Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

Técnicas de Agrupación y de Reducción de la Dimensión. Master en Data Science para Finanzas

# Práctica 1: ¿Qué sabemos de los viajeros?

*El archivo viajeros.csv muestra los datos de 50.000 viajeros de los que tenemos información diversa como su nacionalidad, su sexo, su edad, y, lo que más nos interesa, la valoración que ofrecen a un conjunto muy amplio de características, como su impresión general (en la columna del mismo nombre) o el alojamiento, entre otras (recogidas bajo el genérico valoración_xxx, donde xxx recoge el atributo concreto valorado).*

*Debe emitir un informe detallando si existen grupos de turistas homogéneos a partir de la valoración que dan a los distintos servicios, tal y como se ha señalado. Deberá justificar que el análisis puede llevarse a cabo, el número de clusters a emplear, y la solución alcanzada. Se valorará, además de una correcta presentación, la identificación de posibles relaciones de los clusters alcanzados con una o varias características adicionales (pista: piense en la posible concentración de viajeros en los clusters en función de su nacionalidad, de su profesión, de su sexo, edad o renta, o de una combinación de las mismas).*

## Carga de datos y análisis exploratorio

```
setwd("C:/Users/Manuel/Desktop/CUNEF/Técnicas Agrupacion y Reduccion/PRACTICA VIAJEROS")

viajeros<-read.csv("C:/Users/Manuel/Desktop/CUNEF/Técnicas Agrupacion y Reduccion/PRACTICA VIAJEROS/viajeros.csv", header=T, sep=",")
#View(viajeros)

#ponemos la primera columna del data frame como nombre de las filas:
viajeros<-data.frame(viajeros[,-1], row.names=viajeros[,1])

#creamos una copia de seguridad del data frame original
viajeros_orig=viajeros
# Escalamos la variable impresion
viajeros$IMPRESION <- viajeros$IMPRESION*2
```

Vemos las observaciones que tienen menos NA

```
viajeros_cuantit=viajeros_orig[,c(3:31)]

vacios<-c()
nom<-c()
for (i in (1:length(viajeros_cuantit))) {
        a<- sum(is.na(viajeros_cuantit[,i]))
        vacios<-c(vacios,a)
```

```
        b<-colnames(viajeros_cuantit[i])
        nom<- c(nom,b)
}
df_vacios<-data.frame(vacios,nom)
df_vacios[df_vacios$vacios<30000&df_vacios$vacios>0,]
```

```
##    vacios                          nom
## 1    7311                    IMPRESION
## 2    5327               VALORACION_ALOJ
## 3    8339         VALORACION_TRATO_ALOJ
## 4   14446      VALORACION_GASTRONO_ALOJ
## 5    3894               VALORACION_CLIMA
## 6    7442        VALORACION_ZONAS_BANYO
## 7    5901           VALORACION_PAISAJES
## 8    6627     VALORACION_MEDIO_AMBIENTE
## 9    5504       VALORACION_TRANQUILIDAD
## 10   5088           VALORACION_LIMPIEZA
## 11   9750    VALORACION_CALIDAD_RESTAUR
## 12  13208   VALORACION_OFERTA_GASTR_LOC
## 13  10620      VALORACION_TRATO_RESTAUR
## 14  12199     VALORACION_PRECIO_RESTAUR
## 15  29553            VALORACION_CULTURA
## 23  28041      VALORACION_SERVICIOS_BUS
## 24  25290     VALORACION_SERVICIOS_TAXI
## 25  28390          VALORACION_ALQ_VEHIC
## 26  23160          VALORACION_SEGURIDAD
## 27  14589 VALORACION_ESTADO_CARRETERAS
## 28  16912   VALORACION_CALIDAD_COMERCIO
## 29  14194       VALORACION_HOSPITALIDAD
```

Vemos que algunas variables tienen una gran cantidad de NA. En este sentido, a fin de no perder demasiada información, transformaremos las variables Valoracion Golf y Valoracion Recreo NiNyos en variables dicotómicas.

Asumimos que los viajeros que valoran el recreo de los niños lo hacen porque tienen hijos o similares (sobrinos) y que, por tanto, están en disposición de poder hacer una valoración, del mismo modo que asumimos que las personas que valoran GOLF son jugadores de golf

```
summary(viajeros)
```

```
##     PAIS_RESID_AGRUP
##  Alemania   :11164
##  España     : 8594
##  Otros      :16967
##  Reino Unido:13275
##
##
##
##                                                ALOJ_CATEG_1     IMPRESION
```

```
##  Extrahoteleros                                      :15416   Min.   : 2.0
00
##  Hoteles - apartahoteles de 4 estrellas              :18482   1st Qu.: 8.0
00
##  Hoteles - apartahoteles de 5 estrellas              : 3117   Median : 8.0
00
##  Hoteles - apartahoteles de hasta 3 estrellas        : 7402   Mean   : 8.6
19
##  Otros tipos de alojamientos                         : 1748   3rd Qu.:10.0
00
##  Viviendas propias o casas de amigos o familiares: 3835   Max.   :10.0
00
##                                                               NA's   :7311
##  VALORACION_ALOJ  VALORACION_TRATO_ALOJ VALORACION_GASTRONO_ALOJ
##  Min.   : 1.000   Min.   : 1.000        Min.   : 1.000
##  1st Qu.: 7.000   1st Qu.: 8.000        1st Qu.: 7.000
##  Median : 8.000   Median : 9.000        Median : 8.000
##  Mean   : 7.978   Mean   : 8.292        Mean   : 7.649
##  3rd Qu.: 9.000   3rd Qu.:10.000        3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.000        Max.   :10.000
##  NA's   :5327     NA's   :8339          NA's   :14446
##  VALORACION_CLIMA VALORACION_ZONAS_BANYO VALORACION_PAISAJES
##  Min.   : 1.000   Min.   : 1.000         Min.   : 1.000
##  1st Qu.: 8.000   1st Qu.: 7.000         1st Qu.: 7.000
##  Median : 9.000   Median : 8.000         Median : 9.000
##  Mean   : 8.517   Mean   : 8.068         Mean   : 8.248
##  3rd Qu.:10.000   3rd Qu.: 9.000         3rd Qu.:10.000
##  Max.   :10.000   Max.   :10.000         Max.   :10.000
##  NA's   :3894     NA's   :7442           NA's   :5901
##  VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
##  Min.   : 1.000            Min.   : 1.000          Min.   : 1.00
##  1st Qu.: 7.000            1st Qu.: 7.000          1st Qu.: 7.00
##  Median : 8.000            Median : 8.000          Median : 8.00
##  Mean   : 8.073            Mean   : 8.131          Mean   : 8.04
##  3rd Qu.: 9.000            3rd Qu.: 9.000          3rd Qu.: 9.00
##  Max.   :10.000            Max.   :10.000          Max.   :10.00
##  NA's   :6627              NA's   :5504            NA's   :5088
##  VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
##  Min.   : 1.000             Min.   : 1.000
##  1st Qu.: 7.000             1st Qu.: 7.000
##  Median : 8.000             Median : 8.000
##  Mean   : 7.738             Mean   : 7.465
##  3rd Qu.: 9.000             3rd Qu.: 9.000
##  Max.   :10.000             Max.   :10.000
##  NA's   :9750               NA's   :13208
##  VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR VALORACION_CULTURA
##  Min.   : 1.000           Min.   : 1.00             Min.   : 1.000
##  1st Qu.: 7.000           1st Qu.: 7.00             1st Qu.: 6.000
##  Median : 8.000           Median : 8.00             Median : 8.000
##  Mean   : 8.129           Mean   : 7.64             Mean   : 7.281
```

```
##  3rd Qu.: 9.000           3rd Qu.: 9.00            3rd Qu.: 9.000
##  Max.   :10.000           Max.   :10.00            Max.   :10.000
##  NA's   :10620            NA's   :12199            NA's   :29553
##  VALORACION_DEPORTES VALORACION_GOLF VALORACION_PARQUES_OCIO
##  Min.   : 1.000      Min.   : 1.0    Min.   : 1.00
##  1st Qu.: 7.000      1st Qu.: 5.0    1st Qu.: 7.00
##  Median : 8.000      Median : 7.0    Median : 8.00
##  Mean   : 7.682      Mean   : 6.9    Mean   : 7.66
##  3rd Qu.: 9.000      3rd Qu.: 9.0    3rd Qu.: 9.00
##  Max.   :10.000      Max.   :10.0    Max.   :10.00
##  NA's   :31423       NA's   :40340   NA's   :35044
##  VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES
##  Min.   : 1.00                Min.   : 1.00
##  1st Qu.: 6.00                1st Qu.: 7.00
##  Median : 8.00                Median : 8.00
##  Mean   : 7.33                Mean   : 7.74
##  3rd Qu.: 9.00                3rd Qu.: 9.00
##  Max.   :10.00                Max.   :10.00
##  NA's   :32470                NA's   :33797
##  VALORACION_RECREO_NINYOS VALORACION_SALUD VALORACION_SERVICIOS_BUS
##  Min.   : 1.00            Min.   : 1.0     Min.   : 1.000
##  1st Qu.: 6.00            1st Qu.: 6.0     1st Qu.: 7.000
##  Median : 8.00            Median : 8.0     Median : 8.000
##  Mean   : 7.34            Mean   : 7.5     Mean   : 7.708
##  3rd Qu.: 9.00            3rd Qu.: 9.0     3rd Qu.: 9.000
##  Max.   :10.00            Max.   :10.0     Max.   :10.000
##  NA's   :38298            NA's   :37452    NA's   :28041
##  VALORACION_SERVICIOS_TAXI VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
##  Min.   : 1.000            Min.   : 1.000       Min.   : 1.000
##  1st Qu.: 8.000            1st Qu.: 7.000       1st Qu.: 8.000
##  Median : 9.000            Median : 8.000       Median : 8.000
##  Mean   : 8.285            Mean   : 7.994       Mean   : 8.262
##  3rd Qu.:10.000            3rd Qu.: 9.000       3rd Qu.: 9.000
##  Max.   :10.000            Max.   :10.000       Max.   :10.000
##  NA's   :25290             NA's   :28390        NA's   :23160
##  VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
##  Min.   : 1.000               Min.   : 1.000
##  1st Qu.: 7.000               1st Qu.: 6.000
##  Median : 8.000               Median : 8.000
##  Mean   : 7.735               Mean   : 7.385
##  3rd Qu.: 9.000               3rd Qu.: 9.000
##  Max.   :10.000               Max.   :10.000
##  NA's   :14589                NA's   :16912
##  VALORACION_HOSPITALIDAD     SEXO            EDAD
##  Min.   : 1.000          Hombre:25975   Min.   :16.00
##  1st Qu.: 8.000          Mujer :24025   1st Qu.:32.00
##  Median : 9.000                         Median :44.00
##  Mean   : 8.477                         Mean   :44.29
##  3rd Qu.:10.000                         3rd Qu.:55.00
##  Max.   :10.000                         Max.   :99.00
```

```
##  NA's   :14194
##                          OCUPACION                  INGRESOS
##  Asalariado cargo medio      :12364   De 24001 a 36000: 7435
##  Jubilado <U+0096> retirado         : 5780   De 12000 a 24000: 7002
##  Empresario                  : 5234   De 36001 a 48000: 6302
##  Autónomo - profesión liberal: 4995   De 48001 a 60000: 5444
##  Otros trabajadores y obreros: 4576   Más de 84000    : 5435
##  (Other)                     :10878   (Other)         : 5393
##  NA's                        : 6173   NA's            :12989
```

### valoracion golf

Conseguimos una variable dicotómica que nos sitúa la observación dentro de un grupo concreto – golfistas. Tal y como se afirma en el artículo publicado en la BBC en 2014[1], el golf es un deporte que suele practicar gente con mayor poder adquisitivo que la media. Transformar esta variable en dicotómica nos salvará observaciones (tiene muchos NA).

```
viajeros$VALORACION_GOLF <- replace(viajeros$VALORACION_GOLF, !is.na(viaj
eros$VALORACION_GOLF), "Si")
viajeros$VALORACION_GOLF <- replace(viajeros$VALORACION_GOLF, is.na(viaje
ros$VALORACION_GOLF), "No")
```

### valoracion recreo ninyos

Conseguimos una variable dicotómica que nos dice de forma implícita si la observación forma parte de una unidad familiar o no.

```
viajeros$VALORACION_RECREO_NINYOS <- replace(viajeros$VALORACION_RECREO_N
INYOS, !is.na(viajeros$VALORACION_RECREO_NINYOS), "Si")
viajeros$VALORACION_RECREO_NINYOS <- replace(viajeros$VALORACION_RECREO_N
INYOS, is.na(viajeros$VALORACION_RECREO_NINYOS), "No")
```

## División del análisis

Teniendo en cuenta que el enunciado insta a emitir un informe que responda a *si existen grupos de turistas homogéneos a partir de la valoración que dan a los distintos servicios,*

[1] http://www.bbc.com/mundo/noticias/2014/10/141003_deportes_vert_cap_ricos_nc

centraremos nuestro análisis en las variables de valoración de servicios, siempre que el análisis exploratorio indique la posibilidad de agrupar.

