Clusterización de viajeros empleando el algoritmo CLARA

Álvaro de Prada

9/11/2017

En primer lugar cargamos todos los paquetes que vamos a emplear en la práctica:

```
require(cluster) # para los algoritmos PAM y CLARA
```

```
## Loading required package: cluster
```

```
require(fpc) # idem
```

```
## Loading required package: fpc
```

```
require(factoextra) # para visualizar
```

```
## Loading required package: factoextra
```

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

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

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

```
## Loading required package: NbClust
```

Emplearemos a continuación el Método Clara:

Empezamos cargando los datos:

```
viajerosv5 <- read.csv("viajeros.csv")
```

Nos deshacemos de los valores perdidos:

```
naviajeros <- na.omit(viajerosv5)
```

Y nos quedamos con las variables numéricas:

```
naviajeros <- naviajeros[,4:31]
```

Y comprobamos los distintos valores de cada variable:

```
summary(naviajeros)
```

```
##    IMPRESION     VALORACION_ALOJ  VALORACION_TRATO_ALOJ
##  Min.   :1.000   Min.   : 1.000   Min.   : 1.000
##  1st Qu.:4.000   1st Qu.: 7.000   1st Qu.: 7.000
##  Median :4.000   Median : 8.000   Median : 9.000
```

```
##  Mean   :4.262   Mean   : 7.933   Mean   : 8.191
##  3rd Qu.:5.000   3rd Qu.:10.000   3rd Qu.:10.000
##  Max.   :5.000   Max.   :10.000   Max.   :10.000
##  VALORACION_GASTRONO_ALOJ VALORACION_CLIMA VALORACION_ZONAS_BANYO
##  Min.   : 1.000           Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 6.000           1st Qu.: 8.000   1st Qu.: 7.000
##  Median : 8.000           Median : 9.000   Median : 8.000
##  Mean   : 7.476           Mean   : 8.553   Mean   : 8.102
##  3rd Qu.: 9.000           3rd Qu.:10.000   3rd Qu.:10.000
##  Max.   :10.000           Max.   :10.000   Max.   :10.000
##  VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD
##  Min.   : 1.000      Min.   : 1.000            Min.   : 1.000
##  1st Qu.: 7.000      1st Qu.: 7.000            1st Qu.: 7.000
##  Median : 9.000      Median : 8.000            Median : 8.000
##  Mean   : 8.257      Mean   : 8.105            Mean   : 8.106
##  3rd Qu.:10.000      3rd Qu.: 9.000            3rd Qu.:10.000
##  Max.   :10.000      Max.   :10.000            Max.   :10.000
##  VALORACION_LIMPIEZA VALORACION_CALIDAD_RESTAUR
##  Min.   : 1.000      Min.   : 1.000
##  1st Qu.: 7.000      1st Qu.: 7.000
##  Median : 8.000      Median : 8.000
##  Mean   : 7.991      Mean   : 7.655
##  3rd Qu.: 9.000      3rd Qu.: 9.000
##  Max.   :10.000      Max.   :10.000
##  VALORACION_OFERTA_GASTR_LOC VALORACION_TRATO_RESTAUR
##  Min.   : 1.000              Min.   : 1.000
##  1st Qu.: 6.000              1st Qu.: 7.000
##  Median : 8.000              Median : 8.000
##  Mean   : 7.396              Mean   : 8.033
##  3rd Qu.: 9.000              3rd Qu.: 9.000
##  Max.   :10.000              Max.   :10.000
##  VALORACION_PRECIO_RESTAUR VALORACION_CULTURA VALORACION_DEPORTES
##  Min.   : 1.000            Min.   : 1.000     Min.   : 1.000
##  1st Qu.: 7.000            1st Qu.: 6.000     1st Qu.: 6.000
##  Median : 8.000            Median : 7.000     Median : 8.000
##  Mean   : 7.524            Mean   : 7.136     Mean   : 7.427
##  3rd Qu.: 9.000            3rd Qu.: 8.000     3rd Qu.: 9.000
##  Max.   :10.000            Max.   :10.000     Max.   :10.000
##  VALORACION_GOLF  VALORACION_PARQUES_OCIO VALORACION_AMBIENTE_NOCTURNO
##  Min.   : 1.000   Min.   : 1.000          Min.   : 1.000
##  1st Qu.: 5.000   1st Qu.: 6.000          1st Qu.: 6.000
##  Median : 7.000   Median : 8.000          Median : 8.000
##  Mean   : 6.754   Mean   : 7.193          Mean   : 7.211
##  3rd Qu.: 9.000   3rd Qu.: 9.000          3rd Qu.: 9.000
##  Max.   :10.000   Max.   :10.000          Max.   :10.000
##  VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS 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 : 7.000           Median : 8.000
##  Mean   : 7.367         Mean   : 7.145           Mean   : 7.244
```

```
##  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_VEHIC
##  Min.   : 1.000           Min.   : 1.000           Min.   : 1.000
##  1st Qu.: 6.000           1st Qu.: 7.000           1st Qu.: 7.000
##  Median : 8.000           Median : 8.000           Median : 8.000
##  Mean   : 7.421           Mean   : 7.983           Mean   : 7.686
##  3rd Qu.: 9.000           3rd Qu.: 9.000           3rd Qu.: 9.000
##  Max.   :10.000           Max.   :10.000           Max.   :10.000
##  VALORACION_SEGURIDAD VALORACION_ESTADO_CARRETERAS
##  Min.   : 1.000       Min.   : 1.000
##  1st Qu.: 7.000       1st Qu.: 6.250
##  Median : 8.000       Median : 8.000
##  Mean   : 7.951       Mean   : 7.597
##  3rd Qu.: 9.000       3rd Qu.: 9.000
##  Max.   :10.000       Max.   :10.000
##  VALORACION_CALIDAD_COMERCIO
##  Min.   : 1.000
##  1st Qu.: 6.000
##  Median : 8.000
##  Mean   : 7.428
##  3rd Qu.: 9.000
##  Max.   :10.000
```

Necesitaremos reducir el tamaño de los datos para poder ejecutar más adelante ciertas funciones que requieren bastante poder computacional. Por ello escogemos una muestra de 500 observaciones:

```
set.seed(123)
naviajeros.mas = naviajeros[sample(1:nrow(naviajeros), 500,
replace=FALSE),]
```

Y tipificamos para poder trabajar con variables con medidas equitativas/comparables:

```
naviajeros.tip = scale(naviajeros.mas)
summary(naviajeros.tip)
```

