clusterPaíses

El método empleado para éste análisis será el CLARA para clúster no jerárquicos, ya que la complejidad es muy elevada, teniendo miles de observaciones.

CLARA: Clustering LArge Applications

CLARA es un algoritmo para efectuar análisis cluster sobre grandes conjuntos de varios miles de observaciones; es importante notar que la caracetrística "grande" dependerá tanto de la RAM como de la velocidad de proceso. El algoritmo trabaja en la forma siguiente:

1. División del conjunto original en varios más pequeños de tamaño fijo;

2. Aplicar el algoritmo PAM en cada nuevo subconjunto escogiendo los k mediodides en cada uno y asignando cada observación del conjunto orginal al medioide más cercano;

3. Calcular la media o la suma de las disimilaridades de las observaciones respecto de su medioide más cercano, que se empleará como una medida de la bondad de ajuster de la segmentación;

4. Escoger como solución aquel subconjunto en la que la medida de la disimilaridad es mínima. Se lleva a cabo mediante la función clara() del paquete cluster. clara(x, k, samples=5) . x es la matriz de datos o el data frame, con variables por columnas y observaciones por filas; . k es el número de clusters; . samples es el número de subconjuntos en el que dividir el original; por defecto es 5 pero conviene emplear un valor mucho mayor.

```r
rm(list=ls())

library(cluster)

library(dendextend)

library(fpc)

library(factoextra)

library(NbClust)
library(rminer)  #para valores ausentes

library(dummies) #para poner variables dummy

setwd("C:/Users/usuario/Desktop/TecnicasZafra")
viajeros=read.csv("viajerosNA.csv", header=TRUE, sep=",")
#View(viajeros)
# hacemos que la primera columna del df se transforme en el nombre de las filas
X<-viajeros[,1]
viajeros = data.frame(viajeros[,-1], row.names=viajeros[,1])
```

El planteamiento que voy a seguir en el análisis clúster va a ser:

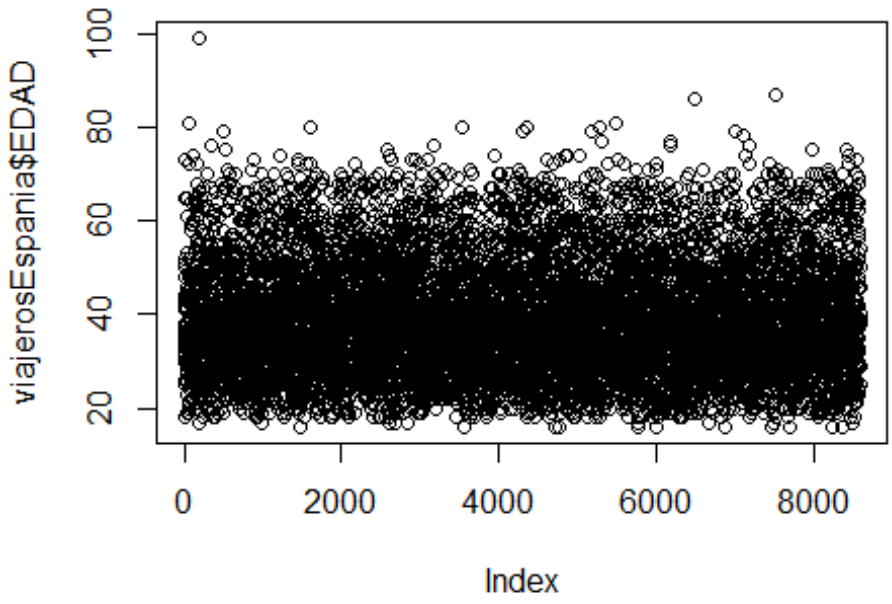
- Fragmentar el datset según los diferentes países que hay: España, Reino Unido, Alemania y otros puesto que vivir en países diferentes genera diferentes comportamientos. Por tanto, realizaré 4 análisis clúster.

Limpieza de datos

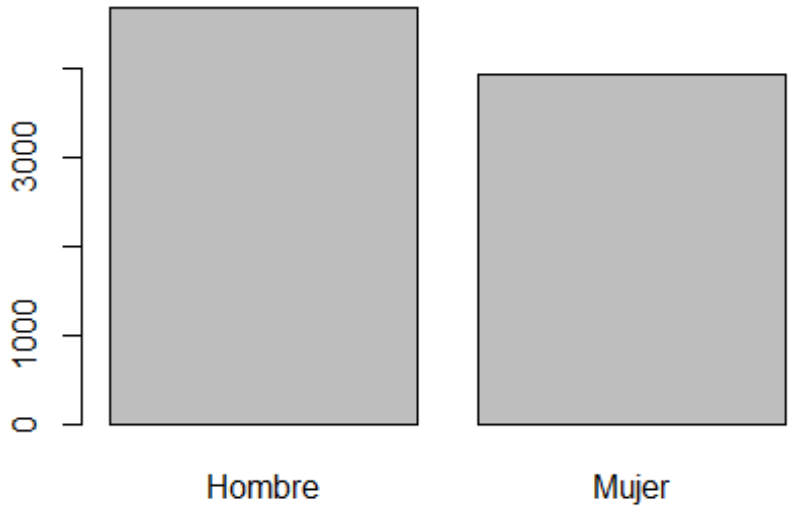
Antes de aplicar el método Clara, voy a preparar la muestra de datos que voy a emplear, ya que como he podido observar, existen muchos NA y variables nominales que hay que decidir cómo tratar:

Hago un dibujo de las variables nominales para ver cómo las puedo tratar:

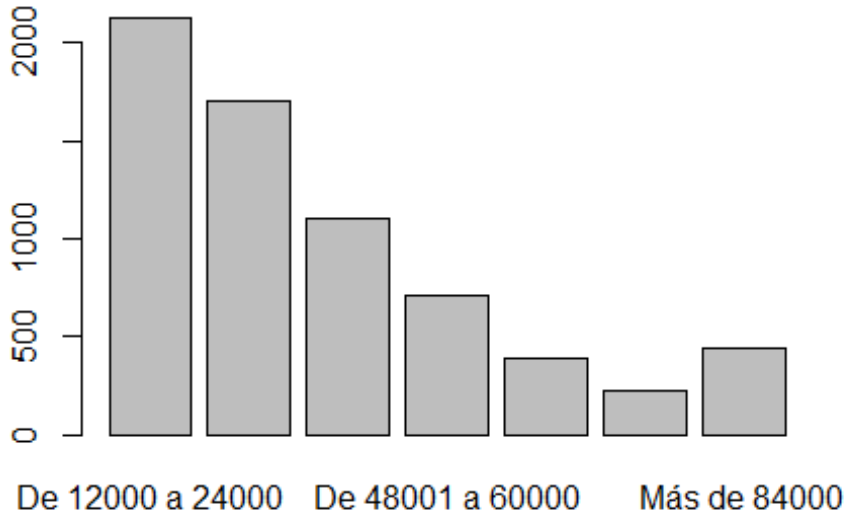
```r
viajerosEspania<-viajeros[viajeros$PAIS_RESID_AGRUP=="España",]

plot(viajerosEspania$EDAD)
```