Una vez tengamos los clusters en función de la puntuación que los viajeros dan a los servicios (satisfacción), prestaremos atención a la composición de los grupos en función de las variables cualitativas (Ingresos, ocupación, procedencia…) para entender si hay relaciones entre grupos sociales y la puntuación.

En este sentido, vamos a analizar la base de datos con dos procedimientos paralelos:

- en primer lugar, vamos a hacer un análisis clúster en función de los servicios generales que cualquier viajero puede puntuar. Alojamiento, clima, medio ambiente o gastronomía local son aspectos que cualquier turista puede puntuar puesto que son inherentes a cualquier visita.
- Por otro lado, vamos a hacer un análisis clúster en función de los servicios específicos que no son atribuíbles a todos los viajeros. A diferencia de los anteriores servicios, la práctica de deportes en general, las excursiones o el ambiente nocturno son actividades que no todo el mundo practica y que, por tanto, han de tratarse de forma distinta.

Si observamos los NA de la muestra, coincide que los servicios generales tienen menos NA que las actividades específicas Con esta metodología conseguimos salvar mucha información para el grupo general que, de otra forma, se perdería al omitir los NA o se distorsionaría al sustituir los NA por vecinos cercanos o la media de la variable.

## ANALISIS DE SATISFACCION GENERAL

```
viajeros_general<-viajeros[,c(1:16,19,23,28:35)]
viajeros_general<-na.omit(viajeros_general)
View(viajeros_general)
```

Nos queda una muestra de 11062 observaciones completas

¿Hay que tipificar? Si las variables presentan fuertes variaciones de rango o una alta variabilidad, conviene tipificarlas. En este caso, las variables cuantitativas no presentan fuertes variaciones de rango ni una alta variabilidad. Ademas, la media está cerca de la mediana. No consideramos necesario tipificar.

Observamos los estadísticos principales de viajeros_general (quito variables no numericas):

```
general_stats <- data.frame(
        Min = apply(viajeros_general[,c(3:16,19:22)], 2, min),
        Med = apply(viajeros_general[,c(3:16,19:22)], 2, median),
# mediana
        Mean = apply(viajeros_general[,c(3:16,19:22)], 2, mean),
# media
        SD = apply(viajeros_general[,c(3:16,19:22)], 2, sd),
```

```
#desv tipica
        Max = apply(viajeros_general[,c(3:16,19:22)], 2, max)
# Maximo
)
general_stats <- round(general_stats, 1)
head(general_stats)
```

```
##                            Min Med Mean  SD Max
## IMPRESION                    2   8  8.6 1.7  10
## VALORACION_ALOJ              1   8  8.0 1.8  10
## VALORACION_TRATO_ALOJ        1   9  8.3 1.8  10
## VALORACION_GASTRONO_ALOJ     1   8  7.5 2.1  10
## VALORACION_CLIMA             1   9  8.6 1.5  10
## VALORACION_ZONAS_BANYO       1   8  8.1 1.8  10
```

Como podemos observar, la distribución se concentra en torno al 8 en general (mediana), quedando un 50% a cada lado. La media es muy próxima a la mediana por lo que la influencia de los outliers no parece ser significativa. Esto de entrada nos da a entender que los viajeros están muy satisfechos o totalmente satisfechos con respecto a los servicios generales.

```
library(factoextra)
```

```
## Warning: package 'factoextra' was built under R version 3.4.2
```

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

```
## Warning: package 'ggplot2' was built under R version 3.4.2
```

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

```
library(cluster)
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.4.2
```

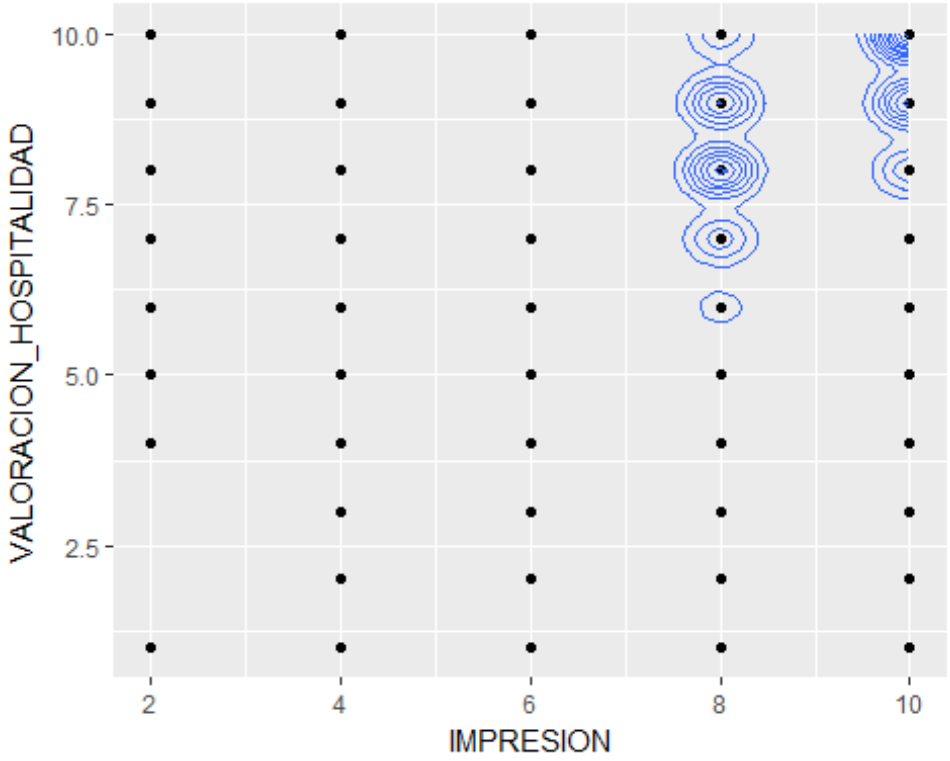
```
library(ggplot2)
```

## ANALISIS PREVIO: ¿TIENE SENTIDO PROCEDER A UN CLUSTER?

Aunque Emplearemos los packages factoextra para visualización, clustertend para evaluar la tendencia de agrupamiento y seriation para una evaluación visual de esa tendencia.

Podemos hacer una primera inspección visual de un par de variables de sec.esc a través de un gráfico de densidad 2D de 'ggplot

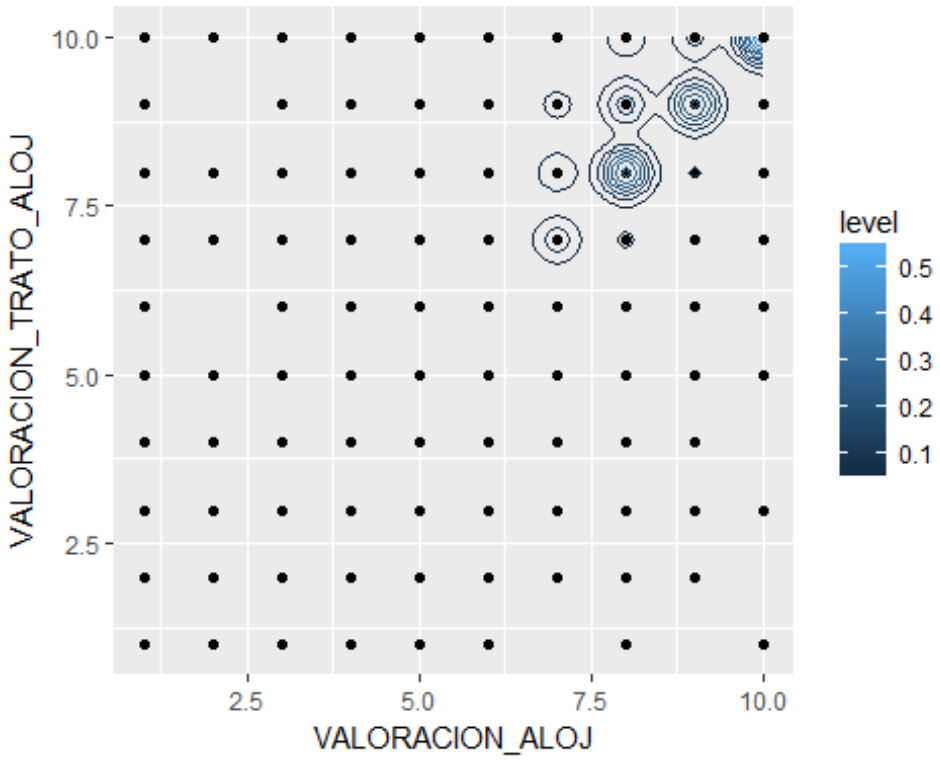
```
ggplot(as.data.frame(viajeros_general), aes(x=IMPRESION, y=VALORACION_HOS
PITALIDAD)) +
        geom_point() + # gráfico de dispersión
        geom_density_2d() # Estimación bidimensional de la densidad
```




Como podemos ver, los valores se concentran en valoraciones altas em ambas variables, tal y como predecían los estadísticos. Podemos hacer un gráfico para comprobar densidades y agrupamientos cruzando otras dos variables, por ejemplo, valoración del alojamiento y valoración del trato en el alojamiento.

```
pruebaG=viajeros_general[,4:5]
graf.datos = ggplot(pruebaG, aes(x = VALORACION_ALOJ, y = VALORACION_TRAT
O_ALOJ)) +
geom_point() +
stat_density2d(aes(color = ..level..))
graf.datos
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

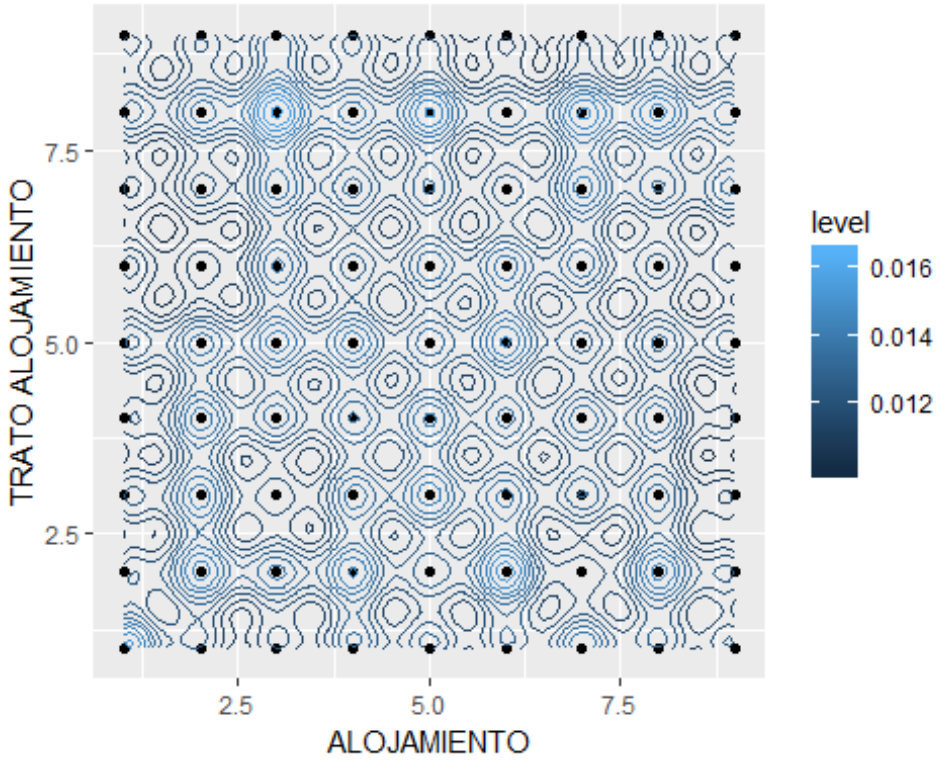



Como podemos observar, los agrupamientos tienen lugar en las valoraciones altas de ambos servicios.

Vamos a comparar la situación anterior con una distribución uniforme aleatoria de los datos, empleando la función runif(n, min, max) como sigue:

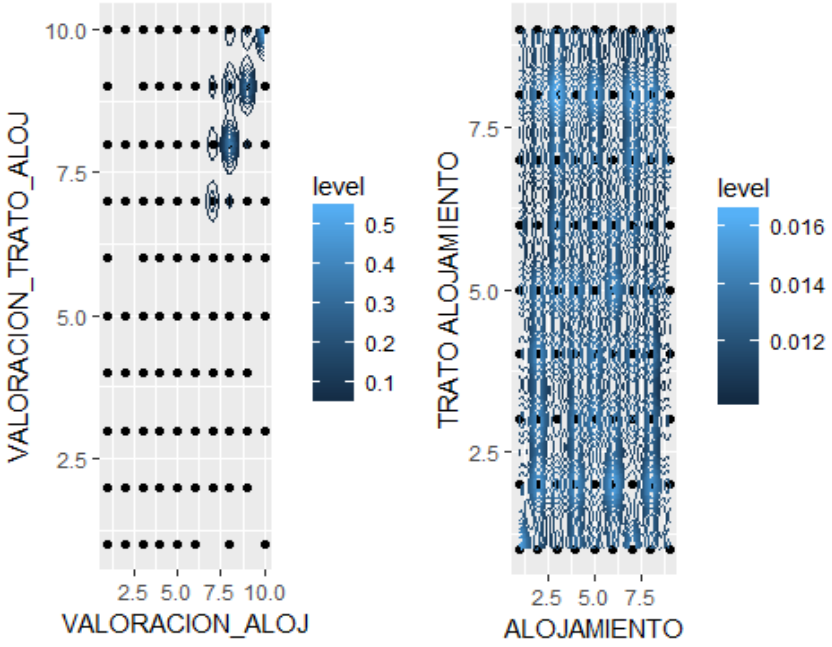
```
# Generamos un conjunto aleatorio de datos para las dos variables
set.seed(123)
n = nrow(pruebaG)
random_df = data.frame(
x = as.integer(runif(nrow(pruebaG), min(pruebaG$VALORACION_ALOJ), max(pru
ebaG$VALORACION_ALOJ))),
y = as.integer(runif(nrow(pruebaG), min(pruebaG$VALORACION_TRATO_ALOJ), m
ax(pruebaG$VALORACION_TRATO_ALOJ))))
# Colocamos en objeto para representación posterior
graf.aleat=ggplot(random_df, aes(x, y)) + geom_point() + labs(x="ALOJAMIE
NTO",y="TRATO ALOJAMIENTO") + stat_density2d(aes(color = ..level..))
graf.aleat
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu