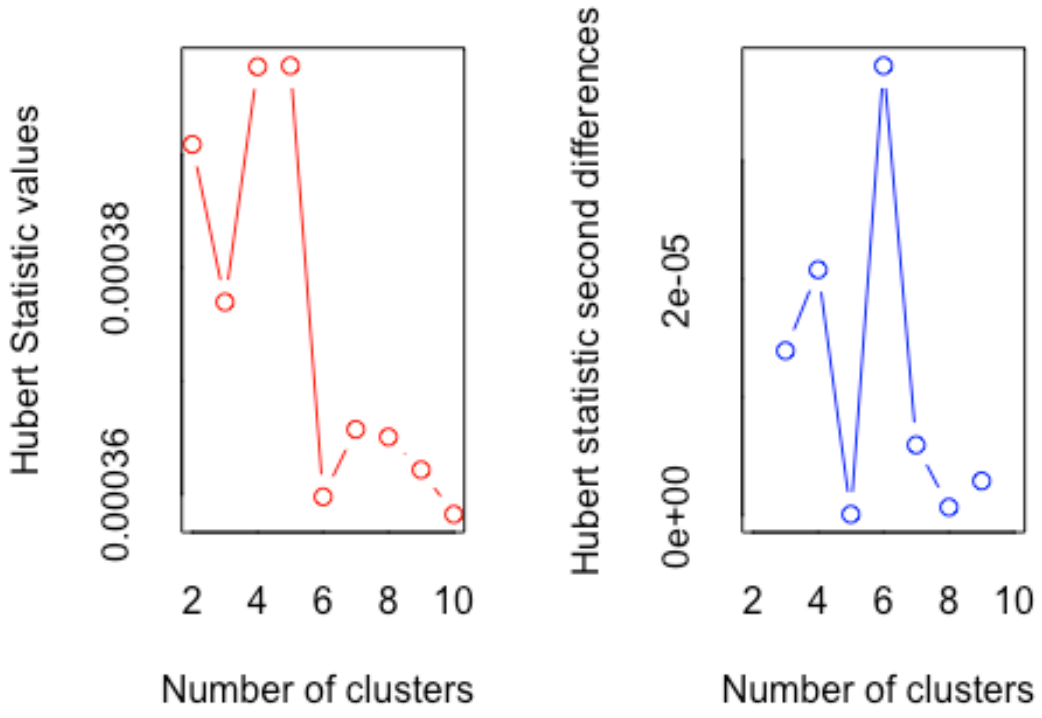
```
##    IMPRESION       VALORACION_ALOJ    VALORACION_TRATO_ALOJ
##  Min.   :-3.8157   Min.   :-3.79691   Min.   :-3.9349
##  1st Qu.:-0.3342   1st Qu.:-0.55446   1st Qu.:-0.1399
##  Median :-0.3342   Median :-0.01405   Median : 0.4023
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.8263   3rd Qu.: 1.06677   3rd Qu.: 0.9444
##  Max.   : 0.8263   Max.   : 1.06677   Max.   : 0.9444
##  VALORACION_GASTRONO_ALOJ VALORACION_CLIMA  VALORACION_ZONAS_BANYO
##  Min.   :-2.9671          Min.   :-4.7470   Min.   :-4.1666
##  1st Qu.:-0.7268          1st Qu.:-0.3540   1st Qu.:-0.7021
##  Median : 0.1694          Median : 0.2736   Median : 0.4527
##  Mean   : 0.0000          Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6174          3rd Qu.: 0.9012   3rd Qu.: 1.0301
```

```
##  Max.   : 1.0655          Max.   : 0.9012   Max.   : 1.0301
##  VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE VALORACION_TRANQUILIDAD
##  Min.   :-3.9327     Min.   :-4.1109           Min.   :-4.0607
##  1st Qu.:-0.6736     1st Qu.:-0.6737           1st Qu.:-0.6871
##  Median : 0.4128     Median :-0.1008           Median : 0.4374
##  Mean   : 0.0000     Mean   : 0.0000           Mean   : 0.0000
##  3rd Qu.: 0.9560     3rd Qu.: 1.0449           3rd Qu.: 0.9997
##  Max.   : 0.9560     Max.   : 1.0449           Max.   : 0.9997
##  VALORACION_LIMPIEZA VALORACION_CALIDAD_RESTAUR
##  Min.   :-3.72545    Min.   :-3.8276
##  1st Qu.:-0.56024    1st Qu.:-0.3855
##  Median :-0.03271    Median : 0.1882
##  Mean   : 0.00000    Mean   : 0.0000
##  3rd Qu.: 0.49483    3rd Qu.: 0.7618
##  Max.   : 1.02236    Max.   : 1.3355
##  VALORACION_OFERTA_GASTR_LOC VALORACION_TRATO_RESTAUR
##  Min.   :-3.1278             Min.   :-3.92255
##  1st Qu.:-0.6611             1st Qu.:-0.54007
##  Median : 0.3256             Median : 0.02368
##  Mean   : 0.0000             Mean   : 0.00000
##  3rd Qu.: 0.8190             3rd Qu.: 0.58742
##  Max.   : 1.3123             Max.   : 1.15117
##  VALORACION_PRECIO_RESTAUR VALORACION_CULTURA VALORACION_DEPORTES
##  Min.   :-3.4197           Min.   :-2.96744   Min.   :-3.3471
##  1st Qu.:-0.2957           1st Qu.:-0.56193   1st Qu.:-0.3807
##  Median : 0.2249           Median :-0.08083   Median : 0.2641
##  Mean   : 0.0000           Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.7456           3rd Qu.: 0.88138   3rd Qu.: 0.7800
##  Max.   : 1.2663           Max.   : 1.36248   Max.   : 1.2959
##  VALORACION_GOLF    VALORACION_PARQUES_OCIO
VALORACION_AMBIENTE_NOCTURNO
##  Min.   :-2.34240   Min.   :-2.6872         Min.   :-2.9529
##  1st Qu.:-0.71743   1st Qu.:-0.4968         1st Qu.:-0.5876
##  Median : 0.09506   Median : 0.3794         Median : 0.3586
##  Mean   : 0.00000   Mean   : 0.0000         Mean   : 0.0000
##  3rd Qu.: 0.50130   3rd Qu.: 0.8175         3rd Qu.: 0.8317
##  Max.   : 1.31379   Max.   : 1.2556         Max.   : 1.3047
##  VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS VALORACION_SALUD
##  Min.   :-2.8792        Min.   :-2.63472         Min.   :-2.8860
##  1st Qu.:-0.5992        1st Qu.:-0.46444         1st Qu.:-0.5526
##  Median : 0.3128        Median :-0.03038         Median : 0.3808
##  Mean   : 0.0000        Mean   : 0.00000         Mean   : 0.0000
##  3rd Qu.: 0.7688        3rd Qu.: 0.83773         3rd Qu.: 0.8475
##  Max.   : 1.2248        Max.   : 1.27179         Max.   : 1.3142
##  VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI
VALORACION_ALQ_VEHIC
##  Min.   :-3.2260          Min.   :-3.70722          Min.   :-3.4051
##  1st Qu.:-0.2817          1st Qu.:-0.54496          1st Qu.:-0.3873
##  Median : 0.2090          Median :-0.01792          Median : 0.1157
##  Mean   : 0.0000          Mean   : 0.00000          Mean   : 0.0000
```

```
##  3rd Qu.: 0.6998          3rd Qu.: 1.03617          3rd Qu.: 0.6186
##  Max.   : 1.1905          Max.   : 1.03617          Max.   : 1.1216
##  VALORACION_SEGURIDAD VALORACION_ESTADO_CARRETERAS
##  Min.   :-3.90791     Min.   :-3.2308
##  1st Qu.:-0.54964     1st Qu.:-0.3298
##  Median : 0.01008     Median : 0.1538
##  Mean   : 0.00000     Mean   : 0.0000
##  3rd Qu.: 0.56979     3rd Qu.: 0.6373
##  Max.   : 1.12950     Max.   : 1.1208
##  VALORACION_CALIDAD_COMERCIO
##  Min.   :-3.2909
##  1st Qu.:-0.7734
##  Median : 0.2336
##  Mean   : 0.0000
##  3rd Qu.: 0.7371
##  Max.   : 1.2406
```

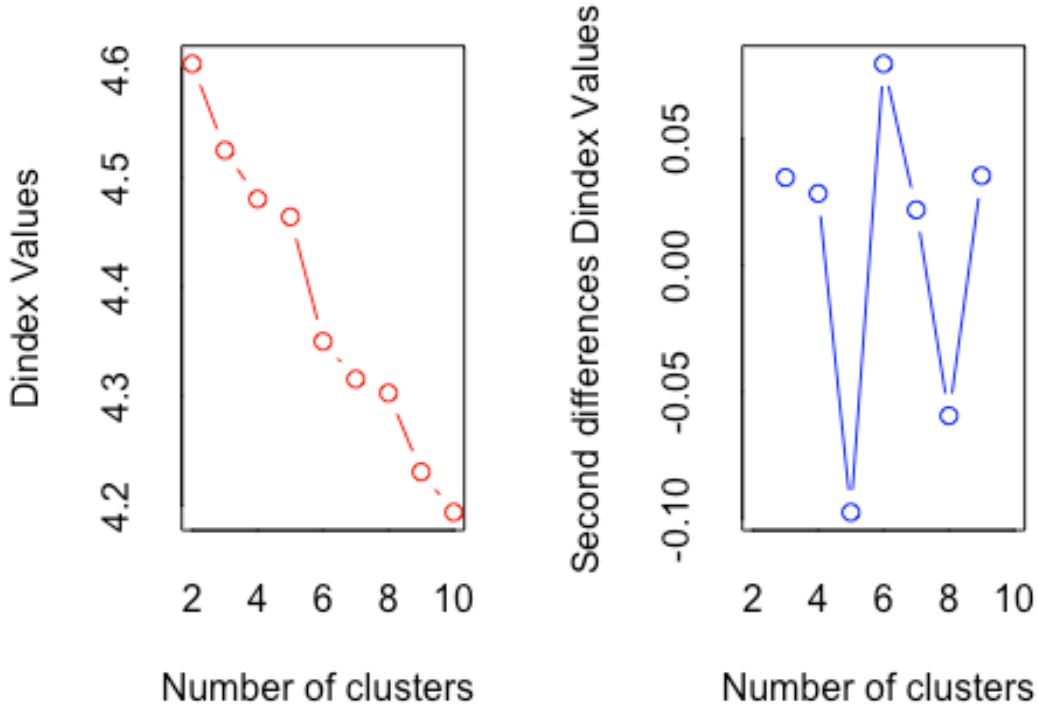
Ejecutamos la función NbClus:

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




```
## *** : The Hubert index is a graphical method of determining the number
of clusters.
```

```
##                 In the plot of Hubert index, we seek a significant
knee that corresponds to a
##                 significant increase of the value of the measure i.e
the significant peak in Hubert
##                 index second differences plot.
##
```