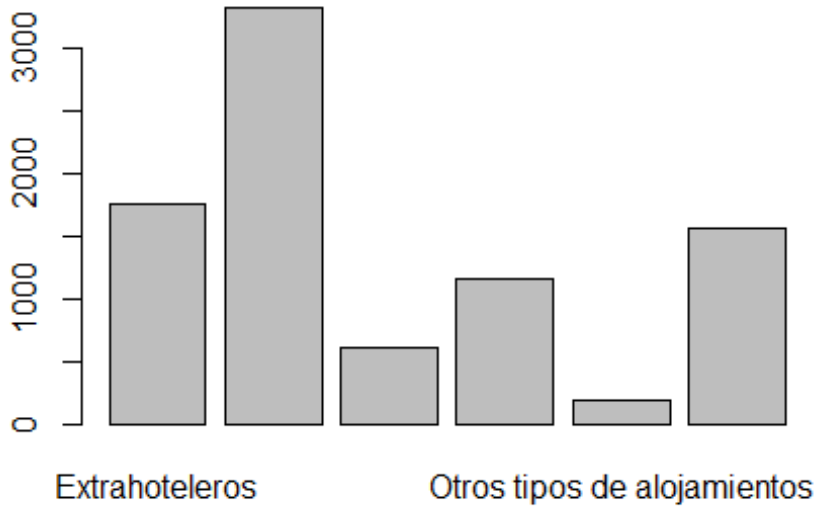

```
plot(viajerosEspania$SEXO)
```




```
plot(viajerosEspania$INGRESOS)
```




```r
plot(viajerosEspania$ALOJ_CATEG_1)
```



A la vista de los resultados de los gráficos, trataré las variables nominales como variables numéricas convirtiéndolas en dummy o agrupándolas:

```
#EDAD
#para que edad este en el intervalo 1-10
viajerosEspania$EDAD[viajerosEspania$EDAD<=35]=1
viajerosEspania$EDAD[36<=viajerosEspania$EDAD &
viajerosEspania$EDAD<=65]=2
viajerosEspania$EDAD[viajerosEspania$EDAD>=60]=3

head(viajerosEspania$EDAD)
```

```
## [1] 1 2 3 2 1 2
```

```
#SEXO

s=dummy(viajerosEspania$SEXO)
head(s)

Sexdummy=s[,2]



#ALOJAMIENTO


aloj=dummy(viajerosEspania$ALOJ_CATEG_1)


aloj1=aloj[,1] #extrahoteleros
aloj2=aloj[,2] #apartahoteles de 4 estrellas
aloj3=aloj[,3] #apartahoteles de 5 estrellas
aloj4=aloj[,4] #apartahoteles de hasta 3 estrellas
#aloj5=aloj[,6] #Viviendas propias o casas de amigos o familiares
Hoteles45=aloj3+aloj2
Hoteles123=aloj4
#viviendas propias o casas de amigos familiares

#INGRESOS



ingresos=dummy(viajerosEspania$INGRESOS)
ingresos=dummy(viajerosEspania$INGRESOS)
length(ingresos[,1])
```

```
## [1] 8594
```

```
i1=ingresos[,1]  #hasta 24000
i2=ingresos[,2]*2  #24000-48000
i3=ingresos[,3]*3  #48000-60000

i5=ingresos[,4]*3  #60000-72000
i7=ingresos[,5]*3  # mas de 84000



ingresos=(i1+i2+i3+i5+i7)

head(ingresos)

## [1] 1 3 0 3 1 2
```

Elimino las variables iniciales nominales:

```
viajeros_orig=viajerosEspania
varnominales=c(1,2,32, 34, 35)
viajerosEspania=viajerosEspania[,-varnominales]
```

Incluyo todas las nuevas variables que he creado nuevas en la tabla "viajerosEspania"

```
#viajerosEspania=data.frame(viajerosEspania,Hoteles45, Hoteles123)
viajerosEspania=data.frame(viajerosEspania,Sexdummy)
viajerosEspania=data.frame(viajerosEspania,ingresos)
viajerosEspania=data.frame(viajerosEspania, Hoteles123, Hoteles45)
```

Pongo todos los NA como 0 ya que el no contestar a una columna lo considero importante. Esto puede ser debido a que la persona en concreto no valora ciertos aspectos de los viajes que aparecen en el cuestionario.

```
viajerosEspania[is.na(viajerosEspania)]=0
```

Por último elimino las filas de "viajerosEspania" que tengan más de 20 NA, esto se puede interpretar como las personas que han dejado en blanco más de 20 preguntas en el cuestionario:

```r
sum(table(viajerosEspania$numNA[viajerosEspania$numNA<=20]))
```

```
## [1] 7699
```

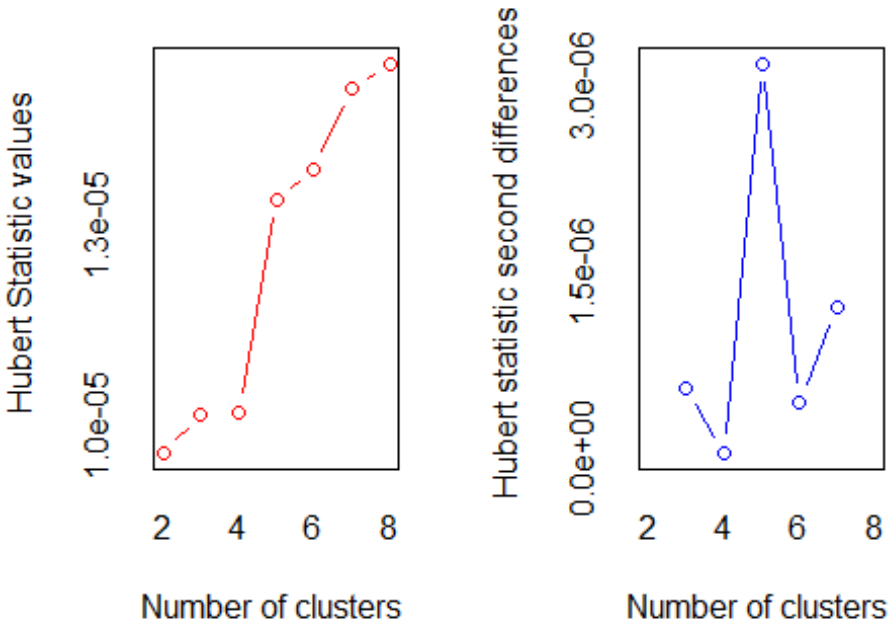
```r
viajeros_orig=viajerosEspania
viajerosEspania<-viajerosEspania[viajerosEspania$numNA<=20,]
```