Observamos que en este gráfico los perfiles son completamente distintos y, por tanto, nos sugiere la posibilidad de grupos frente a la distribución aleatoria de los datos.

```
# Exige haber empaquetado los objetos, como hemos hecho; equivale a par(m
frow=c(f, c))
grid.arrange(graf.datos, graf.aleat, nrow=1, ncol=2)
```



Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

## ANALISIS CLUSTER

Utilizaremos un método no jeráquico puesto que nuestras variables no son variables continuas. Emplearemos el método CLARA (Clustering LArge Applications) para realizar el análisis, puesto que permite trabajar de forma cómoda con grandes conjuntos de varios miles de observaciones.

Tomamos una muestra de 1.000 observaciones,

```
set.seed(555)
muestra_general = viajeros_general[,c(4:16,19:22)][sample(1:nrow(viajeros
_general[,c(4:16,19:22)]), 1000, replace=FALSE),]
summary(muestra_general)
```

```
##  VALORACION_ALOJ  VALORACION_TRATO_ALOJ VALORACION_GASTRONO_ALOJ
##  Min.   : 1.000   Min.   : 1.000        Min.   : 1.000
##  1st Qu.: 7.000   1st Qu.: 8.000        1st Qu.: 6.000
##  Median : 8.000   Median : 9.000        Median : 8.000
##  Mean   : 7.951   Mean   : 8.296        Mean   : 7.473
##  3rd Qu.: 9.000   3rd Qu.:10.000        3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.000        Max.   :10.000
##  VALORACION_CLIMA VALORACION_ZONAS_BANYO VALORACION_PAISAJES
##  Min.   : 1.000   Min.   : 1.000         Min.   : 1.000
##  1st Qu.: 8.000   1st Qu.: 7.000         1st Qu.: 7.000
##  Median : 9.000   Median : 8.000         Median : 9.000
##  Mean   : 8.623   Mean   : 8.159         Mean   : 8.334
##  3rd Qu.:10.000   3rd Qu.: 9.000         3rd Qu.:10.000
##  Max.   :10.000   Max.   :10.000         Max.   :10.000
##  VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
##  Min.   : 1.000            Min.   : 1.000          Min.   : 1.000
##  1st Qu.: 7.000            1st Qu.: 7.000          1st Qu.: 7.000
##  Median : 8.000            Median : 8.000          Median : 8.000
##  Mean   : 8.216            Mean   : 8.146          Mean   : 8.087
##  3rd Qu.: 9.000            3rd Qu.: 9.000          3rd Qu.: 9.000
##  Max.   :10.000            Max.   :10.000          Max.   :10.000
##  VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
##  Min.   : 1.000             Min.   : 1.000
##  1st Qu.: 7.000             1st Qu.: 7.000
##  Median : 8.000             Median : 8.000
##  Mean   : 7.702             Mean   : 7.443
##  3rd Qu.: 9.000             3rd Qu.: 9.000
##  Max.   :10.000             Max.   :10.000
##  VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR VALORACION_SEGURID
AD
##  Min.   : 1.000           Min.   : 1.000            Min.   : 1.000
##  1st Qu.: 7.000           1st Qu.: 7.000            1st Qu.: 8.000
##  Median : 8.000           Median : 8.000            Median : 8.000
##  Mean   : 8.093           Mean   : 7.644            Mean   : 8.219
```

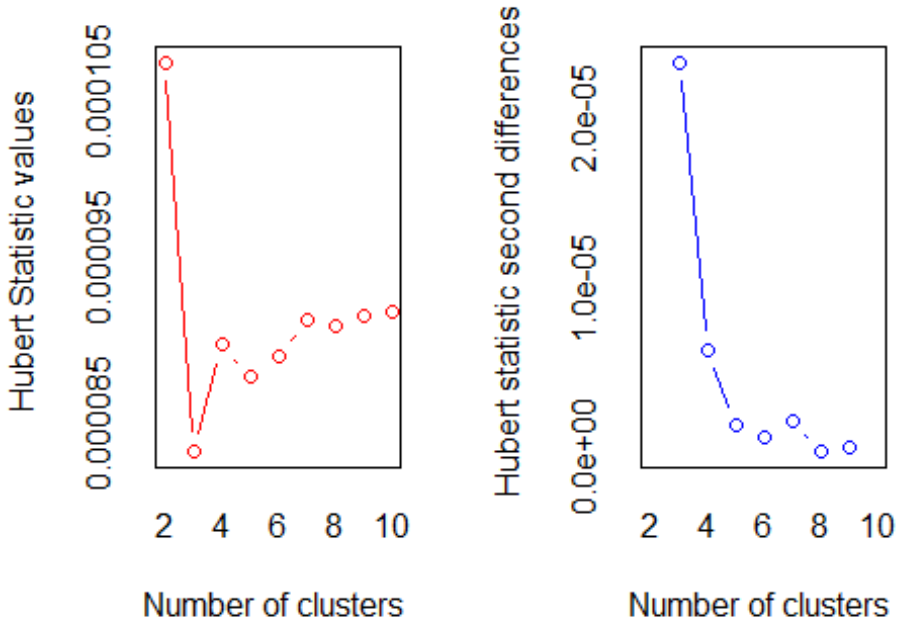
```
##  3rd Qu.: 9.000           3rd Qu.: 9.000           3rd Qu.: 9.000
##  Max.   :10.000           Max.   :10.000           Max.   :10.000
##  VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
##  Min.   : 1.000               Min.   : 1.000
##  1st Qu.: 7.000               1st Qu.: 6.000
##  Median : 8.000               Median : 8.000
##  Mean   : 7.745               Mean   : 7.476
##  3rd Qu.: 9.000               3rd Qu.: 9.000
##  Max.   :10.000               Max.   :10.000
##  VALORACION_HOSPITALIDAD
##  Min.   : 1.000
##  1st Qu.: 8.000
##  Median : 9.000
##  Mean   : 8.454
##  3rd Qu.:10.000
##  Max.   :10.000
```

Aplicamos ahora la función NbClust() para comprobar el número óptimo de clusters.

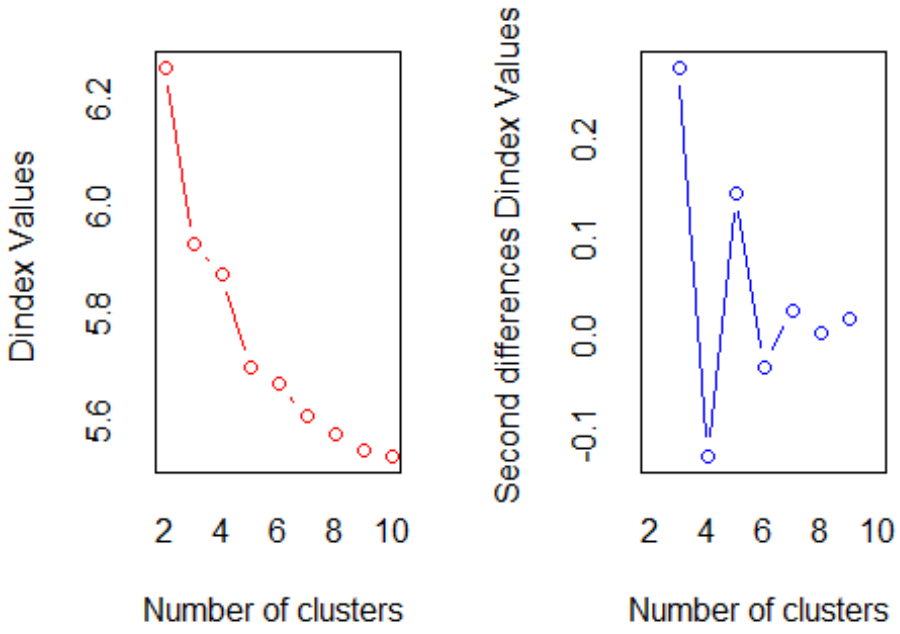
```
require(NbClust)

## Loading required package: NbClust

Nb.viajeros_general=NbClust(muestra_general, 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 kne
e that corresponds to a
##                 significant increase of the value of the measure i.e t
he significant peak in Hubert
##                 index second differences plot.
##
```

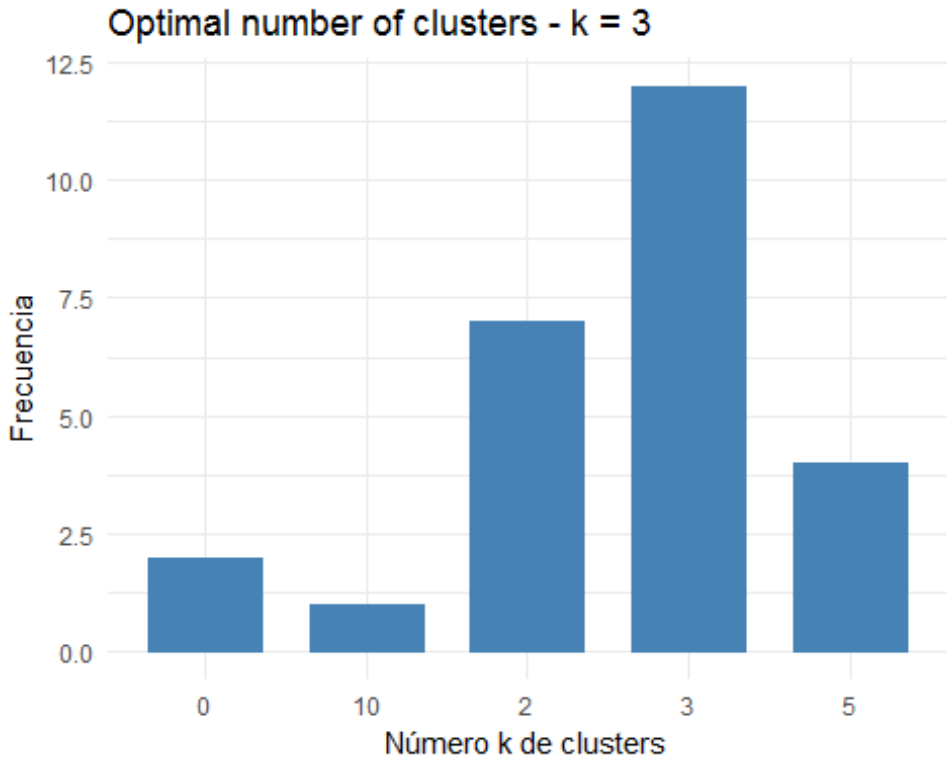



```
## *** : The D index is a graphical method of determining the number of c
lusters.
##                 In the plot of D index, we seek a significant knee (th
e significant peak in Dindex
##                 second differences plot) that corresponds to a signifi
cant increase of the value of
##                 the measure.
##
## *****************************************************************
## * Among all indices:
## * 7 proposed 2 as the best number of clusters
## * 12 proposed 3 as the best number of clusters
## * 4 proposed 5 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
##
##                    ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  3
```

```
##
##
## *******************************************************************
```

```
require(factoextra)
fviz_nbclust(Nb.viajeros_general) + theme_minimal() +
labs(x="Número k de clusters", y="Frecuencia")
```