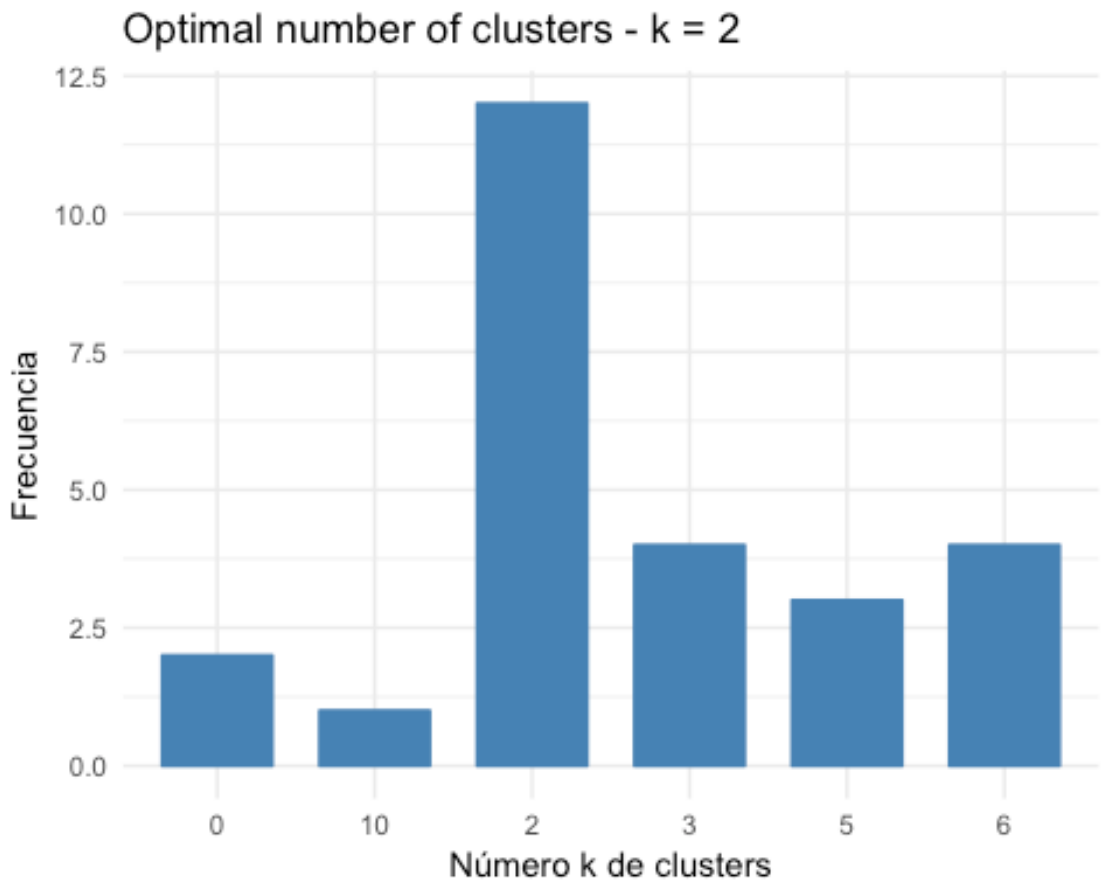



```
## *** : The D index is a graphical method of determining the number of
clusters.
##                 In the plot of D index, we seek a significant knee
(the significant peak in Dindex
##                 second differences plot) that corresponds to a
significant increase of the value of
##                 the measure.
##
## *****************************************************************
## * Among all indices:
## * 12 proposed 2 as the best number of clusters
## * 4 proposed 3 as the best number of clusters
## * 3 proposed 5 as the best number of clusters
## * 4 proposed 6 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
##
##
## *******************************************************************
```

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

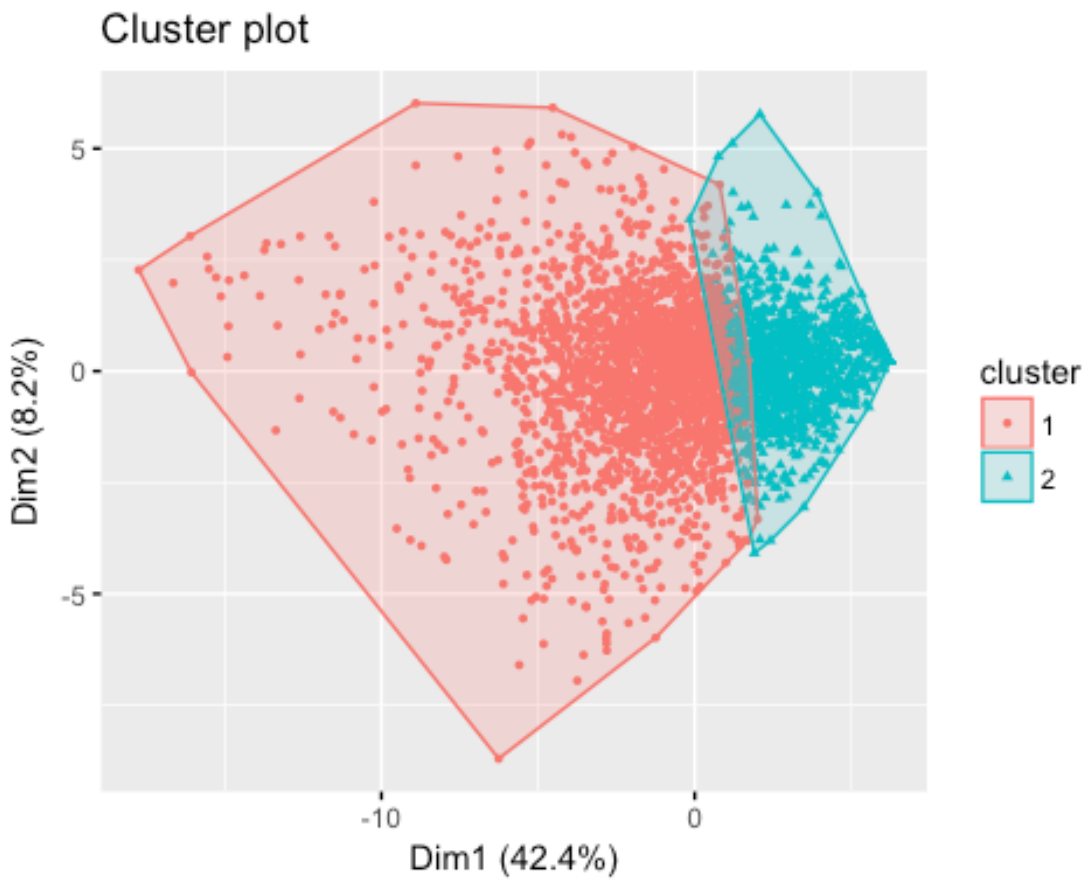
```
## Among all indices:
## ===================
## * 2 proposed  0 as the best number of clusters
## * 12 proposed  2 as the best number of clusters
## * 4 proposed  3 as the best number of clusters
## * 3 proposed  5 as the best number of clusters
## * 4 proposed  6 as the best number of clusters
##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



Por lo tanto la opción mayoritaria señala 2 grupos de CLUSTERS, por lo que procedemos a representar gráficamente estos 2 clusters:

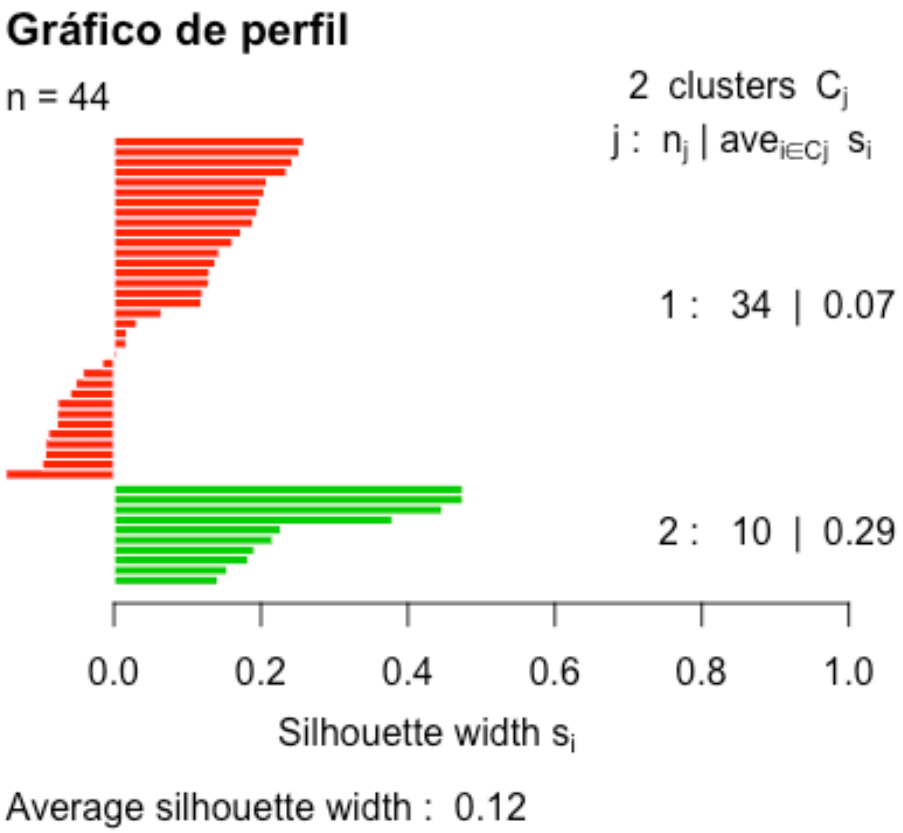
```
require(cluster)
viajeros.clara=clara(naviajeros, 2, samples=200)