Análisis clúster viajeros-España

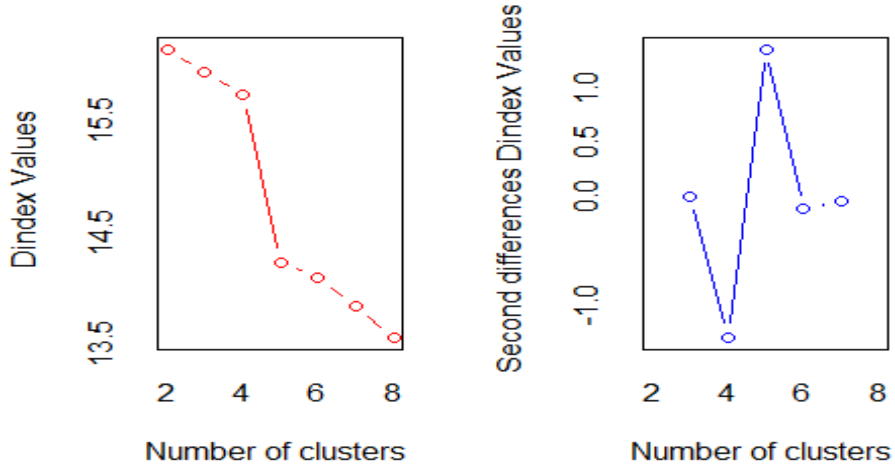
Primero, determino el número de clústers adecuados para luego introducirlo como parámetro en el método clara:

Como se puede observar en los siguientes gráficos y en el script, el número adecuado de clúster para la muestra es 5

```r
set.seed(567)
viajerosEspania.mas = viajerosEspania[sample(1:nrow(viajerosEspania), 1000, replace=FALSE),]

require(NbClust)



Nb.viajeros=NbClust(viajerosEspania.mas, distance = "euclidean", min.nc = 2,max.nc=8, method = "complete", index ="all")
```



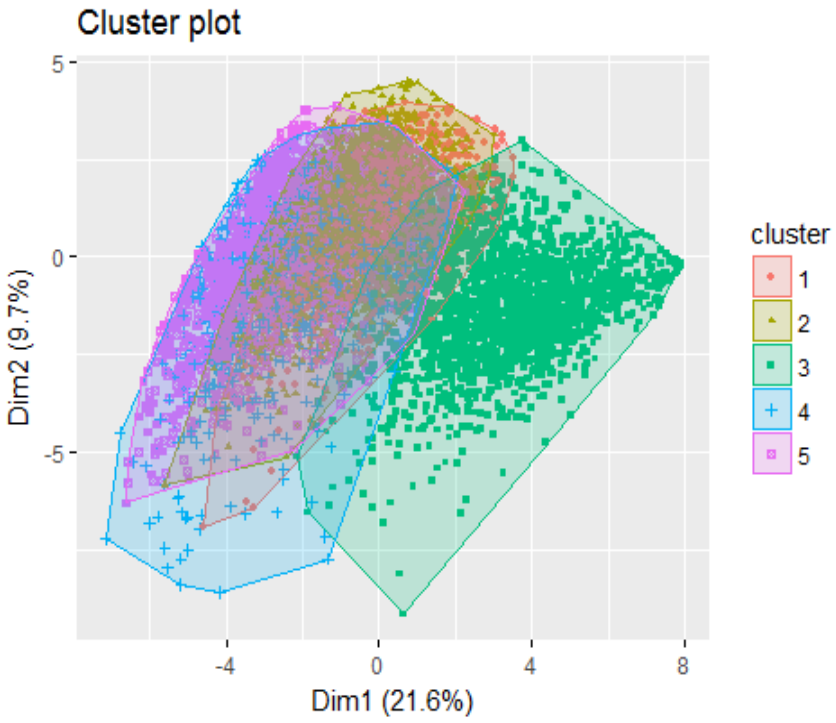

```
## *** : The Hubert index is a graphical method of determining the number
of clusters.
##                 In the plot of Hubert index, we seek a significant
knee that corresponds to a
##                 significant increase of the value of the measure i.e
the significant peak in Hubert
##                 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:
## * 7 proposed 2 as the best number of clusters
## * 1 proposed 3 as the best number of clusters
## * 1 proposed 4 as the best number of clusters
## * 9 proposed 5 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  5
##
##
## *******************************************************************
```

```r
#km.q

#km.q$cluster

#viajeros:todas las variables con na.omit

viajerosEspania.clara=clara(viajerosEspania, 5, samples=200)
require(factoextra)
fviz_cluster(viajerosEspania.clara, stand = TRUE, geom = "point",
pointsize = 1)
```