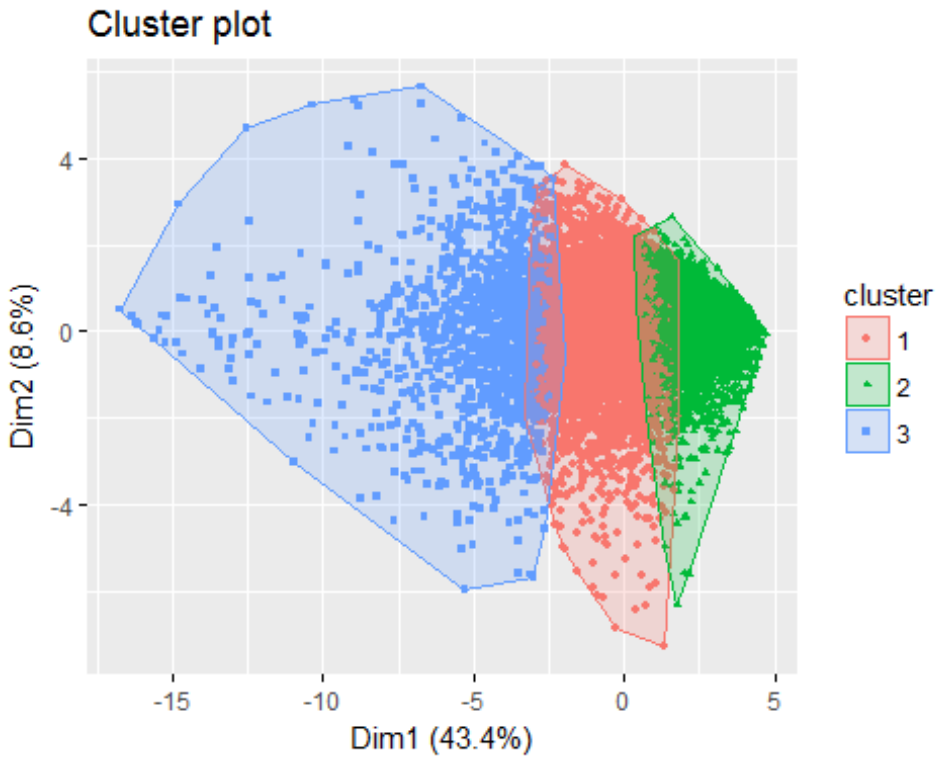
```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 7 proposed  2 as the best number of clusters
## * 12 proposed  3 as the best number of clusters
## * 4 proposed  5 as the best number of clusters
## * 1 proposed  10 as the best number of clusters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



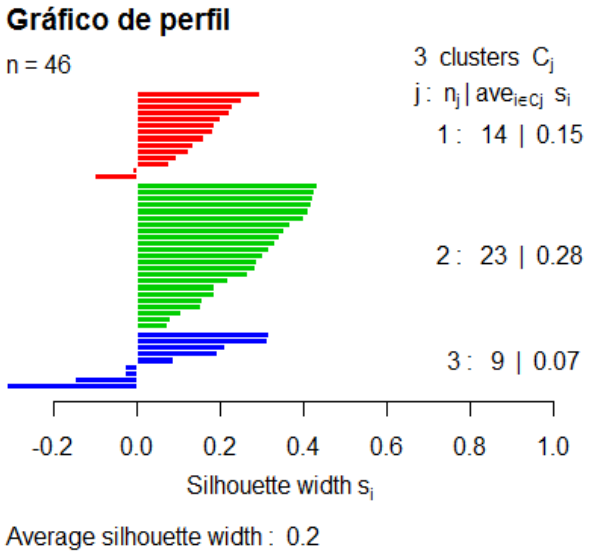

El procedimiento arroja que la opción mayoritaria es crear 3 grupos:

```
require(cluster)
viajeros_general.clara=clara(viajeros_general[,c(4:16,19:22)], 3, samples
=200)
require(factoextra)
```

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




```
plot(silhouette(viajeros_general.clara), col = 2:4, main = "Gráfico de perfil")
```




Como podemos observar, existen malas asignaciones tanto en el cluster 1 como en el 3. El perfil medio del cluster es muy bajo y además las desviaciones del perfil de los cluster difieren bastante del medio, en especial el 1 y el 3. En este sentido, vamos a utilizar las

variables con menos NA dentro de servicios generales asumiendo que de manera generalista las variables que mas gente ha contestado son las mas importantes dentro de la categoria general.

Volvemos a mirar las variables con menos NA como hicimos al principio.

```
viajeros_cuantit=viajeros_orig[,c(3:31)]

vacios<-c()
nom<-c()
for (i in (1:length(viajeros_cuantit))) {
        a<- sum(is.na(viajeros_cuantit[,i]))
        vacios<-c(vacios,a)
        b<-colnames(viajeros_cuantit[i])
        nom<- c(nom,b)
}
df_vacios<-data.frame(vacios,nom)
df_vacios[df_vacios$vacios<30000&df_vacios$vacios>0,]
```

```
##    vacios                          nom
## 1    7311                    IMPRESION
## 2    5327               VALORACION_ALOJ
## 3    8339         VALORACION_TRATO_ALOJ
## 4   14446      VALORACION_GASTRONO_ALOJ
## 5    3894               VALORACION_CLIMA
## 6    7442        VALORACION_ZONAS_BANYO
## 7    5901           VALORACION_PAISAJES
## 8    6627     VALORACION_MEDIO_AMBIENTE
## 9    5504       VALORACION_TRANQUILIDAD
## 10   5088           VALORACION_LIMPIEZA
## 11   9750    VALORACION_CALIDAD_RESTAUR
## 12  13208   VALORACION_OFERTA_GASTR_LOC
## 13  10620      VALORACION_TRATO_RESTAUR
## 14  12199     VALORACION_PRECIO_RESTAUR
## 15  29553             VALORACION_CULTURA
## 23  28041      VALORACION_SERVICIOS_BUS
## 24  25290     VALORACION_SERVICIOS_TAXI
## 25  28390          VALORACION_ALQ_VEHIC
## 26  23160          VALORACION_SEGURIDAD
## 27  14589 VALORACION_ESTADO_CARRETERAS
## 28  16912  VALORACION_CALIDAD_COMERCIO
## 29  14194       VALORACION_HOSPITALIDAD
```

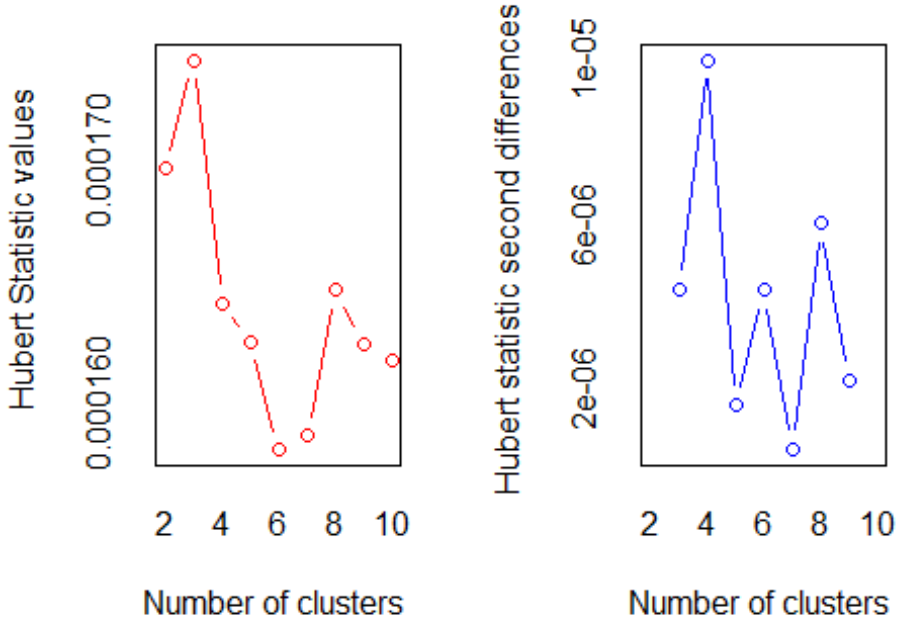
Así:

```
set.seed(555)
muestra_general2 = viajeros_general[,c(3,4,10,11,12,13,15,16,20)][sample(
1:nrow(viajeros_general[,c(3,4,10,11,12,13,15,16,20)]), 1000, replace=FAL
SE),]
summary(muestra_general2)
```

```
##    IMPRESION      VALORACION_ALOJ  VALORACION_MEDIO_AMBIENTE
##  Min.   : 2.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 8.000   1st Qu.: 7.000   1st Qu.: 7.000
##  Median : 8.000   Median : 8.000   Median : 8.000
##  Mean   : 8.642   Mean   : 7.951   Mean   : 8.216
##  3rd Qu.:10.000   3rd Qu.: 9.000   3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.000   Max.   :10.000
##  VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA VALORACION_CALIDAD_RESTAU
R
##  Min.   : 1.000          Min.   : 1.000      Min.   : 1.000
##  1st Qu.: 7.000          1st Qu.: 7.000      1st Qu.: 7.000
##  Median : 8.000          Median : 8.000      Median : 8.000
##  Mean   : 8.146          Mean   : 8.087      Mean   : 7.702
##  3rd Qu.: 9.000          3rd Qu.: 9.000      3rd Qu.: 9.000
##  Max.   :10.000          Max.   :10.000      Max.   :10.000
##  VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR
##  Min.   : 1.000           Min.   : 1.000
##  1st Qu.: 7.000           1st Qu.: 7.000
##  Median : 8.000           Median : 8.000
##  Mean   : 8.093           Mean   : 7.644
##  3rd Qu.: 9.000           3rd Qu.: 9.000
##  Max.   :10.000           Max.   :10.000
##  VALORACION_ESTADO_CARRETERAS
##  Min.   : 1.000
##  1st Qu.: 7.000
##  Median : 8.000
##  Mean   : 7.745
##  3rd Qu.: 9.000
##  Max.   :10.000
```

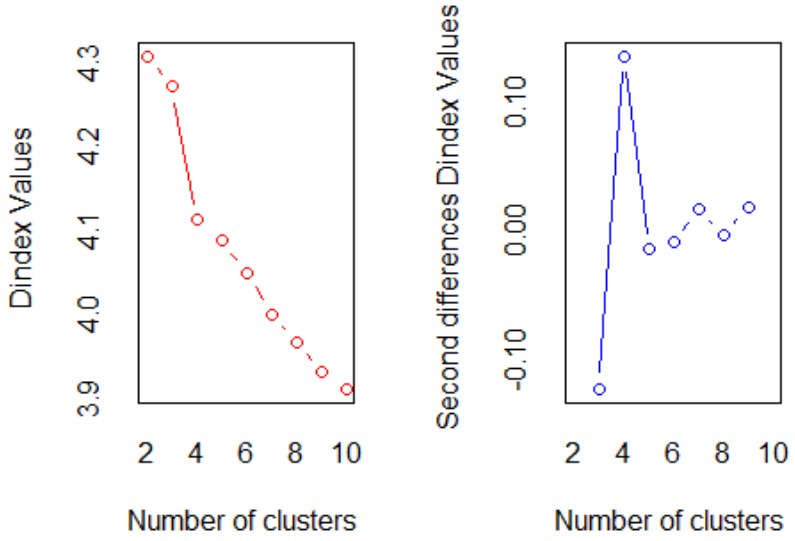
```
require(NbClust)
Nb.viajeros_general2=NbClust(muestra_general2, distance = "euclidean", mi
n.nc = 2,
max.nc = 10, method = "complete", index ="all")
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu




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

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

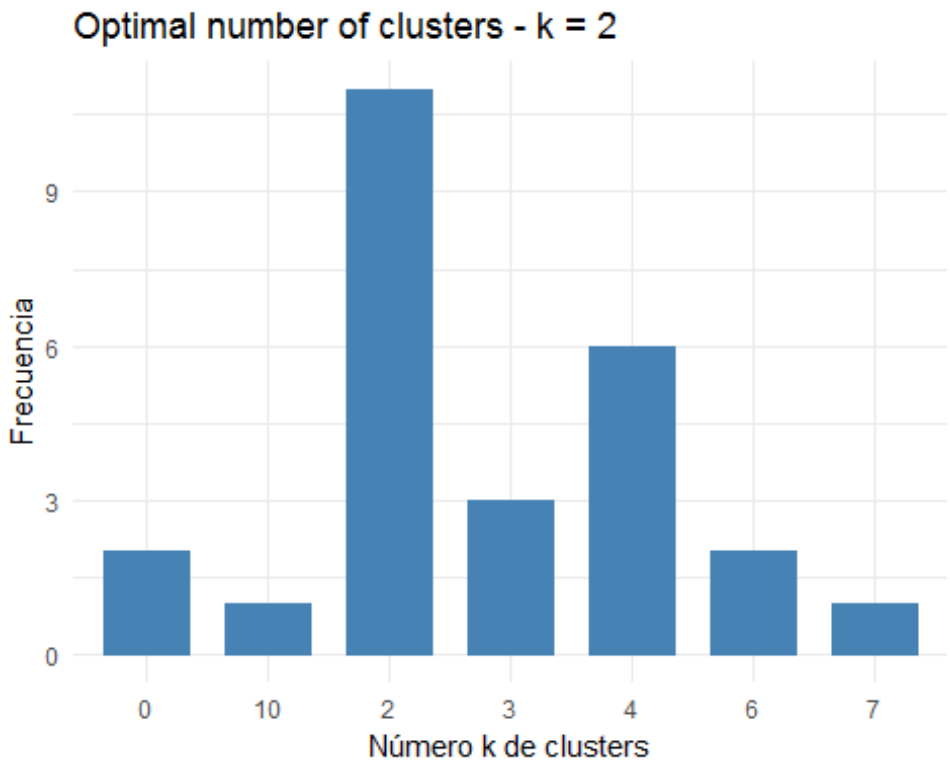



```
## *** : The D index is a graphical method of determining the number of c
lusters.
##                 In the plot of D index, we seek a significant knee (th
e significant peak in Dindex
##                 second differences plot) that corresponds to a signifi
cant increase of the value of
##                 the measure.
##
## *******************************************************************
## * Among all indices:
## * 11 proposed 2 as the best number of clusters
## * 3 proposed 3 as the best number of clusters
## * 6 proposed 4 as the best number of clusters
## * 2 proposed 6 as the best number of clusters
## * 1 proposed 7 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

