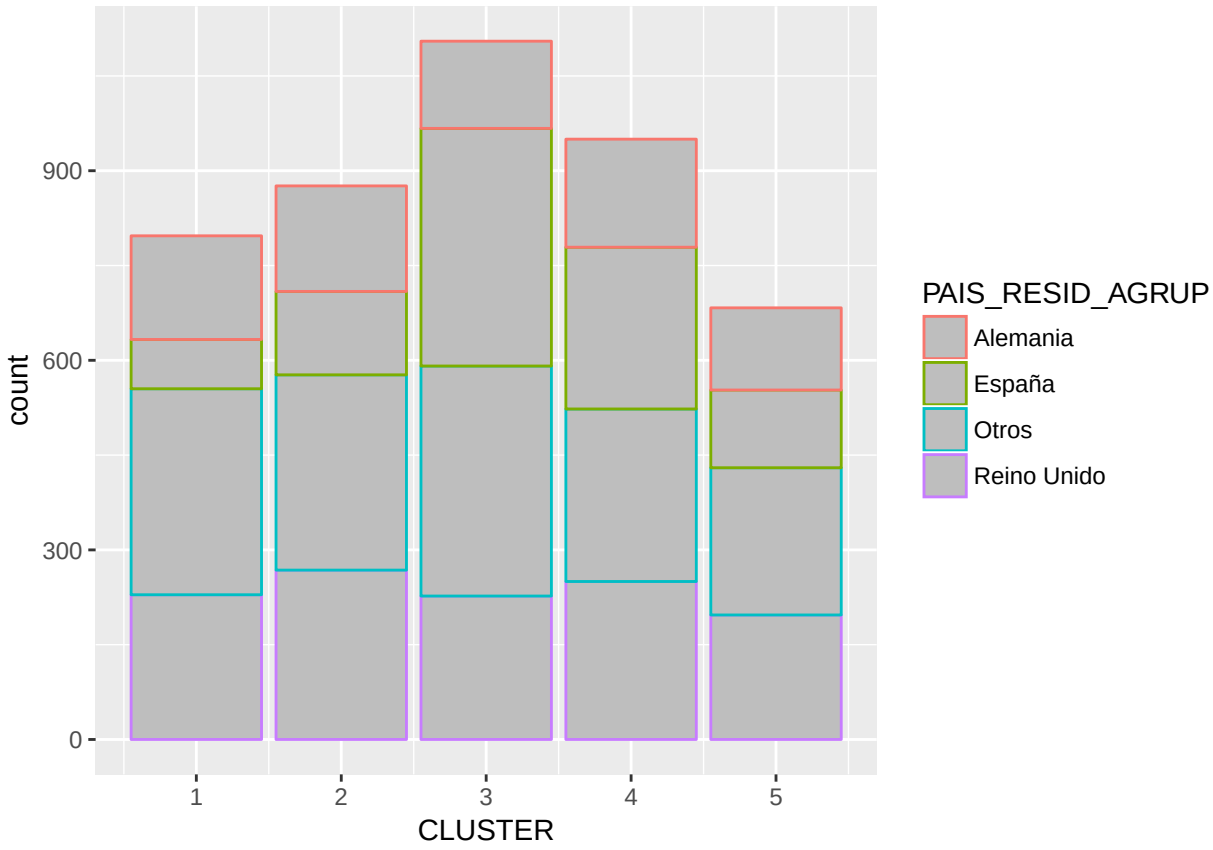
```
## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x1e

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): font
## metrics unknown for character 0x80

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(L_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot substituted for <93>

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x
```

```
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <93>

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call(L_textBounds, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <93>

## Warning in grid.Call.graphics(L_text, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <e2>

## Warning in grid.Call.graphics(L_text, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <80>

## Warning in grid.Call.graphics(L_text, as.graphicsAnnot(x$label), x$x, x
## $y, : conversion failure on 'Jubilado - retirado' in 'mbcsToSbcs': dot
## substituted for <93>
```



Clasificado por su ocupacion, la variedad es bastante grande por cada cluster.

```
ggplot(viajeros_postcluster, aes(x = CLUSTER, color = PAIS_RESID_AGRUP)) +
  geom_bar(fill = 8)
```



Aparentemente es bastante uniforme la composicion de cada cluster por país de residencia.

```
ggplot(viajeros_postcluster, aes(x = CLUSTER, color = ALOJ_CATEG_1)) +
  geom_bar(fill = 8)
```

Finalmente, en cuanto al alojamiento, parece que los clusters también se componen de distintos tipos de alojamiento, no se identifican clusters que pertenezcan a algún tipo de alojamiento específico.