La proyección de cada uno de los clúster obtenidos se refleja en el gráfico siguiente. Se puede observar que el porcentaje explicado entre los dos primeros clústers es de 32% ésto es debido a que se han incluído todas las variables en el clúster.




```
# Medioides
tabla<-viajerosEspania.clara$medoids
#View(tabla)
tabla
```

```
##        IMPRESION VALORACION_ALOJ VALORACION_TRATO_ALOJ
## 135405         4               8                     8
## 211380         5               7                     7
## 84120          4               7                     7
## 76568          4               0                     0
## 258978         4               9                     9
##        VALORACION_GASTRONO_ALOJ VALORACION_CLIMA
VALORACION_ZONAS_BANYO
## 135405                        8                8
8
## 211380                        7                8
7
## 84120                         6                9
8
## 76568                         0                7
7
## 258978                        8                8
7
##        VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE
## 135405                   8                        8
## 211380                   8                        8
## 84120                    8                        8
## 76568                    7                        7
```

```
## 258978                   7                          8
##        VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
## 135405                       8                    8
## 211380                       8                    8
## 84120                        8                    7
## 76568                        8                    6
## 258978                       8                    8
##        VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
## 135405                          8                           8
## 211380                          8                           8
## 84120                           7                           7
## 76568                           7                           6
## 258978                          7                           6
##        VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR
## 135405                        8                         8
## 211380                        8                         7
## 84120                         7                         7
## 76568                         7                         7
## 258978                        7                         7
##        VALORACION_CULTURA VALORACION_DEPORTES VALORACION_GOLF
## 135405                  0                   0              0
## 211380                  0                   0              0
## 84120                   6                   7              7
## 76568                   0                   0              0
## 258978                  0                   0              0
##        VALORACION_PARQUES_OCIO VALORACION_AMBIENTE_NOCTURNO
## 135405                       0                            0
## 211380                       0                            0
## 84120                        8                            7
## 76568                        0                            0
## 258978                       0                            0
##        VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS
VALORACION_SALUD
## 135405                      0                       0
0
## 211380                      0                       0
0
## 84120                       7                       6
7
## 76568                       0                       0
0
## 258978                      0                       0
0
##        VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI
## 135405                        8                         8
## 211380                        0                         0
## 84120                         7                         7
## 76568                         7                         0
## 258978                        0                         0
##        VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
```

```
## 135405                     8                    8
## 211380                     8                    8
## 84120                      7                    7
## 76568                      0                    6
## 258978                     0                    0
##        VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
## 135405                            8                           8
## 211380                            8                           8
## 84120                             7                           8
## 76568                             4                           5
## 258978                            0                           0
##        VALORACION_HOSPITALIDAD EDAD numNA Sexdummy ingresos Hoteles123
## 135405                       8    1     8        0        1          0
## 211380                       8    2    10        0        3          0
## 84120                        8    2     0        1        2          0
## 76568                        6    2    14        0        0          0
## 258978                       0    2    15        0        0          0
##        Hoteles45
## 135405         1
## 211380         0
## 84120          1
## 76568          0
## 258978         1
```

Conclisiones clúster viajeros Españoles

Este clúster nos proporciona 5 grupos. A continuación se describen algunas características de cada uno de ellos:

- **Grupo1:** Se caracteriza por estar formado por individuos menores de 35 años e ingresos inferiores a 24000 euros. En general hacen una buena valoración del viaje (puntuación 8) , sin embargo, no valoran aspectos como cultura, deportes, golf, parques de ocio, recreo niños, es decir, temas más concretos.

- **Grupo2:** La media de edad se encuentra entre 35 y 65 años. Tienen unos ingresos elevados. En general, su valoración es más exigente que la del grupo 1(puntuación 7) y los conceptos que se han valorado en éste grupo, a diferencia del anterior, son el alquiler de vehículos y el estado de las carreteras

y seguridad. Ésto puede ser debido a que éstas personas han ido a sus viajes en coche o han alquilado uno.

- Grupo 3: La media de edad se encuentra entre 35 y 65 años. Los ingresos de éste grupo son medios, entre 24000-48000 euros. Son personas que han estado en hoteles de más de 3 estrellas. Es el único grupo que puntúa variables como cultura, deporte, golf, ambiente nocturno y excursiones, con un nivel aproximado de un 7, es decir, piensa que es posible mejorarlo. Las personas de éste grupo se fijan más en los detalles ya que van a hoteles de alta categoría, con todo tipo de servicios.

- Grupo 4: La media de Edad es también entre 35 y 65 años. El tipo de alojamiento es extrahotelero. Las personas de éste grupo, no tienen mucho interés en completar el cuestionario, pues el número de NA es cercano a 14. Por éste motivo, no se puede valorar la cantidad de ingresos ya que aparece sin valorar (es decir, en ésta clase no se han rellenado). Son personas que valoran pocos concepto, pero específicos. Destaca que como temas particulares valoran el estado de las carreteras, la seguridad y calidad de comercio con una valoración baja, cercana al 5. Estas personas manifiestan una queja en cuanto a las valoraciones específicas dichas.

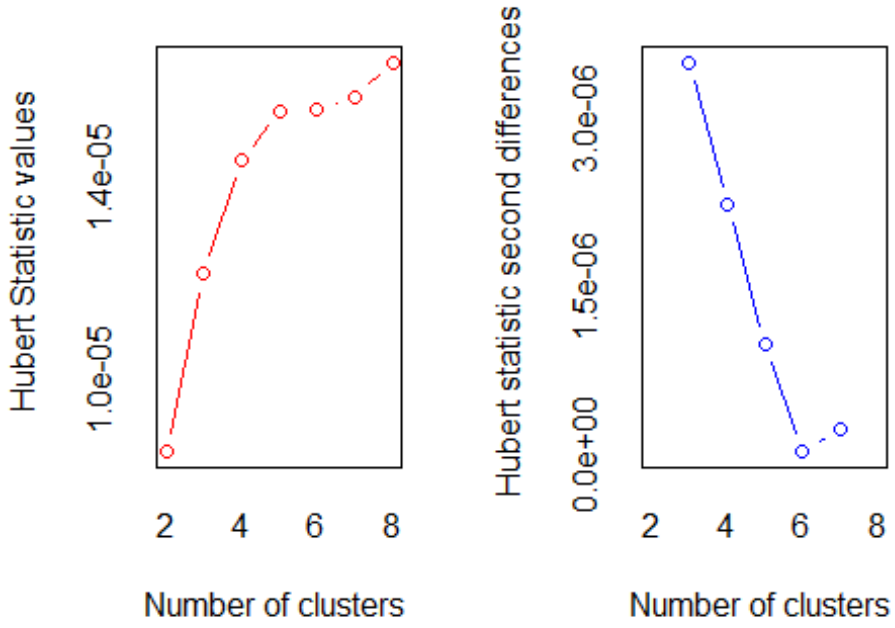
- Grupo 5: La media de Edad es también entre 35 y 65 años. Se alojan en hoteles de 4 a 5 estrellas. Las personas de éste grupo, no tienen mucho interés en completar el cuestionario, pues el número de NA es cercano a 15. Por éste motivo, no se conoce el nivel de ingresos, sin embargo, el tipo de alojamiento es de hoteles de 4 y 5 estrellas. En la valoración genérica no son exigentes.