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



Podemos comprobar como existe una clara segementación de 2 grupos bastante heterogéneos, a pesar de que existeuna pequeña zona de convergencia entre los grupos.

```
plot(silhouette(viajeros.clara),col = 2:3, main = "Gráfico de perfil")
```



Obtenemos los medioides:

```
viajeros.clara$medoids
```

```
##       IMPRESION VALORACION_ALOJ VALORACION_TRATO_ALOJ
## 11910         4               7                     7
## 10992         5               9                     9
##       VALORACION_GASTRONO_ALOJ VALORACION_CLIMA VALORACION_ZONAS_BANYO
## 11910                        7                8                      8
## 10992                        9                9                      9
##       VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE
## 11910                   8                        8
## 10992                   9                        9
##       VALORACION_TRANQUILIDAD VALORACION_LIMPIEZA
## 11910                       8                   8
## 10992                       9                   9
##       VALORACION_CALIDAD_RESTAUR VALORACION_OFERTA_GASTR_LOC
## 11910                          7                           7
## 10992                          9                           9
##       VALORACION_TRATO_RESTAUR VALORACION_PRECIO_RESTAUR
## 11910                        7                         7
## 10992                        9                         9
##       VALORACION_CULTURA VALORACION_DEPORTES VALORACION_GOLF
## 11910                  7                   7               7
```

```
## 10992                  9                   9               9
##       VALORACION_PARQUES_OCIO VALORACION_AMBIENTE_NOCTURNO
## 11910                       7                           7
## 10992                       9                           9
##       VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS VALORACION_SALUD
## 11910                      7                       7                7
## 10992                      9                       9                9
##       VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI
## 11910                        7                         7
## 10992                        9                         9
##       VALORACION_ALQ_VEHIC VALORACION_SEGURIDAD
## 11910                    7                    7
## 10992                    9                    9
##       VALORACION_ESTADO_CARRETERAS VALORACION_CALIDAD_COMERCIO
## 11910                            7                          7
## 10992                            9                          9
```

Tras haber clusterizado a los viajeros, creamos una tabla en la que se indique en qué cluster se ha introducido a cada viajero para poder ver si tienen características comunes como por ejemplo ingresos, edad...:

```
clusters<-viajeros.clara$clustering
clusters<-as.data.frame(clusters)
resultado<-na.omit(viajerosv5)
resultado<-data.frame(clusters,resultado)

head(resultado)
```

```
##    clusters      X PAIS_RESID_AGRUP
ALOJ_CATEG_1
## 7         1 242037      Reino Unido Hoteles - apartahoteles de 4
estrellas
## 11        1 161764      Reino Unido
Extrahoteleros
## 12        1 228332            Otros Hoteles - apartahoteles de 4
estrellas
## 23        1 146449        Espa\xf1a Hoteles - apartahoteles de 4
estrellas
## 28        1 219486      Reino Unido
Extrahoteleros
## 35        2 254647        Espa\xf1a Hoteles - apartahoteles de 4
estrellas
##    IMPRESION VALORACION_ALOJ VALORACION_TRATO_ALOJ
## 7          1               7                    7
## 11         5              10                    7
## 12         4               7                    9
## 23         4               9                    9
## 28         4               8                    7
## 35         4              10                   10
##    VALORACION_GASTRONO_ALOJ VALORACION_CLIMA VALORACION_ZONAS_BANYO
## 7                         1                8                      8
```

```
## 11                       10                 9                       10
## 12                        6                10                        8
## 23                        9                10                        9
## 28                        9                10                        9
## 35                        9                10                        9
##    VALORACION_PAISAJES VALORACION_MEDIO_AMBIENTE
VALORACION_TRANQUILIDAD
## 7                   10                       10
9
## 11                  10                       10
10
## 12                   9                        8
6
## 23                  10                        9
8
## 28                   9                        9
5
## 35                   8                        9
10
##    VALORACION_LIMPIEZA VALORACION_CALIDAD_RESTAUR
## 7                    7                         10
## 11                   5                         10
## 12                   6                          8
## 23                   9                          6
## 28                   9                         10
## 35                  10                          9
##    VALORACION_OFERTA_GASTR_LOC VALORACION_TRATO_RESTAUR
## 7                            7                       10
## 11                          10                       10
## 12                           8                        9
## 23                           8                        4
## 28                           8                        9
## 35                           8                        9
##    VALORACION_PRECIO_RESTAUR VALORACION_CULTURA VALORACION_DEPORTES
## 7                          8                  7                    7
## 11                        10                  1                    1
## 12                        10                  1                   10
## 23                         5                  7                    6
## 28                         9                  5                    5
## 35                         9                  8                    8
##    VALORACION_GOLF VALORACION_PARQUES_OCIO
VALORACION_AMBIENTE_NOCTURNO
## 7                3                      10
10
## 11               1                      10
10
## 12              10                       8
8
## 23               7                       6
8
```

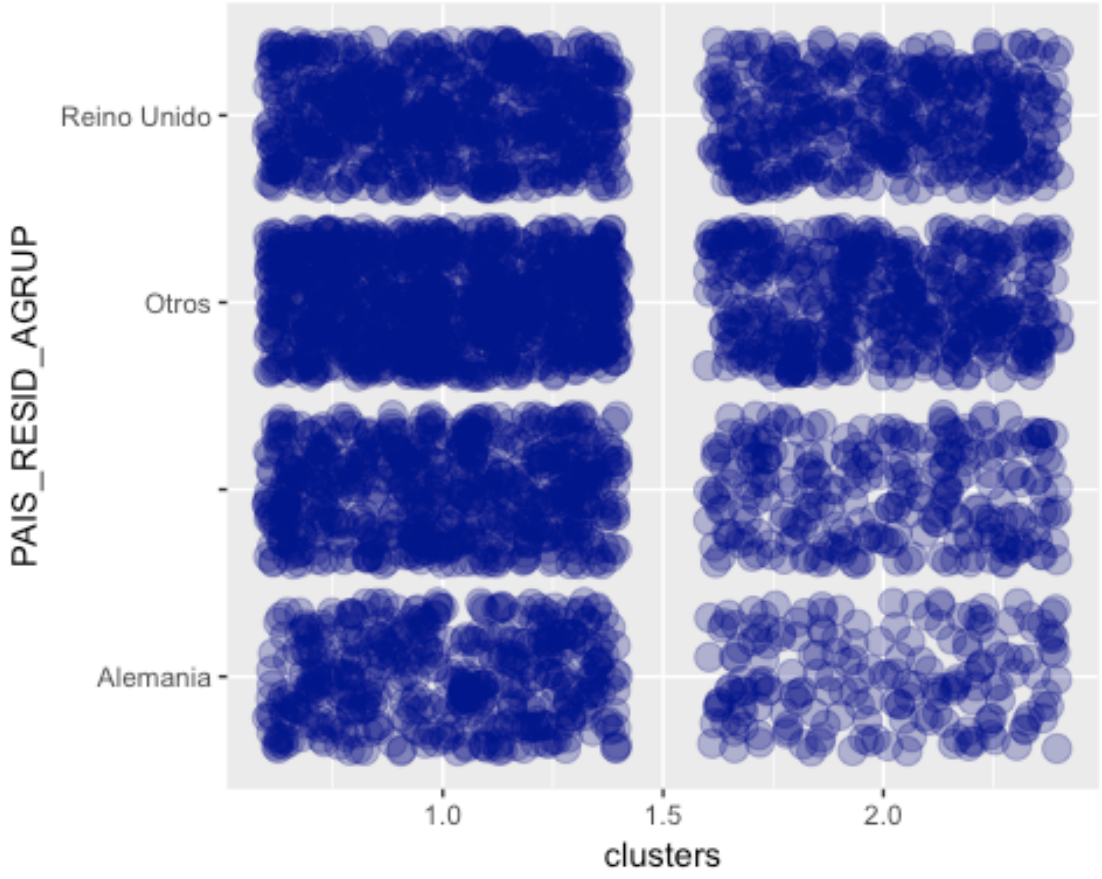
```
## 28              10                       5
8
## 35               9                       9
8
##    VALORACION_EXCURSIONES VALORACION_RECREO_NINYOS VALORACION_SALUD
## 7                       8                        8                8
## 11                     10                        1                1
## 12                      6                        9                4
## 23                      7                        6                7
## 28                      5                        9                5
## 35                      9                        9                9
##    VALORACION_SERVICIOS_BUS VALORACION_SERVICIOS_TAXI
VALORACION_ALQ_VEHIC
## 7                         3                        5
5
## 11                       10                       10
10
## 12                        8                        9
6
## 23                        8                        7
6
## 28                        8                        8
10
## 35                       10                        9
9
##    VALORACION_SEGURIDAD VALORACION_ESTADO_CARRETERAS
## 7                     7                           5
## 11                   10                          10
## 12                    7                           4
## 23                    8                           5
## 28                    8                          10
## 35                    9                           7
##    VALORACION_CALIDAD_COMERCIO VALORACION_HOSPITALIDAD   SEXO EDAD
## 7                            7                       7 Hombre   28
## 11                          10                      10 Hombre   25
## 12                           8                      10 Hombre   38
## 23                           8                       9 Hombre   25
## 28                          10                      10  Mujer   18
## 35                           9                       9 Hombre   30
##                            OCUPACION         INGRESOS
## 7  Aut\xf3nomo - profesi\xf3n liberal  M\xe1s de 84000
## 11       Asalariado alta direcci\xf3n De 60001 a 72000
## 12                          Estudiante  M\xe1s de 84000
## 23       Otros trabajadores y obreros De 12000 a 24000
## 28       Otros trabajadores y obreros De 12000 a 24000
## 35 Aut\xf3nomo - profesi\xf3n liberal De 24001 a 36000
```