```
require(factoextra)
fviz_nbclust(Nb.viajeros_general2) + theme_minimal() +
labs(x="Número k de clusters", y="Frecuencia")
```

```
## Among all indices:
## ==================
```

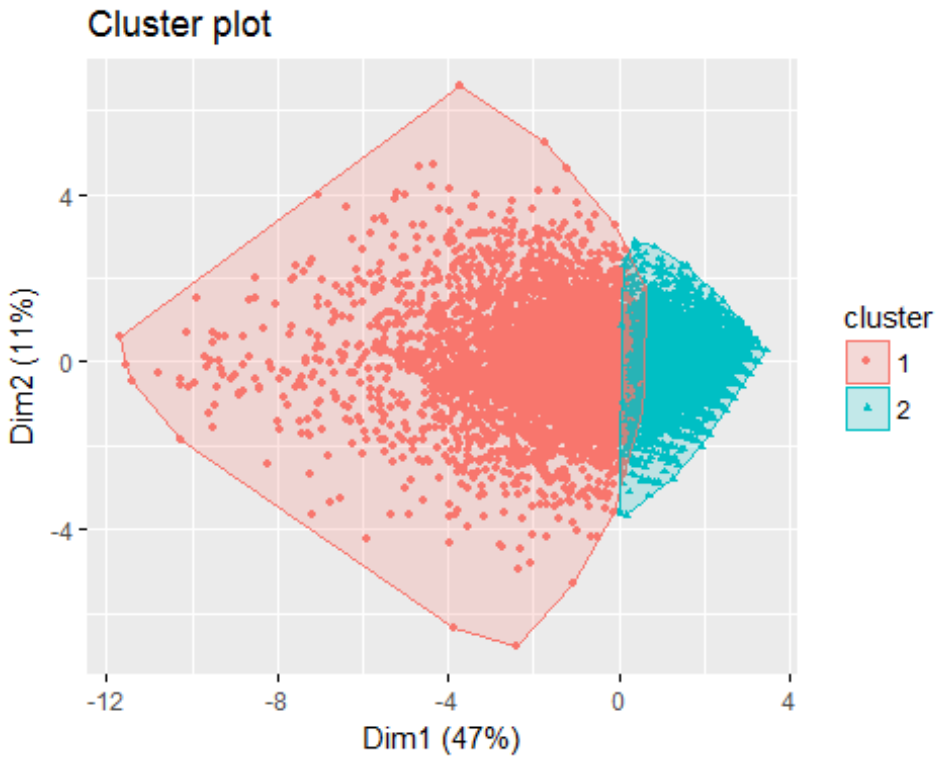
```
## * 2 proposed  0 as the best number of clusters
## * 11 proposed  2 as the best number of clusters
## * 3 proposed  3 as the best number of clusters
## * 6 proposed  4 as the best number of clusters
## * 2 proposed  6 as the best number of clusters
## * 1 proposed  7 as the best number of clusters
## * 1 proposed  10 as the best number of clusters
##
## Conclusion
## =========================
## * According to the majority rule, the best number of clusters is  2 .
```




En este caso, el conjunto de los métodos de decisión apuntan a que el número óptimo de custers es 2.

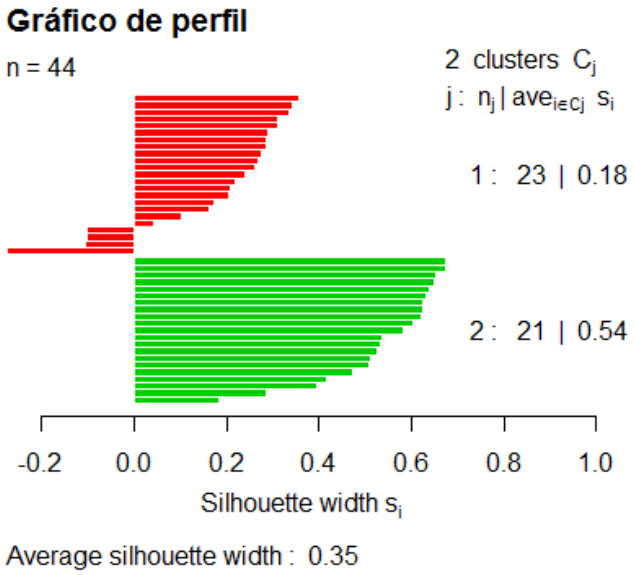
```
require(cluster)
viajeros_general.clara2=clara(viajeros_general[,c(3,4,10,11,12,13,15,16,20)], 2, samples=200)
require(factoextra)
fviz_cluster(viajeros_general.clara2, stand = TRUE, geom = "point", pointsize = 1)
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu




El grupo 1 corresponde a los individuos que están bastante satisfechos y a los pocos que están relativamente satisfechos. El grupo 2 corresponde a los individuos que están muy satisfechos. En la dimension 1 tenemos la variable mas representativa (47% de representación). La siguiente variable (Dim2) representa un 11%.

```
plot(silhouette(viajeros_general.clara2), col = 2:3, main = "Gráfico de p
erfil")
```



Como podemos observar, aunque nos mejora mucho el perfil, sigue sin ser demasiado bueno y sigue habiendo malas asignaciones (en el cluster 1 en este caso). Parece que la hipótesis planteada era correcta, aunque no consiga un resultado óptimo.

Aun sin ser un resultado demasiado satisfactorio. Concluimos que tenemos 2 clusters principales dentro de viajeros según el nivel de satisfacción respecto de los servicios y actividades mas generales.

*Matiz: es preciso destacar que durante el proceso hemos probado a eliminar outliers de distintas maneras, sin conseguir una mejora palpable del clustering.*

```
# require(data.table)
# viajeros<-data.table(viajeros)
#
#
# viajeros=viajeros[,ToKeep:= abs(EDAD-mean(EDAD)) < 3*sd(EDAD)][ToKeep =
= TRUE]
# viajeros=viajeros[,ToKeep:= abs(VALORACION_HOSPITALIDAD-mean(VALORACION
_HOSPITALIDAD)) < 3*sd(VALORACION_HOSPITALIDAD)][ToKeep == TRUE]
# viajeros=viajeros[,ToKeep:= abs(VALORACION_CALIDAD_COMERCIO-mean(VALORA
CION_CALIDAD_COMERCIO)) < 3*sd(VALORACION_CALIDAD_COMERCIO)][ToKeep == TR
UE]
# (… y así con todas las variables)
```

## COMPOSICION Y DESCRIPCION DE LOS CLUSTERS DE SATISFACCION GENERAL

En este apartado vamos a describir cómo se componen los clusters que hemos formado, teniendo en cuenta las variables cualitativas de viajeros_general, con el objetivo de buscar relaciones entre la satisfacción y los valores de estas variables.

Lo primero es crear una columna que indique el cluster al que pertenece cada observación.

```
CLUSTER<-viajeros_general.clara2$clustering #vector
viajeros_general<-cbind(viajeros_general,CLUSTER)
```

Vemos la composicion en nº de cada cluster:

```
viajeros_general.clara2
```

```
## Call:     clara(x = viajeros_general[, c(3, 4, 10, 11, 12, 13, 15, 16,
20)], k = 2, samples = 200)
## Medoids:
##        IMPRESION VALORACION_ALOJ VALORACION_MEDIO_AMBIENTE
## 145728         8               7                         7
## 4995          10               9                         9
##        VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
## 145728                       7                   7
## 4995                         9                   9
```

```
##        VALORACION_CALIDAD_RESTAUR VALORACION_TRATO_RESTAUR
## 145728                          7                        7
## 4995                            9                        9
##        VALORACION_PRECIO_RESTAUR VALORACION_ESTADO_CARRETERAS
## 145728                         7                           7
## 4995                           9                           9
## Objective function:   4.03106
## Clustering vector:    Named int [1:11062] 1 2 1 1 1 2 1 2 1 1 1 2 1 1
2 1 1 1 ...
##  - attr(*, "names")= chr [1:11062] "242037" "161764" "228332" "218161"
"146449" "219486" "265250" ...
## Cluster sizes:            5617 5445
## Best sample:
##  [1] 107753 19986  54921  132125 35717  36263  143112 257381 214535 19
5730
## [11] 31246  175823 173204 13882  252112 231426 77059  9304   5859   21
3025
## [21] 28099  91029  231028 216842 259718 95971  220023 45009  145728 12
647
## [31] 49280  253444 62902  119311 142724 247817 112829 188764 169210 49
95
## [41] 241581 44622  157126 52583
##
## Available components:
##  [1] "sample"     "medoids"    "i.med"      "clustering" "objective"
##  [6] "clusinfo"   "diss"       "call"       "silinfo"    "data"
```

Como vemos, los grupos son muy parecidos en tamaño. El cluster 1 tiene 5617 observaciones, mientras que el cluster 2 tiene 5445 observaciones.

### Composicion por pais

```
table(viajeros_general$PAIS_RESID_AGRUP, viajeros_general$CLUSTER)
```

```
##
##                   1    2
##   Alemania    1211  885
##   España      1309 1104
##   Otros       1925 1752
##   Reino Unido 1172 1704
```

Como podemos observar, el grupo de l tiene más presencia absoluta de alemanes, españoles y otras nacionalidades que el grupo 2, que tiene más presencia británica (31% del total del cluster) que el cluster 1. Se podría interpretar que, en cierto sentido, los británicos son más fáciles de satisfacer o menos exigentes. El grupo mas exigente son los alemanes, con un 42% en el grupo de los más satisfechos.

### Composicion por ingresos

```
table(viajeros_general$INGRESOS, viajeros_general$CLUSTER)
```

```
##
##                        1    2
##   De 12000 a 24000 1086 1074
##   De 24001 a 36000 1157 1120
##   De 36001 a 48000  919  930
##   De 48001 a 60000  803  800
##   De 60001 a 72000  482  470
##   De 72001 a 84000  324  303
##   Más de 84000      846  748
```

Como vemos, existe una composición bastante homogénea respecto a los ingresos. No parece una variable demasiado determinante. Si bien es cierto, los viajeros con mas ingresos tienen levemente a estar mas insatisfechos, pudiendo ser debido a sus altas expectativas.

**Relacion entre ingresos y golf**

```
table(viajeros_general$INGRESOS, viajeros_general$VALORACION_GOLF)
```

```
##
##                        No   Si
##   De 12000 a 24000  978 1182
##   De 24001 a 36000 1256 1021
##   De 36001 a 48000 1084  765
##   De 48001 a 60000  954  649
##   De 60001 a 72000  573  379
##   De 72001 a 84000  366  261
##   Más de 84000      901  693
```

Sorprendentemente, descubrimos que las personas que menores ingresos tienen son las que más han puntuado el golf (por tanto, interpretamos que las que mas han jugado).

**Composición por alojamiento**

```
table(viajeros_general$ALOJ_CATEG_1, viajeros_general$CLUSTER)
```

```
##
##                                                    1    2
##   Extrahoteleros                                1558 1379
##   Hoteles - apartahoteles de 4 estrellas        2320 2327
##   Hoteles - apartahoteles de 5 estrellas         341  517
##   Hoteles - apartahoteles de hasta 3 estrellas  1047  749
##   Otros tipos de alojamientos                    155  159
##   Viviendas propias o casas de amigos o familiares  196  314
```

En proporción, puede observarse como los alojamientos que mas satisfacción producen en terminos relativos son los hoteles de 5 estrellas y las viviendas propias o casas de amigos o familiares. Por el contrario, los menos satisfactorios son los hoteles y apartahoteles de hasta 3 estrellas, seguido de los extrahoteleros.

### Composición por sexo

```
table(viajeros_general$SEXO, viajeros_general$CLUSTER)
```

```
##
##             1    2
##   Hombre 3281 2844
##   Mujer  2336 2601
```

Puede observarse que, el 46% de los hombres se sitúa en el grupo de individuos que reportan mas satisfacción, mientras que en este grupo se incluyen el 52% de las mujeres. Así,aunque las diferencias no son muy relevantes, se podría afirmar que en términos relativos los hombres son mas exigentes que las mujeres.

### Composición por ocupación

```
table(viajeros_general$OCUPACION, viajeros_general$CLUSTER)
```

```
##
##                                1    2
##   Ama de casa                 48   64
##   Asalariado alta dirección  629  554
##   Asalariado cargo medio    1760 1622
##   Asalariado nivel auxiliar  374  385
##   Autónomo - profesión liberal  644  695
##   Empresario                 733  724
##   Estudiante                 369  317
##   Jubilado <U+0096> retirado          313  361
##   Otros trabajadores y obreros  656  645
##   Parado                      91   78
```

En cuanto a la ocupación, los viajeros que trabajan de amas de casa son los mas satisfechos, mientras que los menos satisfechos son los parados y los estudiantes.