Análisis clúster viajeros-Alemania

Se han realizado los mismos pasos que para el caso de España. Los resultados obtenidos son los siguientes:

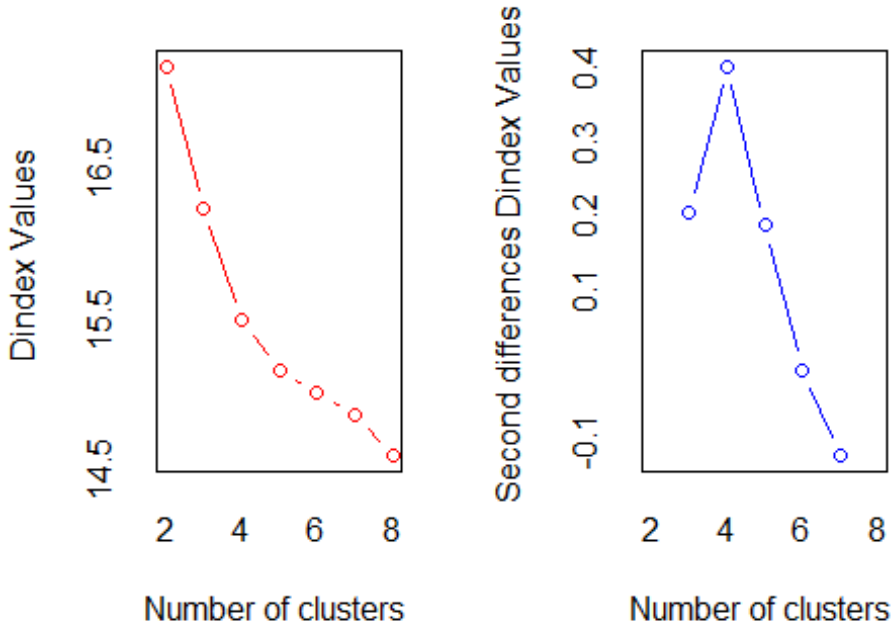
-Se escoge 2 como el número óptimo de clústers

```
set.seed(567)
viajerosAlemania.mas = viajerosAlemania[sample(1:nrow(viajerosAlemania),
1000, replace=FALSE),]

require(NbClust)



Nb.viajeros=NbClust(viajerosAlemania.mas, distance = "euclidean", min.nc
= 2,max.nc=8, method = "complete", index ="all")
```



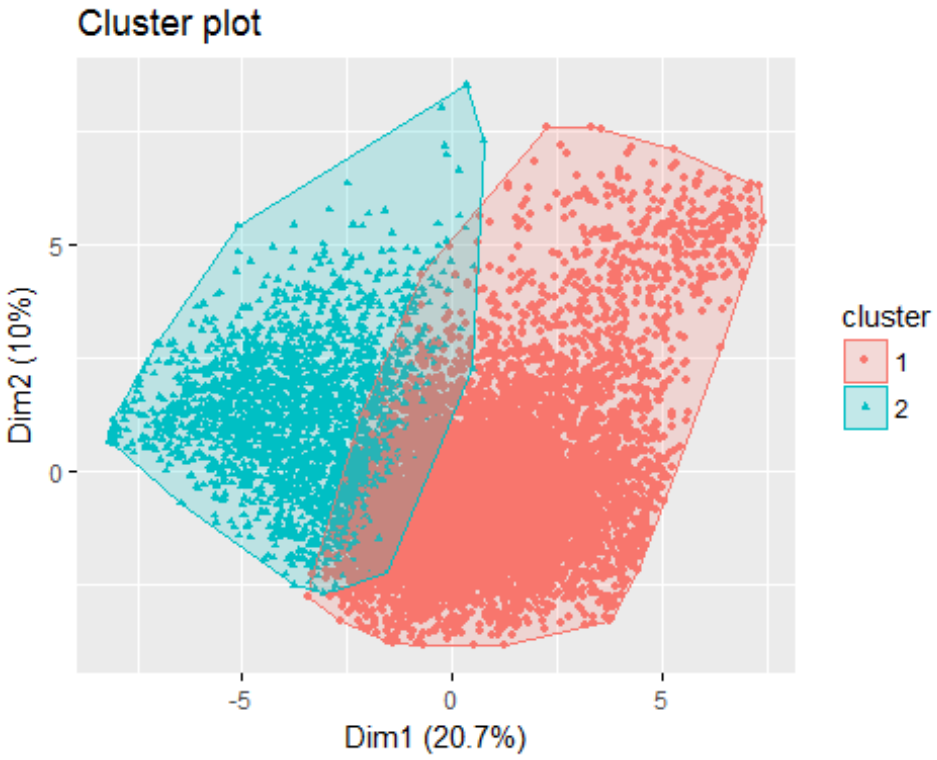

```
## *** : The Hubert index is a graphical method of determining the number
of clusters.
##                 In the plot of Hubert index, we seek a significant
knee that corresponds to a
##                 significant increase of the value of the measure i.e
the significant peak in Hubert
```

```
##                  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:
## * 6 proposed 2 as the best number of clusters
## * 2 proposed 3 as the best number of clusters
## * 6 proposed 4 as the best number of clusters
## * 2 proposed 5 as the best number of clusters
## * 4 proposed 6 as the best number of clusters
## * 2 proposed 7 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

```r
#km.q

#km.q$cluster

#viajeros:todas las variables con na.omit

viajerosAlemania.clara=clara(viajerosAlemania, 2, samples=200)
require(factoextra)
fviz_cluster(viajerosAlemania.clara, stand = TRUE, geom = "point",
pointsize = 1)
```




```r
# Medioides
tabla<-viajerosAlemania.clara$medoids
View(tabla)
```

```
IMPRESION VALORACION_ALOJ VALORACION_TRATO_ALOJ VALORACION_GASTRONO_ALOJ
96322          4                8                     8
6
134286         4                9                     9
9
       VALORACION_CLIMA VALORACION_ZONAS_BANYO VALORACION_PAISAJES
96322                 8                      6                   6
134286                8                      8                   8
       VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD
VALORACION_LIMPIEZA
96322                          7                       8
8
```