A continuaciñon representaremos gráficamente las distribuciones de las variables nominales en cada cluster, para poder estimar el tipo de viajero (pais de origen, ingresos, edad...) del que se compone cada cluster:

```
table(resultado$clusters, resultado$PAIS_RESID_AGRUP)

##
##     Alemania Espa\xf1a Otros Reino Unido
##   1      420       567   823         635
##   2      189       264   477         439

ggplot(aes(x = clusters, y = PAIS_RESID_AGRUP), data = resultado, colour
= 'red') +
   geom_point(size = 4, alpha=0.3 , position="jitter",  colour = 'dark
blue')
```



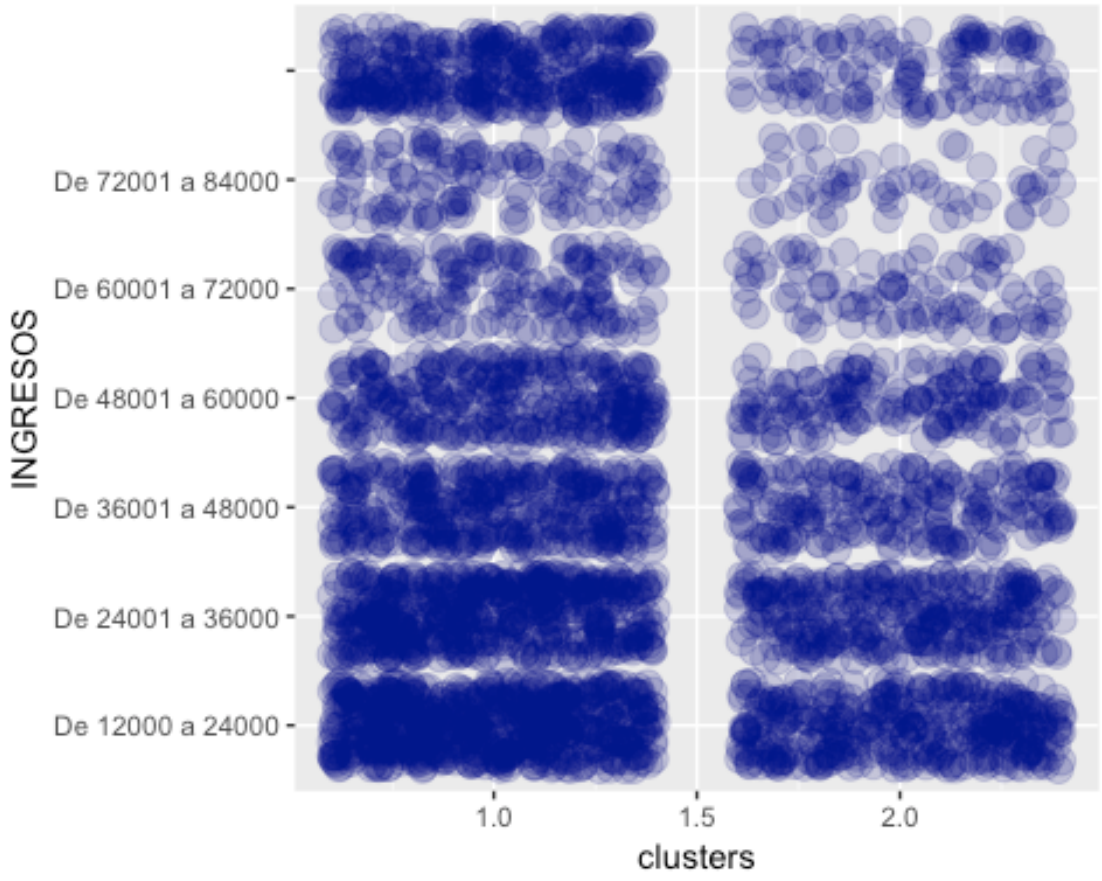
En este gráfico podemos ver que en los datos obtenidos, la procedencia mayoritaria en ambos clusters son de UK y Otros, pero no podemos confirmar a simple vista que exista una diferenciación de clusters en función del pais de procedencia.

```
table(resultado$clusters, resultado$INGRESOS)

##
##     De 12000 a 24000 De 24001 a 36000 De 36001 a 48000 De 48001 a
60000
##   1              597              500              360
311
##   2              371              307              221
177
```

```
##
##     De 60001 a 72000 De 72001 a 84000 M\xe1s de 84000
##   1              176              131             370
##   2              100               60             133
```

```
ggplot(aes(x = clusters, y = INGRESOS), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```

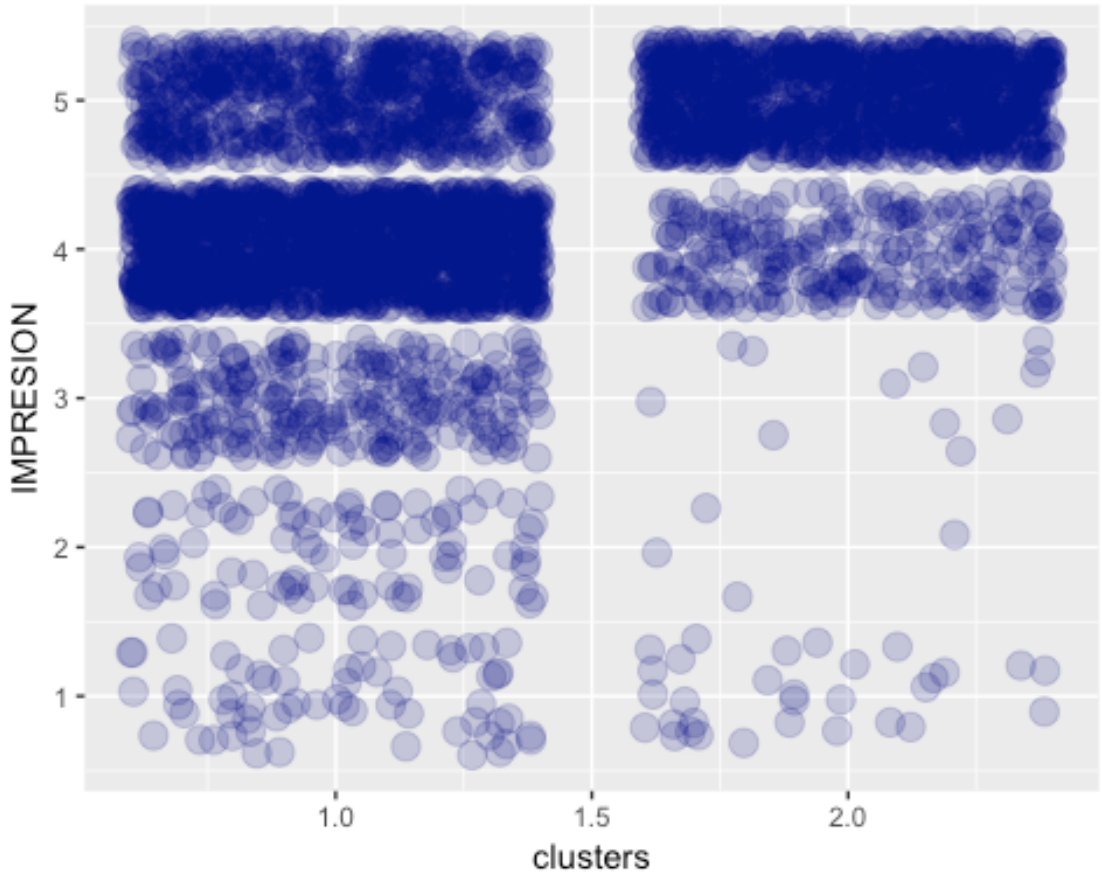


En el caso de los ingresos sucede algo similar. La mayor parte de los viajeros se encuentra en rangos salariales entre 12.000-36.000 y 72.001-84.000, pero no se aprecia segmentación entre clusters.