### Composición por familias

```
table(viajeros_general$VALORACION_RECREO_NINYOS, viajeros_general$CLUSTER
)
```

```
##
##         1    2
##   No 2680 2775
##   Si 2937 2670
```

En el grupo 1 ha habido mas observaciones sobre las opciones de recreo de los niños. Entendemos por tanto, que en el grupo 1 hay mas unidades familiares con hijos.

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

## ANALISIS DE SATISFACCION EN ACTIVIDADES ESPECIFICAS

Matiz: En este análisis tenemos los viajeros que han participado en las actividades específicas. Una característica de las observaciones en este caso, es que TODOS los individuos han participado en TODAS las actividades/servicios específicos, dejando fuera a aquellos que solo han participado en algunos.

```
viajeros_esp<-viajeros[,c(1,2,17:28,32:35)]
viajeros_esp<-na.omit(viajeros_esp)
View(viajeros_esp)
```

Conseguimos una muestra de 5241 observaciones completas

Observo los estadísticos principales de viajeros_esp (quito variables no numericas):

```
esp_stats <- data.frame(
        Min = apply(viajeros_esp[,c(3,4,6:8,10:14)], 2, min),
        Med = apply(viajeros_esp[,c(3,4,6:8,10:14)], 2, median),
# mediana
        Mean = apply(viajeros_esp[,c(3,4,6:8,10:14)], 2, mean),
# media
        SD = apply(viajeros_esp[,c(3,4,6:8,10:14)], 2, sd),
#desv tipica
        Max = apply(viajeros_esp[,c(3,4,6:8,10:14)], 2, max)
# Maximo
)
esp_stats <- round(esp_stats, 1)
head(esp_stats)
```

```
##                               Min Med Mean  SD Max
## VALORACION_CULTURA              1   7  7.1 2.0  10
## VALORACION_DEPORTES             1   8  7.4 2.0  10
## VALORACION_PARQUES_OCIO         1   8  7.1 2.3  10
## VALORACION_AMBIENTE_NOCTURNO    1   8  7.1 2.2  10
## VALORACION_EXCURSIONES          1   8  7.3 2.1  10
## VALORACION_SALUD                1   8  7.2 2.1  10
```

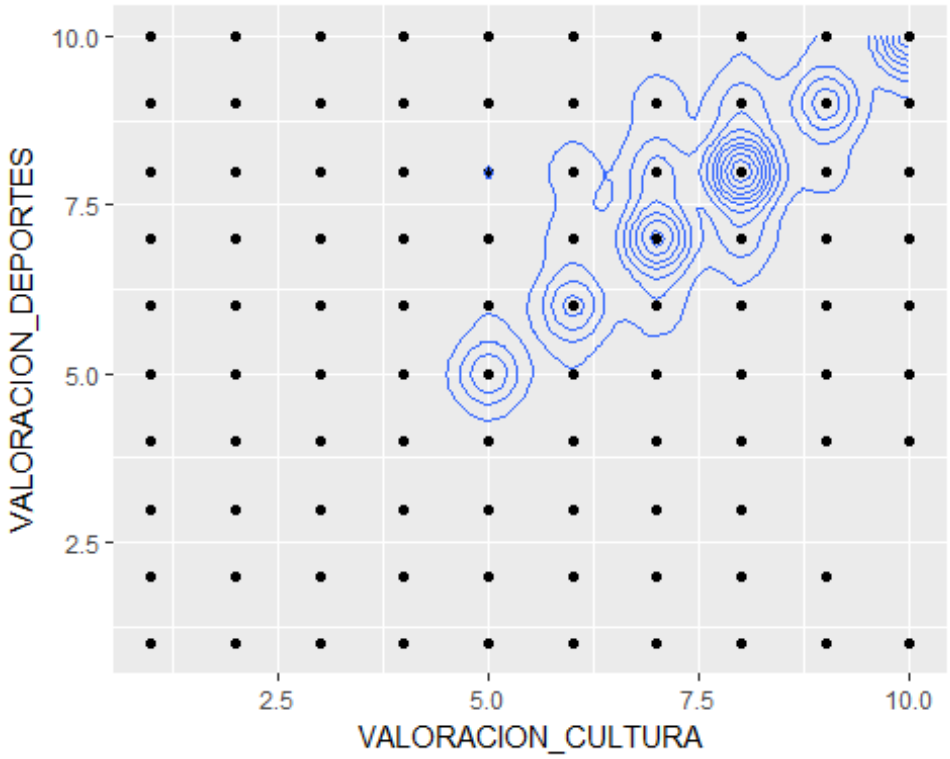
Como podemos observar, la distribución se concentra en torno al 8 en general (mediana), quedando un 50% a cada lado. La media es muy próxima a la mediana por lo que la influencia de los outliers no parece ser significativa. Esto de entrada nos da a entender que los viajeros están muy satisfechos o totalmente satisfechos con respecto a los servicios específicos. En estos servicios, sin embargo, los outliers si son mas notables pues la media varía mas respecto de la mediana que en el caso de los servicios generales y la desviación típica es mayor.

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

## ANALISIS PREVIO: ¿TIENE SENTIDO PROCEDER A UN CLUSTER?

Volvemos a realizar el mismo proceso que con viajeros_general, pero con viajeros_esp

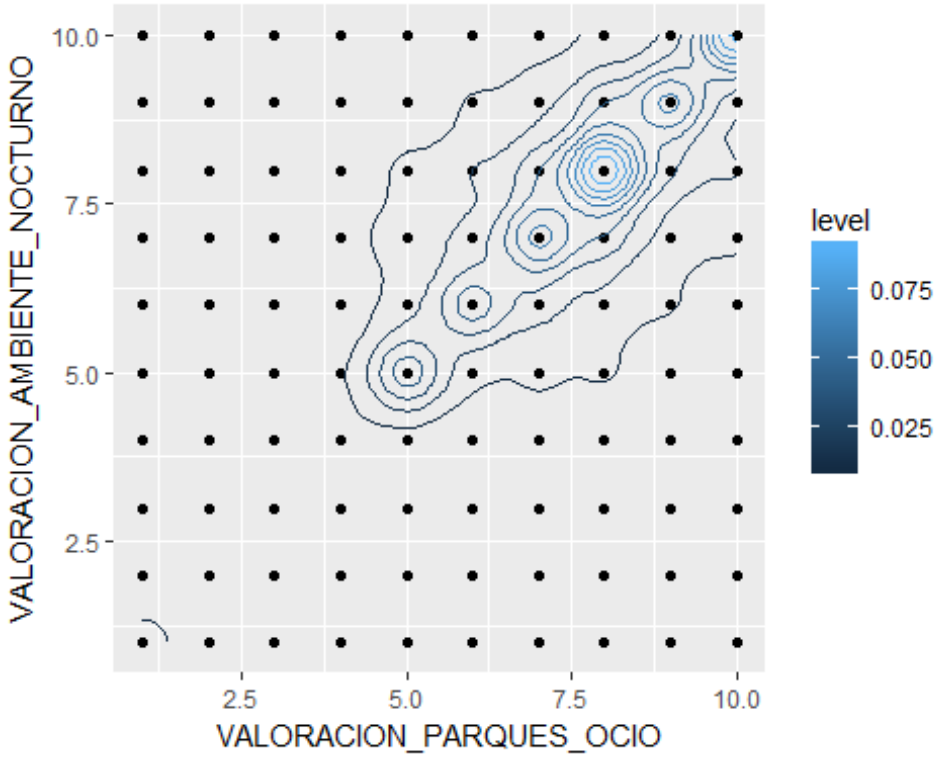
```
ggplot(as.data.frame(viajeros_esp), aes(x=VALORACION_CULTURA, y=VALORACIO
N_DEPORTES)) +
        geom_point() + # gráfico de dispersión
        geom_density_2d() # Estimación bidimensional de la densidad
```




Como se puede observar, en las actividades específicas haymas diferencias en cuanto a la satisfaccion, habiendo una mayor concentracion entre el aprobado y el sobresaliente.

Podemos hacer un gráfico para comprobar densidades y agrupamientos cruzando otras dos variables, por ejemplo, Valoración de parques de ocio y valoración de ambiente nocturno.

```
pruebaE=viajeros_esp[,6:7]
graf.datos2 = ggplot(pruebaE, aes(x = VALORACION_PARQUES_OCIO, y = VALORA
CION_AMBIENTE_NOCTURNO)) +
geom_point() +
stat_density2d(aes(color = ..level..))
graf.datos2
```

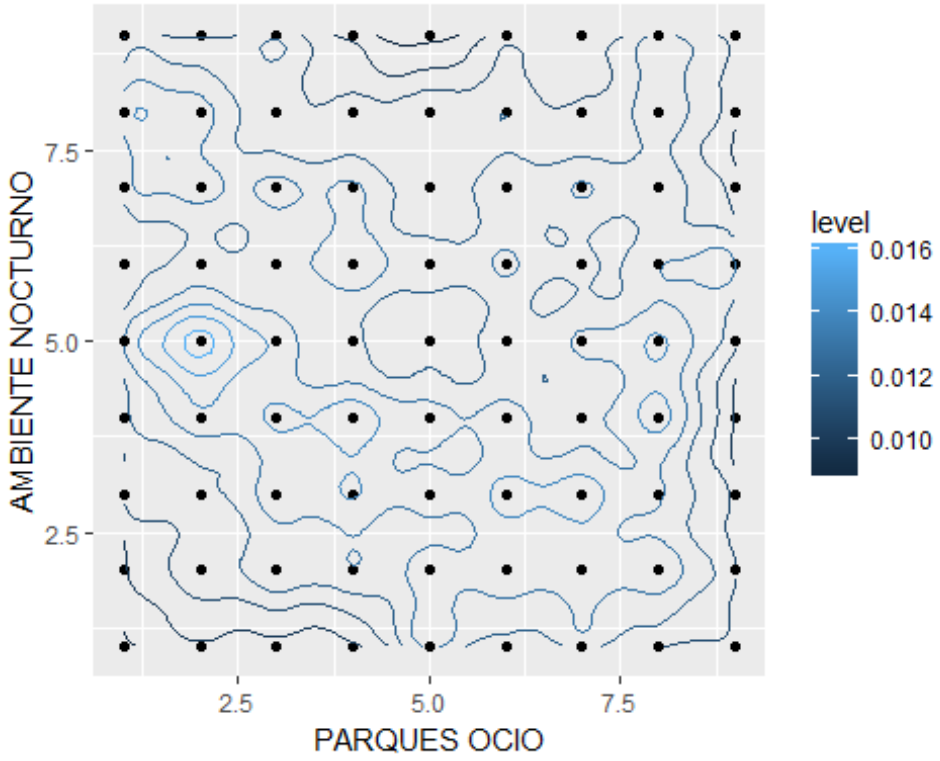



Como podemos observar, los agrupamientos tienne lugar entre el 5 y el 10, con una mayor concentración en torno al 8.

Podemos comparar la situación anterior con una distribución uniforme aleatoria de los datos, empleando la función runif(n, min, max) como sigue:

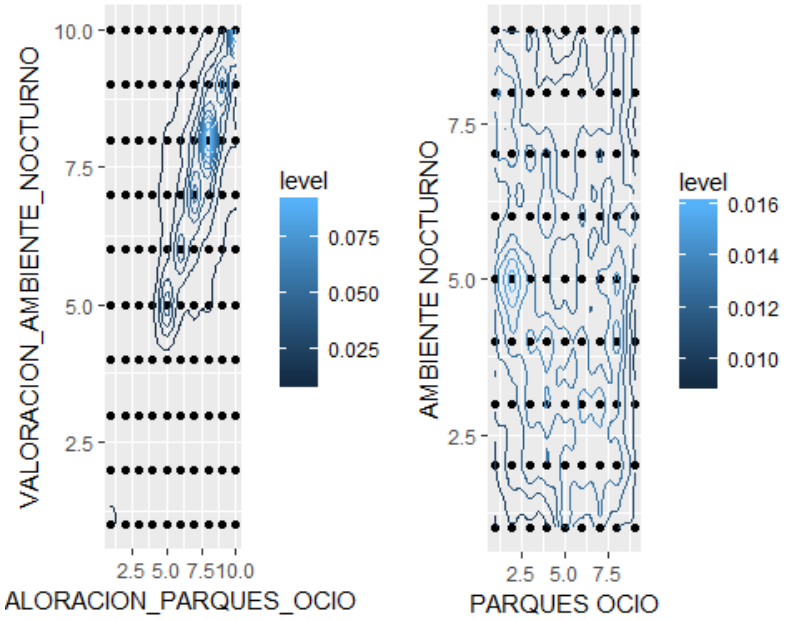
```
# Generamos un conjunto aleatorio de datos para las dos variables
set.seed(123)
n = nrow(pruebaE)
random_df2 = data.frame(
x = as.integer(runif(nrow(pruebaE), min(pruebaE$VALORACION_PARQUES_OCIO),
max(pruebaE$VALORACION_PARQUES_OCIO))),
y = as.integer(runif(nrow(pruebaE), min(pruebaE$VALORACION_AMBIENTE_NOCTU
RNO), max(pruebaE$VALORACION_AMBIENTE_NOCTURNO))))
# Colocamos en objeto para representación posterior
graf.aleat2=ggplot(random_df2, aes(x, y)) + geom_point() + labs(x="PARQUE
S OCIO",y="AMBIENTE NOCTURNO") + stat_density2d(aes(color = ..level..))
graf.aleat2
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu




Como podemos observar en el gráfico, los perfiles son completamente distintos y, por tanto, nos sugiere la posibilidad de grupos frente a la distribución aleatoria de los datos.