```
134286                          8                         8
8
       VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
96322                           6                           5
134286                          8                           8
       VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR
VALORACION_CULTURA
96322                         6                         5
0
134286                        8                         8
7
       VALORACION_DEPORTES VALORACION_GOLF VALORACION_PARQUES_OCIO
96322                    0              0                       0
134286                   7              7                       7
       VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES
96322                             0                      0
134286                            7                      7
       VALORACION_RECREO_NINYOS VALORACION_SALUD VALORACION_SERVICIOS_BUS
96322                        0                0                         5
134286                       7                7                         8
       VALORACION_SERVICIOS_TAXI VALORACION_ALQ_VEHIC
VALORACION_SEGURIDAD
96322                           5                    0
5
134286                          8                    8
8
       VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
96322                             7                           5
134286                            8                           8
       VALORACION_HOSPITALIDAD EDAD numNA Sexdummy ingresos Hoteles123
Hoteles45
96322                        7    3     9        1        3          1
0
134286                       8    2     0        1        2          0
1
```

Conclusiones análisis clúster Alemania

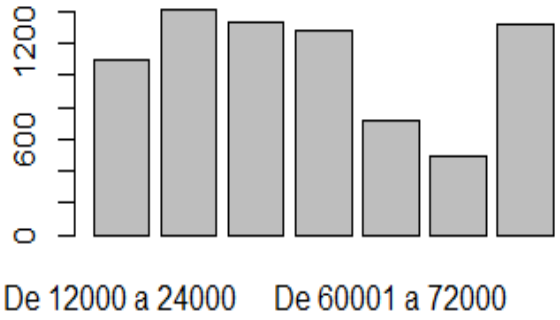
Se han creado dos clúster muy bien diferenciados con las siguientes características:

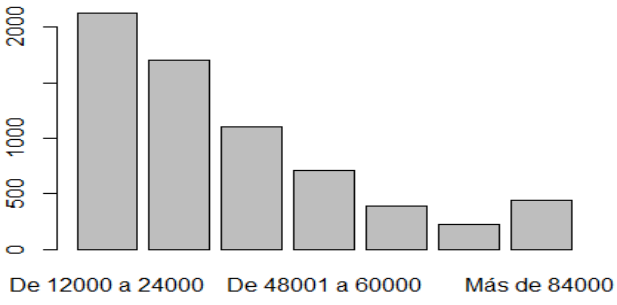
- Grupo 1: La media de edad es elevada, mayor de 65 años y con unos ingresos también elevados, mayor de 60000euros. Sin embargo el tipo de alojamiento es más modesto de lo que cabría esperar para éstos ingresos, quizás debido a la edad (y el tipo de viajes organizados para mayores). La valoración general del alojamiento es elevada, sin embargo algunos conceptos presentan valoraciones

muy bajas, por ejemplo servicio bus o taxi (asociado a la edad). Además, por la edad, no les importan aspectos como la cultura, los deportes o el ocio, dejándolos sin valorar.

- Grupo 2: La media de edad es entre 35 y 60 años. Se alojan en mejores hoteles, a pesar de que sus ingresos son elevados pero inferiores al grupo 1. Son menos exigentes en su valoración y valoran aspectos más concretos como el deporte, la cultuta, el ambiente nocturno... .

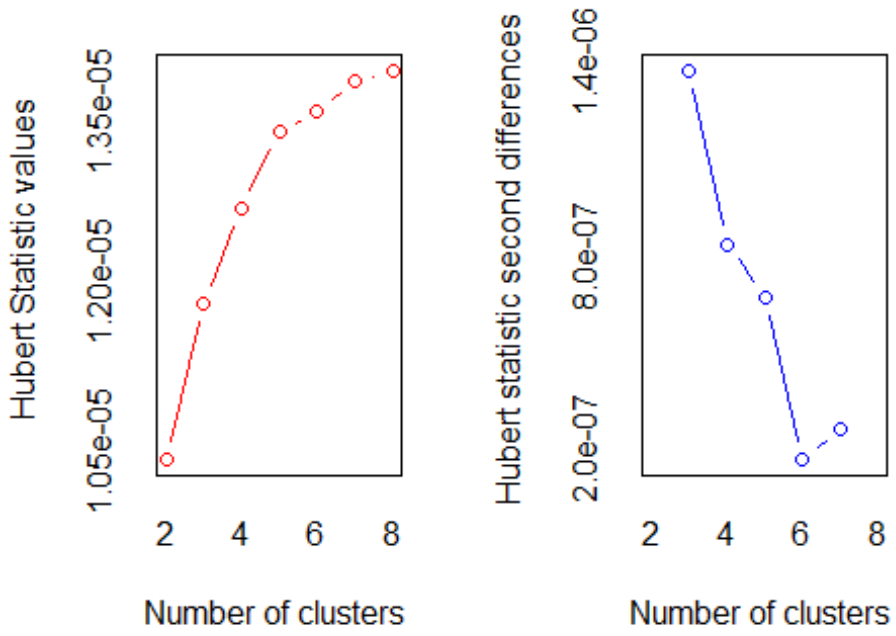
❖ En general, como comparativa con el clúster de España, se observa que los sueldos en Alemania son más elevados que en España, y prueba de ello :es el histograma (el primero de Alemania y el segundo de España):



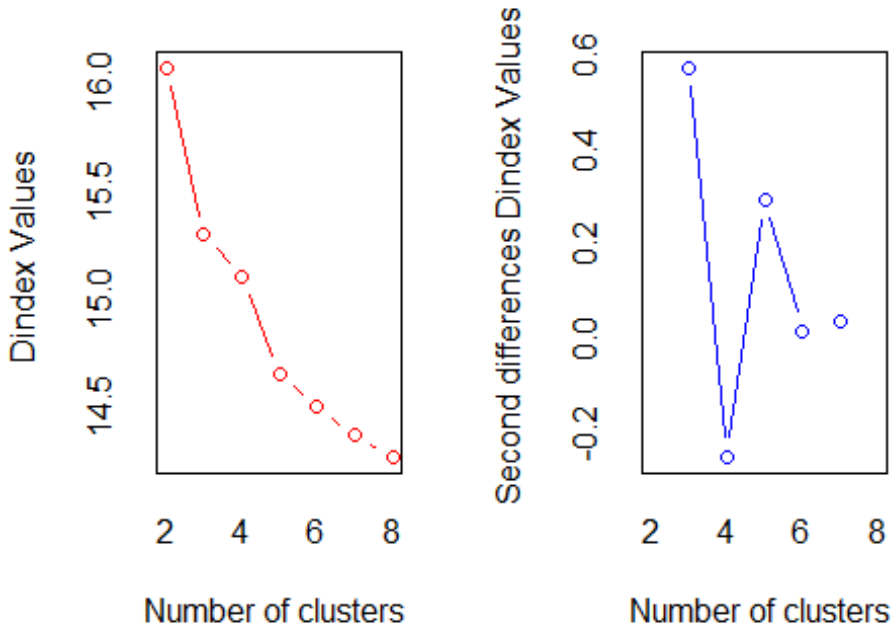


Análisis Clúster Reino Unido