```
table(resultado$clusters, resultado$IMPRESION)
```

```
##
##        1    2    3    4    5
##   1   63   82  249 1395  656
##   2   31    4   12  263 1059
```

```
ggplot(aes(x = clusters, y = IMPRESION), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



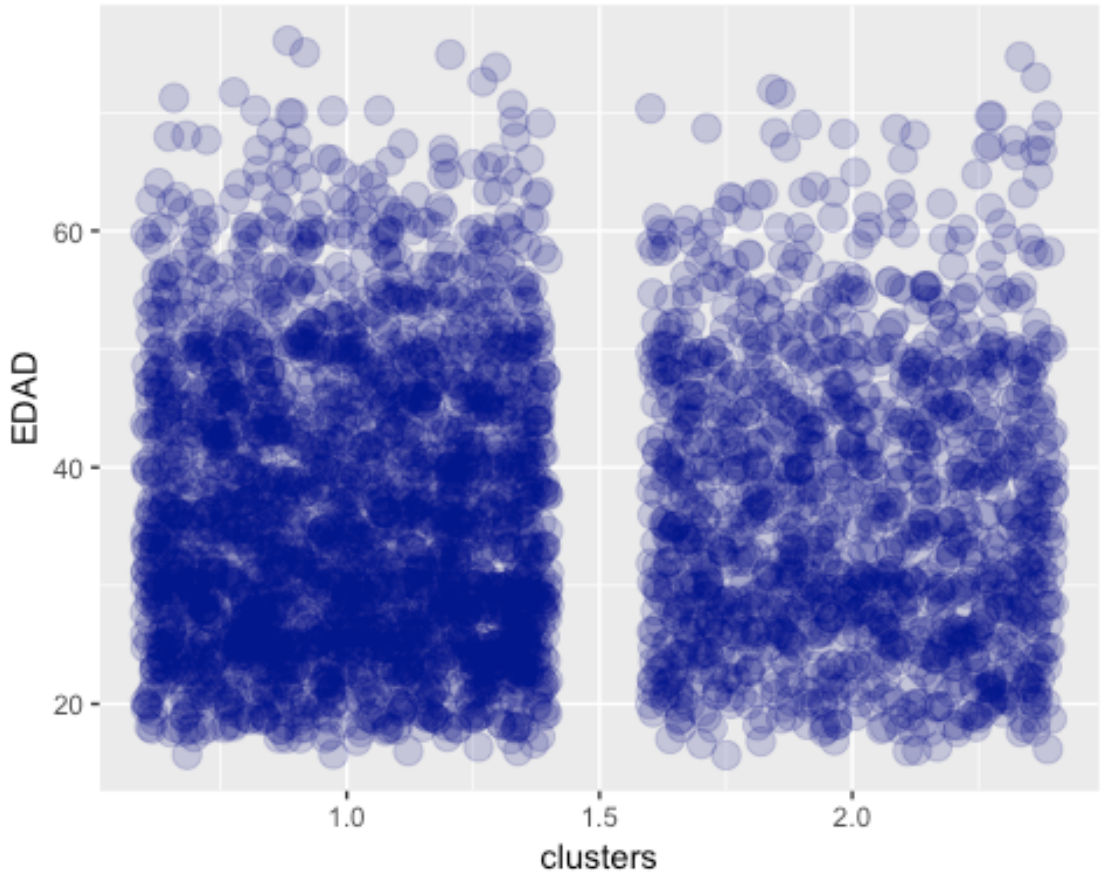
En la categoría de impresión, si hemos comprobado que existe una segmentación entre clusters, en la que podemos ver que en el cluster 1 hay una tendencia a dar una impresión de nota 4 y en el cluster dos, la nota que predomina es el 5.

```
table(resultado$EDAD, resultado$clusters)
```

```
##
##        1   2
##   16   5   5
##   17  11   4
##   18  37  22
##   19  51  23
##   20  45  36
##   21  51  35
##   22  50  36
##   23  86  31
##   24  78  38
##   25 116  40
##   26  74  52
##   27  72  40
##   28  87  51
##   29  79  42
##   30 104  57
##   31  71  43
```

```
##    32  58  36
##    33  70  31
##    34  55  37
##    35  73  39
##    36  68  34
##    37  66  31
##    38  52  32
##    39  57  31
##    40  64  50
##    41  46  22
##    42  50  43
##    43  55  24
##    44  52  32
##    45  53  40
##    46  66  28
##    47  45  22
##    48  45  34
##    49  40  32
##    50  69  39
##    51  38  15
##    52  29  20
##    53  29  10
##    54  29  15
##    55  31  22
##    56  21  12
##    57  20   5
##    58  14  12
##    59  18   9
##    60  30  10
##    61  13   7
##    62  13   5
##    63  10   8
##    64   8   1
##    65   8   3
##    66   8   2
##    67   4   5
##    68   6   5
##    69   2   3
##    70   5   4
##    71   2   0
##    72   1   2
##    73   1   1
##    74   1   0
##    75   2   1
##    76   1   0
```

```
ggplot(aes(x = clusters, y = EDAD), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```

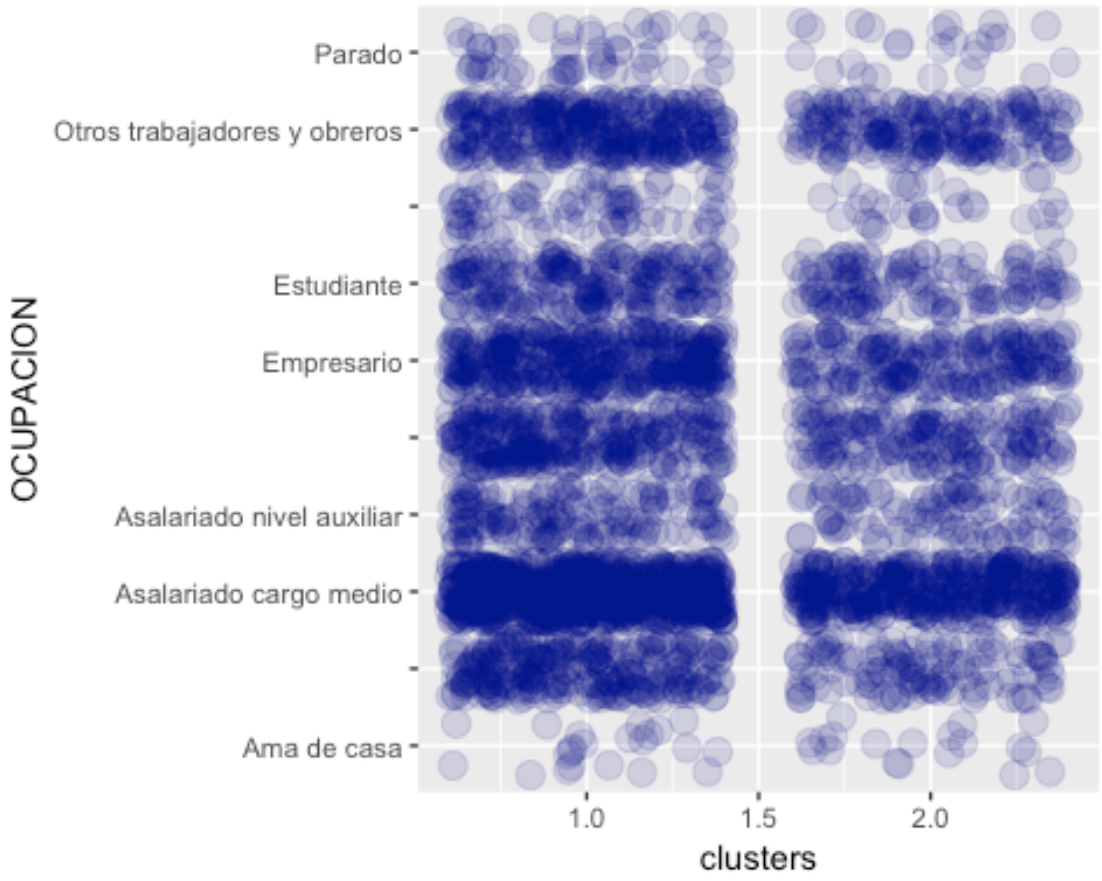


En el caso de la edad sucede algo similar que con el sueldo y la procedencia, no se puede definir una clara segmentación.