```
grid.arrange(graf.datos2, graf.aleat2, nrow=1, ncol=2)
```



Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

## ANALISIS CLUSTER

Utilizaremos un método no jeráquico puesto que nuestras variables no son variables continuas.

Empleamos el método CLARA (Clustering LArge Applications) para realizar el análisis, puesto que permite trabajar de forma cómoda con grandes conjuntos de varios miles de observaciones.

Tomamos una muestra de 1.000 observaciones,

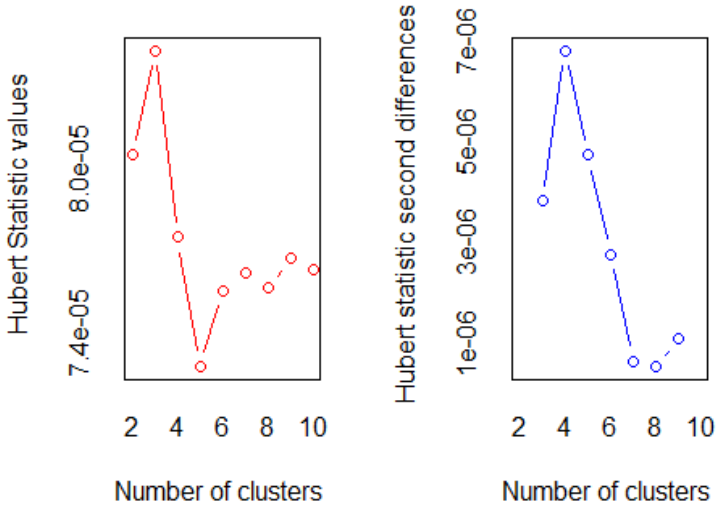
```
set.seed(555)
muestra_esp = viajeros_esp[,c(3,4,6:8,10:14)][sample(1:nrow(viajeros_esp[
,c(3,4,6:8,10:14)]), 1000, replace=FALSE),]
summary(muestra_esp)
```

```
##  VALORACION_CULTURA VALORACION_DEPORTES VALORACION_PARQUES_OCIO
##  Min.   : 1.00      Min.   : 1.000      Min.   : 1.000
##  1st Qu.: 6.00      1st Qu.: 6.000      1st Qu.: 6.000
##  Median : 7.00      Median : 8.000      Median : 8.000
##  Mean   : 7.08      Mean   : 7.451      Mean   : 7.141
##  3rd Qu.: 9.00      3rd Qu.: 9.000      3rd Qu.: 9.000
##  Max.   :10.00      Max.   :10.000      Max.   :10.000
##  VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES VALORACION_SALUD
##  Min.   : 1.000               Min.   : 1.000         Min.   : 1.000
##  1st Qu.: 6.000               1st Qu.: 6.000         1st Qu.: 6.000
##  Median : 8.000               Median : 8.000         Median : 8.000
##  Mean   : 7.183               Mean   : 7.292         Mean   : 7.239
##  3rd Qu.: 9.000               3rd Qu.: 9.000         3rd Qu.: 9.000
##  Max.   :10.000               Max.   :10.000         Max.   :10.000
##  VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI VALORACION_ALQ_VEH
IC
##  Min.   : 1.00            Min.   : 1.000            Min.   : 1.000
##  1st Qu.: 6.00            1st Qu.: 7.000            1st Qu.: 7.000
##  Median : 8.00            Median : 8.000            Median : 8.000
##  Mean   : 7.44            Mean   : 7.991            Mean   : 7.737
##  3rd Qu.: 9.00            3rd Qu.: 9.000            3rd Qu.: 9.000
##  Max.   :10.00            Max.   :10.000            Max.   :10.000
##  VALORACION_SEGURIDAD
##  Min.   : 1.000
##  1st Qu.: 7.000
##  Median : 8.000
##  Mean   : 8.009
##  3rd Qu.: 9.000
##  Max.   :10.000
```

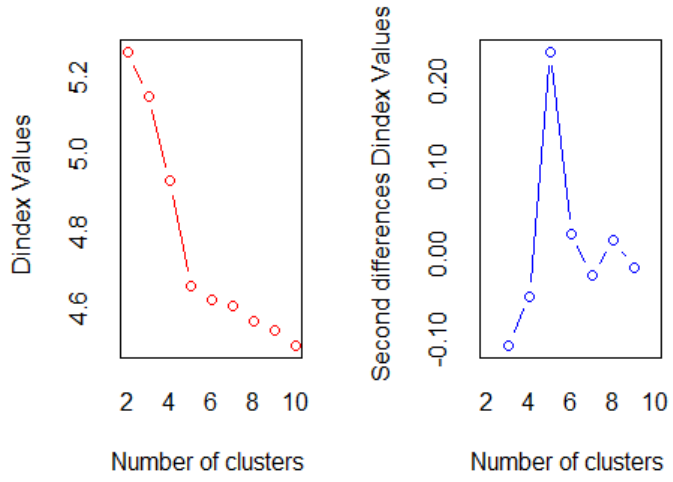
Aplicamos ahora la función NbClust()

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

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu




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

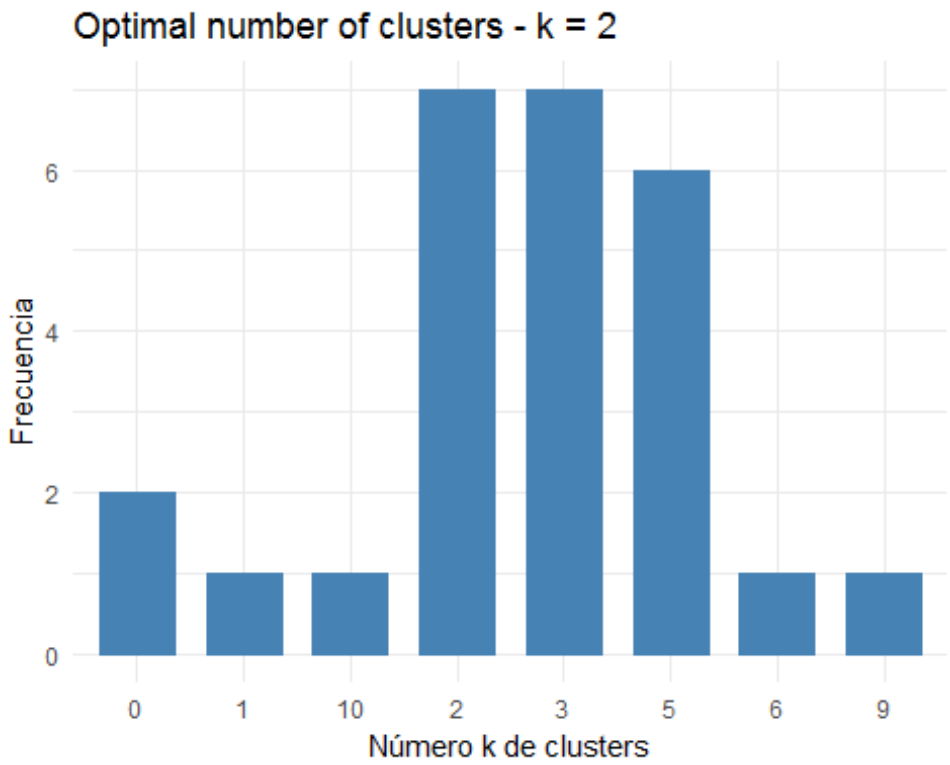



```
## *** : The D index is a graphical method of determining the number of c
lusters.
##                 In the plot of D index, we seek a significant knee (th
e significant peak in Dindex
```

```
##                  second differences plot) that corresponds to a signifi
cant increase of the value of
##                  the measure.
##
## *******************************************************************
## * Among all indices:
## * 7 proposed 2 as the best number of clusters
## * 7 proposed 3 as the best number of clusters
## * 6 proposed 5 as the best number of clusters
## * 1 proposed 6 as the best number of clusters
## * 1 proposed 9 as the best number of clusters
## * 1 proposed 10 as the best number of clusters
##
##                     ***** Conclusion *****
##
## * According to the majority rule, the best number of clusters is  2
##
##
## *******************************************************************
```

```
require(factoextra)
fviz_nbclust(Nb.viajeros_esp) + theme_minimal() +
labs(x="Número k de clusters", y="Frecuencia")
```

```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 1 proposed  1 as the best number of clusters
## * 7 proposed  2 as the best number of clusters
## * 7 proposed  3 as the best number of clusters
## * 6 proposed  5 as the best number of clusters
## * 1 proposed  6 as the best number of clusters
## * 1 proposed  9 as the best number of clusters
## * 1 proposed  10 as the best number of clusters
##
## Conclusion
## ==========================
## * According to the majority rule, the best number of clusters is  2 .
```

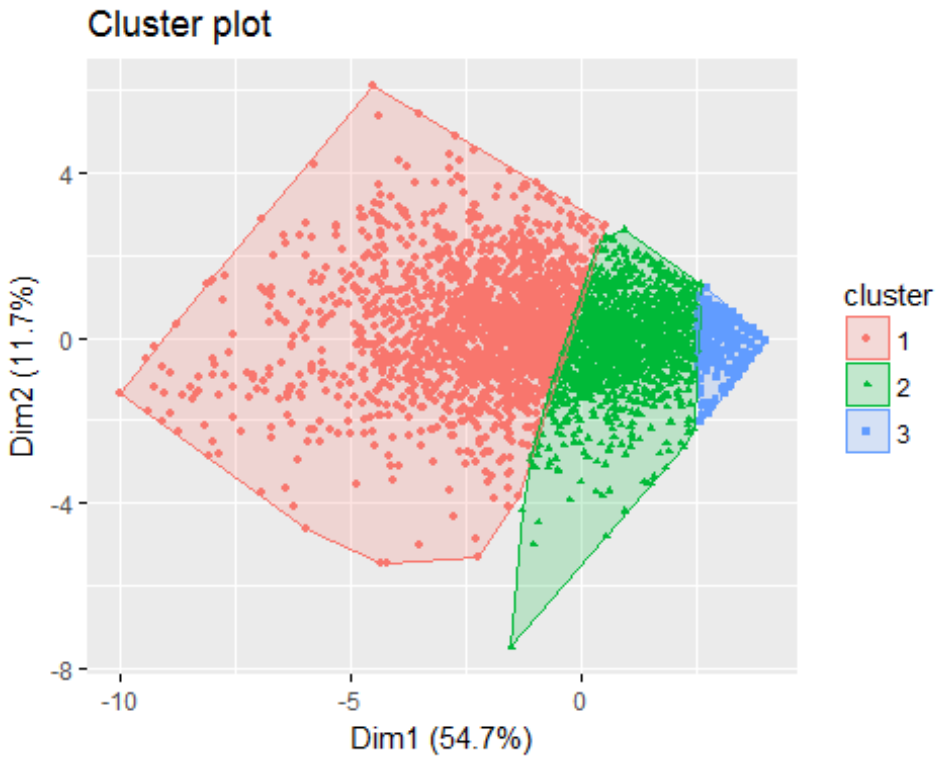



Obtenemos que el número de clusters óptimo es 2 según 7 métodos, aunque el mismo número de métodos indican que 3 también es el número óptimo.

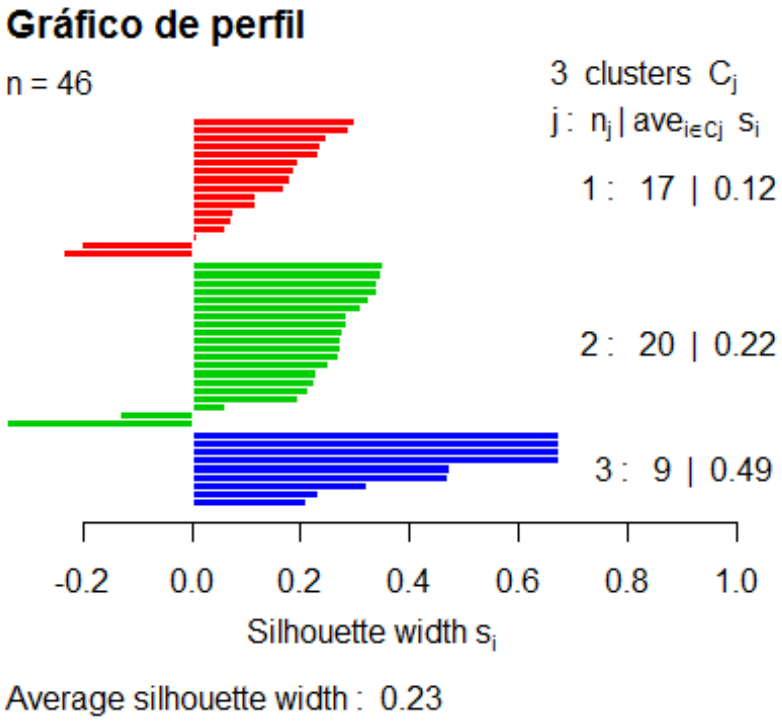
Probamos con 3 clusters