```
set.seed(567)
viajerosRU.mas = viajerosRU[sample(1:nrow(viajerosRU), 1000,
replace=FALSE),]

require(NbClust)



Nb.viajeros=NbClust(viajerosRU.mas, distance = "euclidean", min.nc =
2,max.nc=8, method = "complete", index ="all")
```

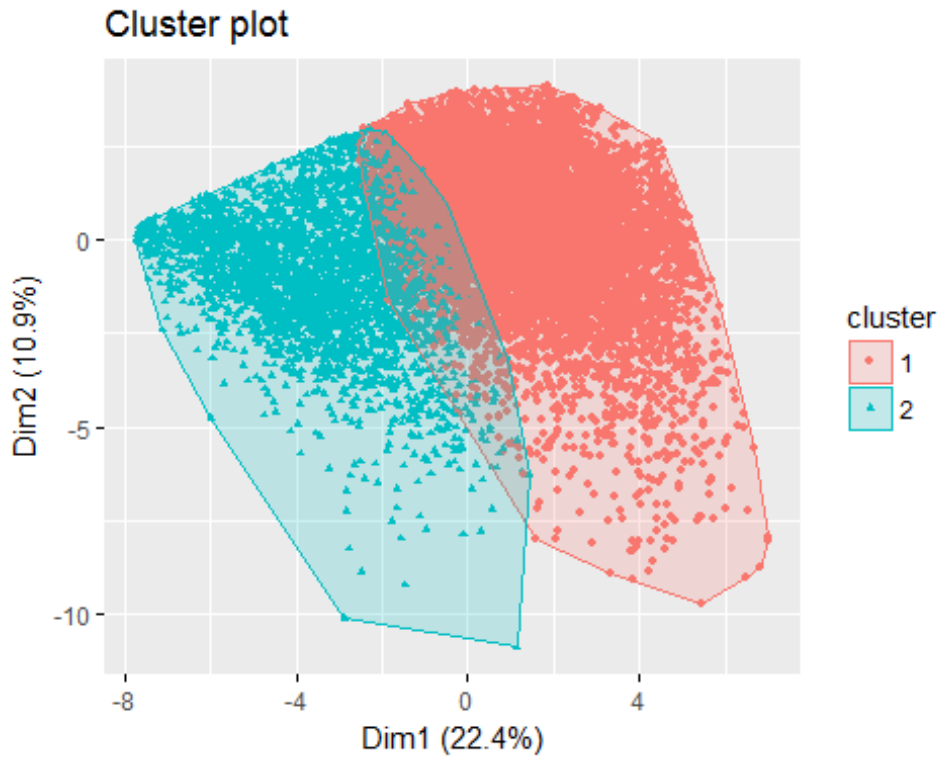



```
## *** : The Hubert index is a graphical method of determining the number
of clusters.
##                 In the plot of Hubert index, we seek a significant
knee that corresponds to a
##                 significant increase of the value of the measure i.e
the significant peak in Hubert
##                 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
## * 7 proposed 3 as the best number of clusters
## * 7 proposed 5 as the best number of clusters
## * 1 proposed 8 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************

#km.q

#km.q$cluster
```

```r
#viajeros:todas las variables con na.omit

viajerosRU.clara=clara(viajerosRU, 2, samples=200)
require(factoextra)
fviz_cluster(viajerosRU.clara, stand = TRUE, geom = "point", pointsize =
1)
```




```r
# Medioides
tabla<-viajerosRU.clara$medoids
#View(tabla)

tabla
```

```
##        IMPRESION VALORACION_ALOJ VALORACION_TRATO_ALOJ
## 172339         4               8                     8
## 179983         4               8                    10
##        VALORACION_GASTRONO_ALOJ VALORACION_CLIMA
VALORACION_ZONAS_BANYO
## 172339                        7                8
8
## 179983                        8                9
8
##        VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE
## 172339                   7                        7
## 179983                   7                        8
##        VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
## 172339                       7                   8
```

```
## 179983                        8                    9
##        VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
## 172339                          8                           8
## 179983                          8                           8
##        VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR
## 172339                        8                         8
## 179983                       10                         8
##        VALORACION_CULTURA VALORACION_DEPORTES VALORACION_GOLF
## 172339                  0                   0               0
## 179983                  7                   7               7
##        VALORACION_PARQUES_OCIO VALORACION_AMBIENTE_NOCTURNO
## 172339                       0                           0
## 179983                       7                           7
##        VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS
VALORACION_SALUD
## 172339                      0                       0
0
## 179983                      7                       7
7
##        VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI
## 172339                        0                         7
## 179983                        7                         7
##        VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
## 172339                    0                    0
## 179983                    7                    8
##        VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
## 172339                            6                           5
## 179983                            7                           7
##        VALORACION_HOSPITALIDAD EDAD numNA Sexdummy ingresos Hoteles123
## 172339                       7    2    11        0        2          0
## 179983                       9    2     0        0        3          0
##        Hoteles45
## 172339         1
## 179983         0
```

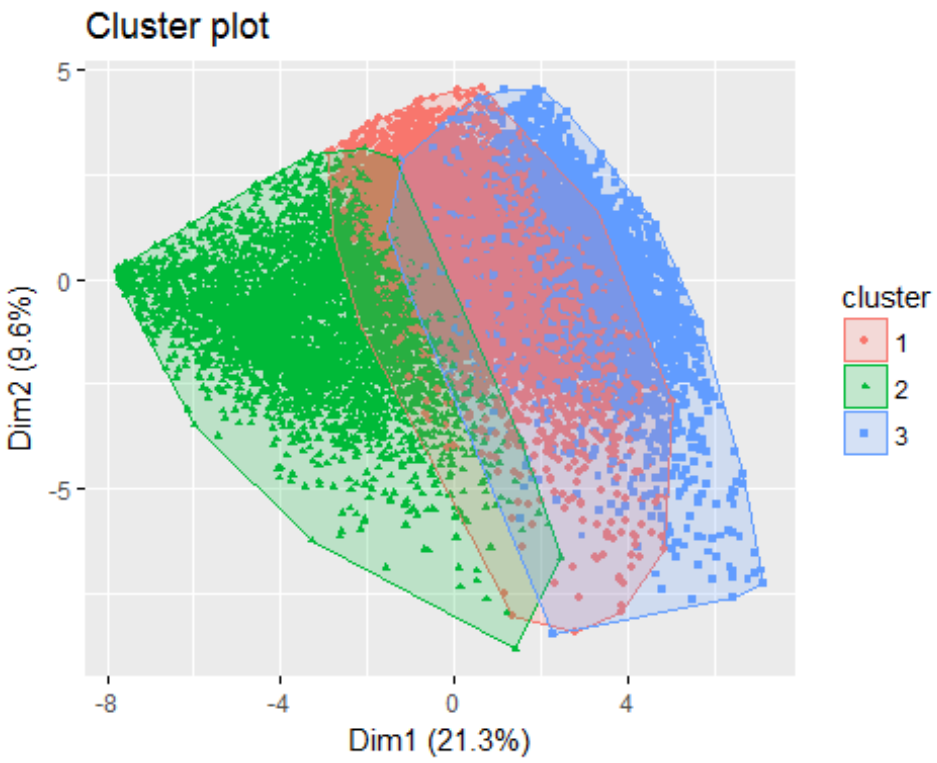
Conclusiones del análisis clúster para Reino Unido

- Grupo 1: el nivel de ingresos del grupo 1 es moderado alto. Este grupo está compuesto por las personas que se alojan en hoteles de 4,5 estrellas de categoría elevada. Sólo valoran conceptos generales con una puntuación alta (entre 7 y 9), no valorando aspectos más concretos como deportes, cultura,... .

- Grupo 2: tienen una media de edad parecida al grupo anterior, pero más ingresos, por lo que valoran otros servicios más concretos como salud, niños, nocturno, cultura... No dejan ninguna pregunta sin responder. Su valoración es también alta, aunque de los conceptos particulares es un poco más baja.

Conclusiones del análisis clúster para Otros países