```
table(resultado$clusters, resultado$OCUPACION)
```

```
##
##     Ama de casa Asalariado alta direcci\xf3n Asalariado cargo medio
##   1          22                          259                    701
##   2          19                          117                    378
##
##     Asalariado nivel auxiliar Aut\xf3nomo - profesi\xf3n liberal
##   1                       158                                 280
##   2                       101                                 174
##
##     Empresario Estudiante Jubilado \x96 retirado
##   1        365        214                     81
##   2        202        128                     30
##
##     Otros trabajadores y obreros Parado
##   1                          311     54
##   2                          195     25
```

```
ggplot(aes(x = clusters, y = OCUPACION), data = resultado) +
   geom_point(size = 4, alpha=0.15 , position="jitter",  colour = 'dark
blue')
```



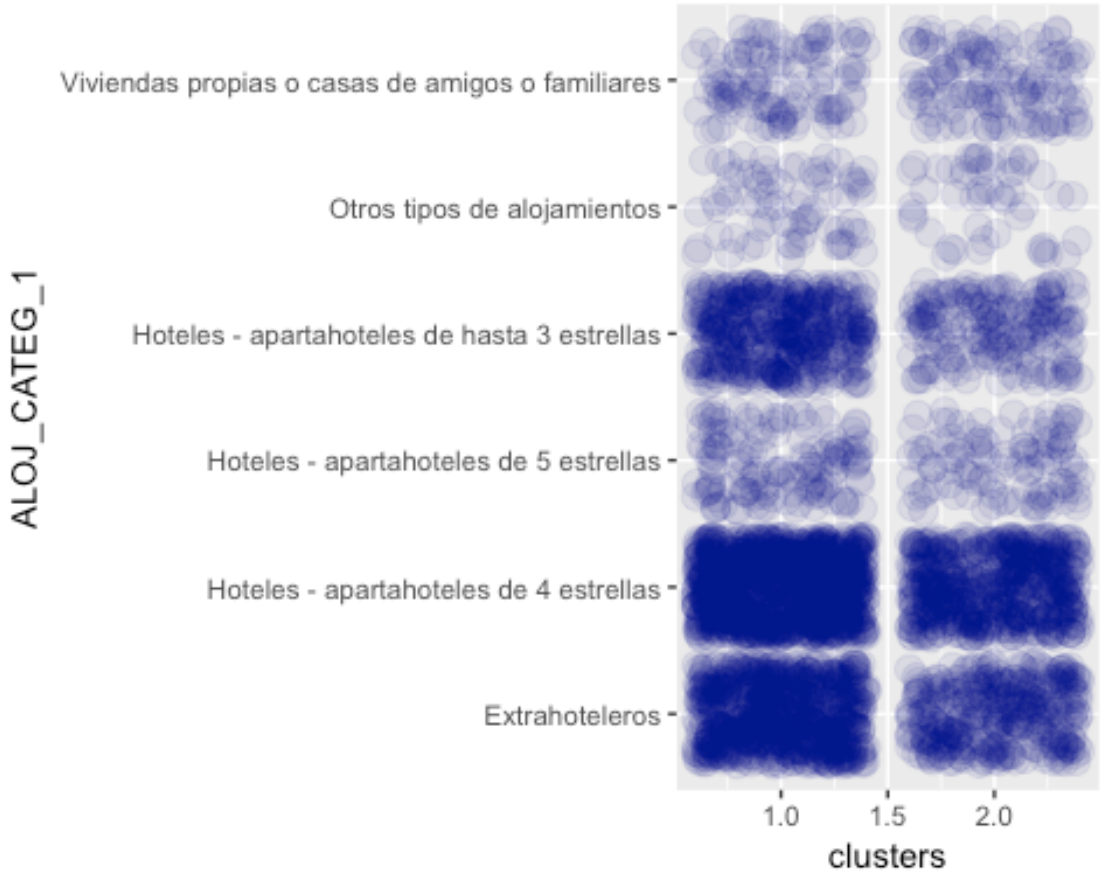
En el caso de la ocupación sucede algo similar que con el sueldo y la procedencia, no se puede definir una clara segmentación.

```
table(resultado$clusters, resultado$ALOJ_CATEG_1)
```

```
##
##     Extrahoteleros Hoteles - apartahoteles de 4 estrellas
##   1            692                                    985
##   2            344                                    557
##
##     Hoteles - apartahoteles de 5 estrellas
##   1                                    138
##   2                                    106
##
##     Hoteles - apartahoteles de hasta 3 estrellas
##   1                                          448
##   2                                          187
##
##     Otros tipos de alojamientos
##   1                          71
```

```
##   2                           49
##
##     Viviendas propias o casas de amigos o familiares
##   1                                                111
##   2                                                126
```

```
ggplot(aes(x = clusters, y = ALOJ_CATEG_1), data = resultado) +
   geom_point(size = 4, alpha=0.09 , position="jitter",  colour = 'dark
blue')
```

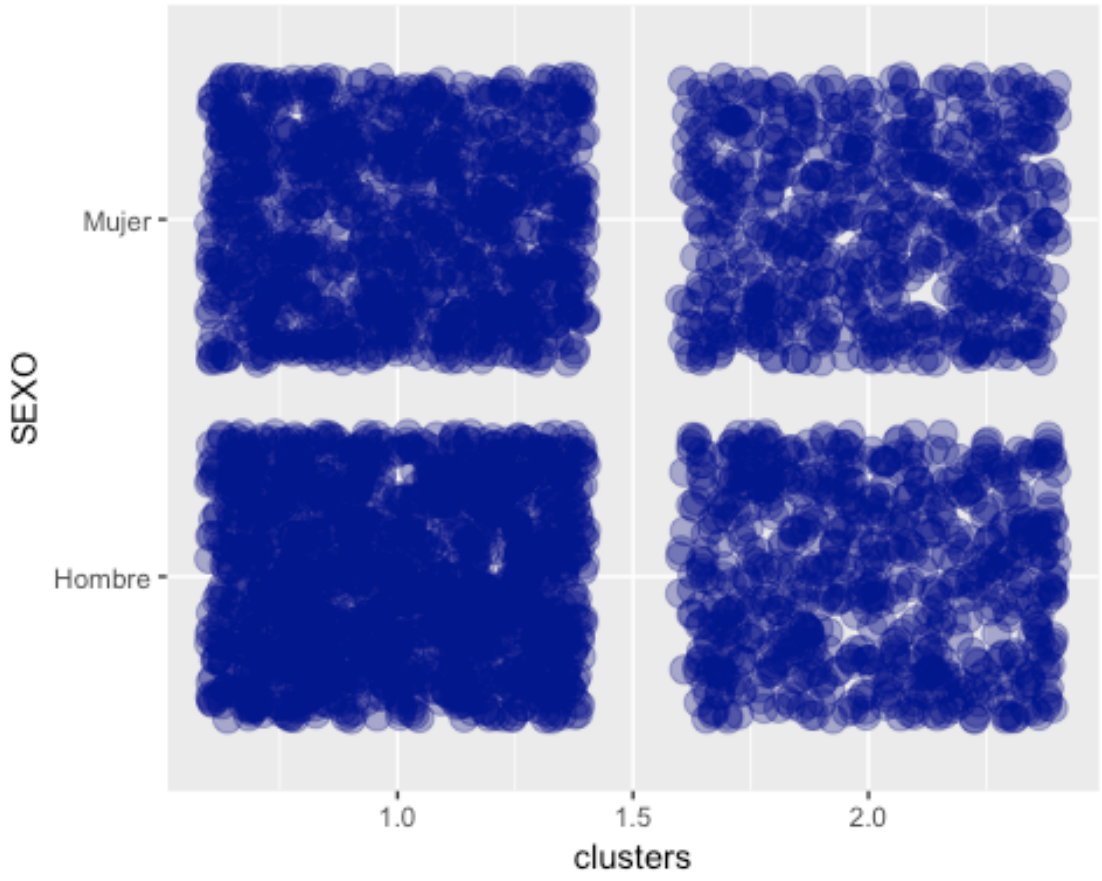


En el caso de la categoría de alojamiento podemos decir que