```
require(cluster)
viajeros_esp.clara=clara(viajeros_esp[,c(3,4,6:8,10:14)], 3, samples=200)
require(factoextra)
fviz_cluster(viajeros_esp.clara, stand = TRUE, geom = "point", pointsize
= 1)
```

Manuel Carmona Cabello de Alba
manuel.carmona@cunef.edu

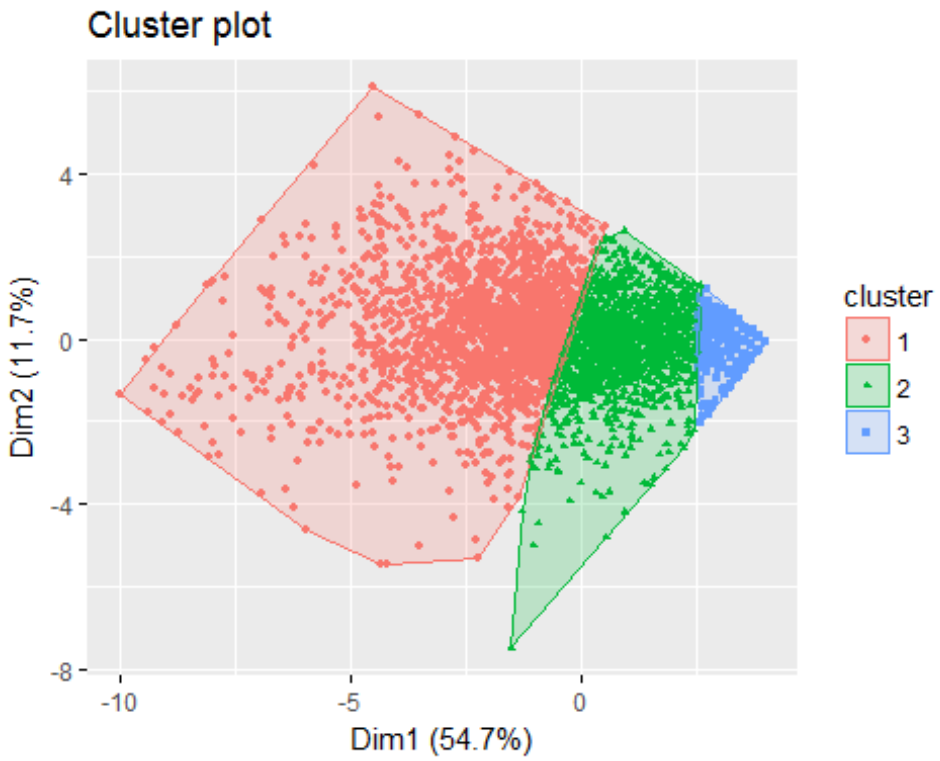



```
plot(silhouette(viajeros_esp.clara), col = 2:4, main = "Gráfico de perfil
")
```



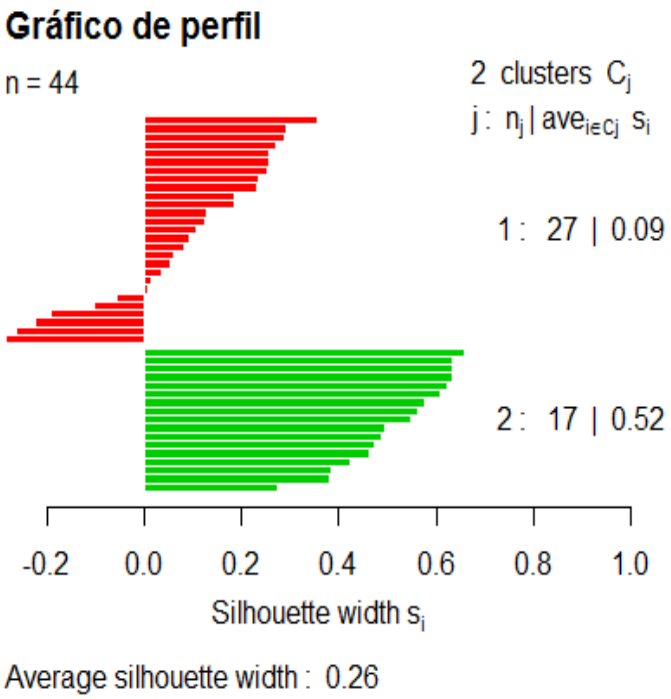
Existen malas asignaciones en el cluster 1 y el 2. El mejor cluster es el 3, con un perfil de 0.49.

```
require(cluster)
viajeros_esp.clara2=clara(viajeros_esp[,c(3,4,6:8,10:14)], 2, samples=200
)
require(factoextra)
fviz_cluster(viajeros_esp.clara, stand = TRUE, geom = "point", pointsize
= 1)
```




Desde esta perspectiva no se aprecia demasiado solapamiento.

```
plot(silhouette(viajeros_esp.clara2), col = 2:3, main = "Gráfico de perfi
l")
```




Como podemos observar, existen malas asignaciones en el cluster 1, que tiene un perfil muy bajo. Sin embargo, el perfil medio ha mejorado respecto al anterior, pasando de 0.23 a 0.26. Vamos a quedarnos con la división en 2 clusters ya que la desviación del perfil de los clusters respecto del perfil medio es parecida y el perfil medio es mejor.

## COMPOSICION Y DESCRIPCION DE LOS CLUSTERS DE SATISFACCION ESPECIFICA

En este apartado vamos a describir cómo se componen los clusters que hemos formado, teniendo en cuenta las variables cualitativas de viajeros_general, con el objetivo de buscar relaciones entre la satisfacción y los valores de estas variables.

Lo primero es crear una columna que indique el cluster al que pertenece cada observación.

```
CLUSTER<-viajeros_esp.clara2$clustering #vector
viajeros_esp<-cbind(viajeros_esp,CLUSTER)
```

Vemos la composicion en nº de cada cluster

```
viajeros_esp.clara2

## Call:     clara(x = viajeros_esp[, c(3, 4, 6:8, 10:14)], k = 2, sample
s = 200)
## Medoids:
```

```
##        VALORACION_CULTURA VALORACION_DEPORTES VALORACION_PARQUES_OCIO
## 36999                   6                   6                      6
## 203315                  9                   9                      9
##        VALORACION_AMBIENTE_NOCTURNO VALORACION_EXCURSIONES
## 36999                             6                     6
## 203315                            9                     9
##        VALORACION_SALUD VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_
TAXI
## 36999                 6                       7
7
## 203315                9                       9
9
##        VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
## 36999                     7                    7
## 203315                    9                    9
## Objective function:   4.806416
## Clustering vector:    Named int [1:5241] 1 1 1 1 1 1 2 2 1 2 2 1 2 1 1
2 2 2 ...
##  - attr(*, "names")= chr [1:5241] "251229" "242037" "161764" "228332"
"146449" "219486" "254647" ...
## Cluster sizes:            2864 2377
## Best sample:
##  [1] 228332 215237 90868  100807 129247 231560 10553  8860   208792 20
5717
## [11] 93865  245056 9199   203315 213752 164063 177737 50417  153502 25
9214
## [21] 212191 22223  120861 173330 10454  141595 155104 240451 227286 24
8583
## [31] 34339  186859 126895 36999  13228  264448 31343  247879 68184  12
4006
## [41] 147588 80350  201883 150944
##
## Available components:
##  [1] "sample"     "medoids"    "i.med"      "clustering" "objective"
##  [6] "clusinfo"   "diss"       "call"       "silinfo"    "data"
```

Como vemos, los grupos son muy parecidos en tamaño. El cluster 1 tiene 2864 observaciones, mientras que el cluster 2 tiene 2377 observaciones.

### Composicion por pais

```
table(viajeros_esp$PAIS_RESID_AGRUP, viajeros_esp$CLUSTER)
```

```
##
##                1   2
##   Alemania    524 404
##   España      658 458
##   Otros       978 795
##   Reino Unido 704 720
```

En este caso cambian las tornas, los mas satisfechos por lo general son los británicos, mientras que los españoles son los menos satisfechos.

**Composicion por ingresos**

```
table(viajeros_esp$INGRESOS, viajeros_esp$CLUSTER)
```

```
##
##                       1   2
##   De 12000 a 24000 679 606
##   De 24001 a 36000 589 544
##   De 36001 a 48000 448 375
##   De 48001 a 60000 356 301
##   De 60001 a 72000 215 182
##   De 72001 a 84000 158 112
##   Más de 84000     419 257
```

Como vemos, existe una composición menos homogénea respecto a los ingresos. Los viajeros con mas ingresos tienden a estar menos satisfechos en términos relativos.

**Composición por alojamiento**

```
table(viajeros_esp$ALOJ_CATEG_1, viajeros_esp$CLUSTER)
```

```
##
##                                                    1    2
##   Extrahoteleros                                 879  657
##   Hoteles - apartahoteles de 4 estrellas        1064  934
##   Hoteles - apartahoteles de 5 estrellas         149  158
##   Hoteles - apartahoteles de hasta 3 estrellas   476  364
##   Otros tipos de alojamientos                     93   72
##   Viviendas propias o casas de amigos o familiares 203 192
```

Se vuelve a repetir el patrón. En proporción, puede observarse como los alojamientos que mas satisfacción producen en terminos relativos son los hoteles de 5 estrellas y las viviendas propias o casas de amigos o familiares. Por el contrario, los menos satisfactorios son ahora los extrahoteleros seguidos de los hoteles y apartahoteles de hasta 3 estrellas.

**Composición por sexo**

```
table(viajeros_esp$SEXO, viajeros_esp$CLUSTER)
```

```
##
##              1    2
##   Hombre 1713 1250
##   Mujer  1151 1127
```

Como vemos, teniendo en cuenta los servicios específicos existe un porcentaje menor de individuos en el grupo de los mas satisfechos. Se incluyen en este grupo el 42% de

los hombres y el 49% de las mujeres. Esto podría llevarnos a pensar que los servicios mas específicos están mas descuidados.

### Composición por ocupación

```
table(viajeros_esp$OCUPACION, viajeros_esp$CLUSTER)
```

```
##
##                                1   2
##   Ama de casa                 26  32
##   Asalariado alta dirección  288 212
##   Asalariado cargo medio     795 648
##   Asalariado nivel auxiliar  177 158
##   Autónomo - profesión liberal 361 297
##   Empresario                 420 372
##   Estudiante                 258 206
##   Jubilado <U+0096> retirado       123  75
##   Otros trabajadores y obreros 349 329
##   Parado                      67  48
```

En cuanto a la ocupación, los jubilados o retirados son los menos satisfechos con las actividades y servicios específicos, mientras que los mas satisfechos vuelven a ser las amas de casa.

### Composición por familias

```
table(viajeros_esp$VALORACION_RECREO_NINYOS, viajeros_esp$CLUSTER)
```

```
##
##         1    2
##   No   79   69
##   Si 2785 2308
```

Podemos observar una clarísima diferencia con la composición de los clusters de servicios y actividades generales. La mayoría de los individuos que ha participado en actividades o servicios específicos también forma parte de una unidad familiar o tiene hijos, ya que asumimos que los individuos que evaluan una actividad o servicio es porque han accedido a el y, por tanto, los que evaluan recreo de los niños tienen hijos o similares (sobrinos, por ejemplo).

### Composición por familias/ingresos

Con la siguiente tabla podemos comprobar la estructura de renta de las familias. Como podemos observar, las familias con mas ingresos son las menos satisfechas con una clara diferencia.

```
table(viajeros_esp$VALORACION_RECREO_NINYOS, viajeros_esp$CLUSTER, viajer
os_esp$INGRESOS)
```

```
## , ,  = De 12000 a 24000
##
##        1   2
##   No  15  16
##   Si 664 590
##
## , ,  = De 24001 a 36000
##
##        1   2
##   No  19  15
##   Si 570 529
##
## , ,  = De 36001 a 48000
##
##        1   2
##   No  17   7
##   Si 431 368
##
## , ,  = De 48001 a 60000
##
##        1   2
##   No   9   9
##   Si 347 292
##
## , ,  = De 60001 a 72000
##
##        1   2
##   No   6   9
##   Si 209 173
##
## , ,  = De 72001 a 84000
##
##        1   2
##   No   4   2
##   Si 154 110
##
## , ,  = Más de 84000
##
##        1   2
##   No   9  11
##   Si 410 246
```

## Conclusión composición clusters

Hemos observado que las variables cualitativas, por lo general, no explican mucho los clusters que hemos obtenido tanto para los servicios generales como para los específicos. Si bien es cierto que señalan algunas diferencias, no lo hacen con una intensidad suficiente como para formar grupos homogéneos en función de ellas.