```
#viajeros:todas las variables con na.omit

viajesOTROS.clara=clara(viajesOTROS, 3, samples=200)
require(factoextra)
fviz_cluster(viajesOTROS.clara, stand = TRUE, geom = "point", pointsize =
1)
```




```
# Medioides
tabla<-viajesOTROS.clara$medoids
#View(tabla)

tabla
```

```
##       IMPRESION VALORACION_ALOJ VALORACION_TRATO_ALOJ
## 11457         5              8                     8
## 47545         4              8                     8
## 8355          4              7                     8
##       VALORACION_GASTRONO_ALOJ VALORACION_CLIMA VALORACION_ZONAS_BANYO
## 11457                        7                8                      8
## 47545                        8                8                      8
## 8355                         7                8                      8
##       VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE
## 11457                   8                        8
## 47545                   8                        8
## 8355                    7                        7
##       VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
## 11457                       8                   8
## 47545                       8                   8
## 8355                        7                   7
##       VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
## 11457                          7                           6
## 47545                          7                           8
## 8355                           7                           6
##       VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR
## 11457                        7                        7
## 47545                        8                        8
## 8355                         7                        7
##       VALORACION_CULTURA VALORACION_DEPORTES VALORACION_GOLF
## 11457                  0                   0              0
## 47545                  7                   7              7
## 8355                   0                   0              0
##       VALORACION_PARQUES_OCIO VALORACION_AMBIENTE_NOCTURNO
## 11457                       0                           0
## 47545                       7                           7
## 8355                        0                           0
##       VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS VALORACION_SALUD
## 11457                      0                       0                0
## 47545                      7                       7                7
## 8355                       0                       0                0
##       VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI
## 11457                        0                        7
## 47545                        7                        7
## 8355                         0                        0
##       VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
## 11457                    0                    7
## 47545                    7                    7
## 8355                     0                    0
##       VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
## 11457                            7                          7
## 47545                            7                          7
## 8355                             0                          0
##       VALORACION_HOSPITALIDAD EDAD numNA Sexdummy ingresos Hoteles123
## 11457                       7    2    10        0        1          1
```

```
## 47545                      7    1     0        0        1          0
## 8355                       0    2    15        1        3          0
##       Hoteles45
## 11457         0
## 47545         0
## 8355          0
```

- Grupo 1: Éste clúster se caracteriza por tener una media de edad entre 35 y 65 años. Los ingresos son bajos entre 12000-24000 euros. Se alojan en hoteles de hasta 3 estrellas, acorde con su nivel de ingresos. Éste clúster corresponde a gente que no valoran los temas concretos (cultura, deportes, golf..). Son personas que se mueven en bus o en coche.

- Grupo 2: Éste clúster se caracteriza por tener una media de edadmás joven, menores de 35 años. Los ingresos son bajos entre 12000-24000 euros. Su alojamiento es extrahotelero y presentan interés por todos los conceptos puesto que responden a todas las preguntas.

- Grupo 3: Es un grupo de edad de 35 a 65 años, tienen ingresos altos (más que en los otros gruposÇ) y su modo de alojamiento es extrahotelero. No tienen demasiado interés en completar el cuestionario, dejan si contestar 15 preguntas, principalmente las correspondientes a conceptos más detallados.

Comentarios finales

- El inicio de éste estudio ha sido bastante arduo:

  Primeramente, ha habido que tratar los datos para conseguir una tabla coherente.

  En segundo lugar, he intentado hacer el análisis clúster con la tabla completa, sin tomar las muestras de cada país y diferenciándolos en la tabla de todas las variables. Sin embargo, los países no se diferenciaban el el análisis clúster aún

eliminando variables que consideraba redundantes y volviendo a realizar el análisis. Los resultados de éste análisis me llevaron a sacar las siguientes conclusiones:

- En las elecciones de variables no era posible observar demasiada diferencia entre clústers en algunas valoraciones como el alojamiento, impresión o las de los restaurantes, teniendo todas una media de entre 7 y 8. Por lo que no es posible interpretar las valoraciones en función del sexo, de los países, edad o número de ingresos. Ésto podría ser debido a que la mayoría de las variables que había escogido para el análisis clúster tenían puntuaciones muy homogéneas (la mayoría entre 8 y 10), por lo que no marcan diferencia entre clústers, teniendo todos una media parecida de valoraciones.
- Como solución a ésto, realicé un nuevo análisis clúster, en el que sólo estuvieran variables en las que se marcara diferencia. Éstas variables son las que tienen muchos NA y que sólo han contestado personas que de verdad valoran los aspectos representados por ellas. Sin embargo, éstas no generaban diferenciación en las características que nos interesaban de los individuos, como son edad, alojamiento, ingresos...
- Por último se ha considerado el análisis explicado en éste documento diferenciando un dataset por cada país.