```
table(resultado$clusters, resultado$SEXO)
```

```
##
##     Hombre Mujer
##   1   1402  1043
##   2    701   668
```

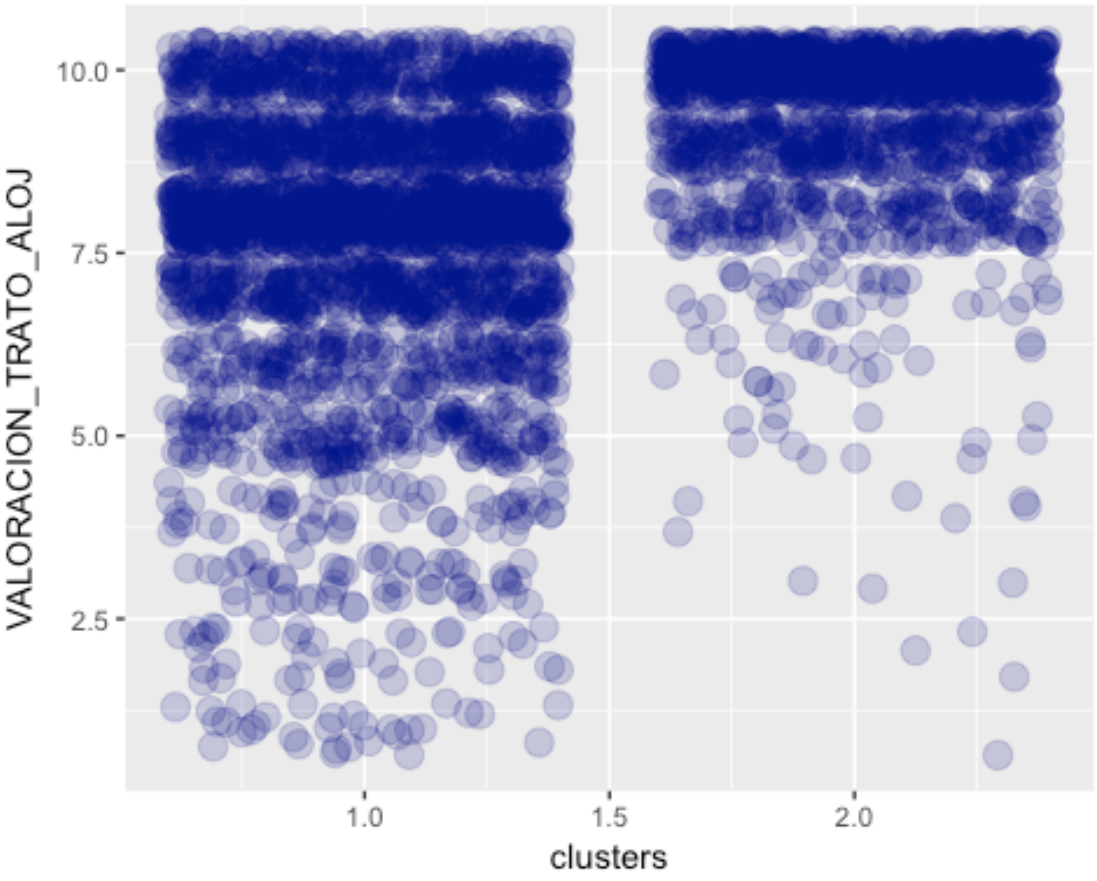
```
ggplot(aes(x = clusters, y = SEXO), data = resultado) +
   geom_point(size = 4, alpha=0.35 , position="jitter",  colour = 'dark
blue')
```



En el caso del sexo, podríamos decir que en el cluster 1 hay unmayor numero de hombres que en el cluster 2.

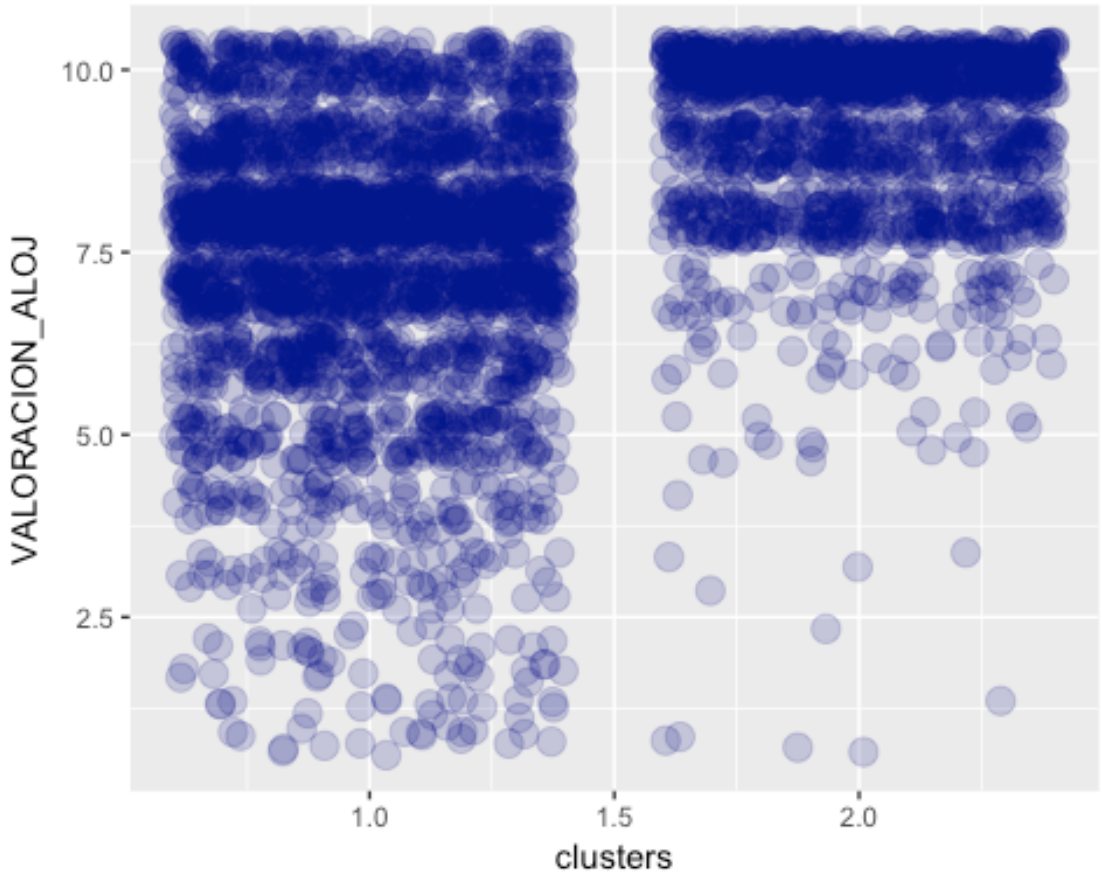
Por último vamos a comprobar si existe una segmentación entre VALORACIONES en cada cluster:

```
ggplot(aes(x = clusters, y = VALORACION_TRATO_ALOJ), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



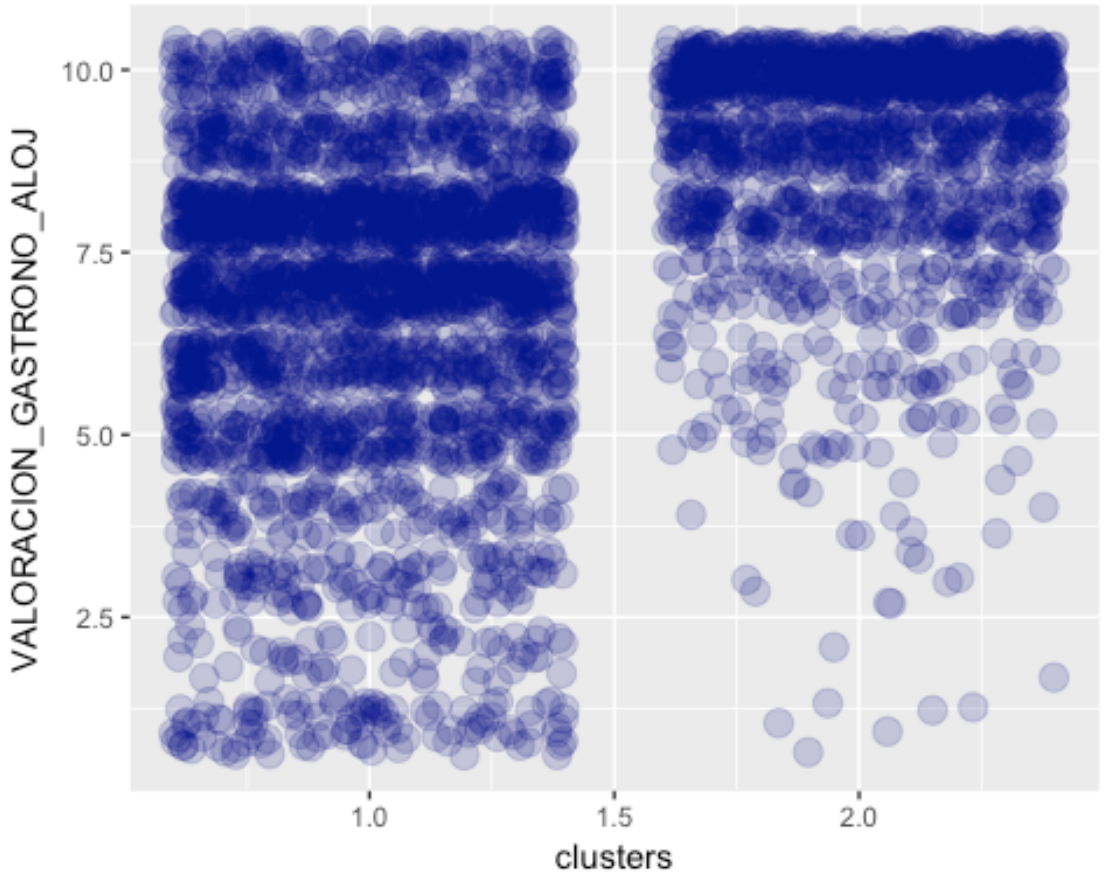
En el CLUSTER 2 valoran más el trato del alojamiento que en el CLUSTER 1.

```
ggplot(aes(x = clusters, y = VALORACION_ALOJ), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



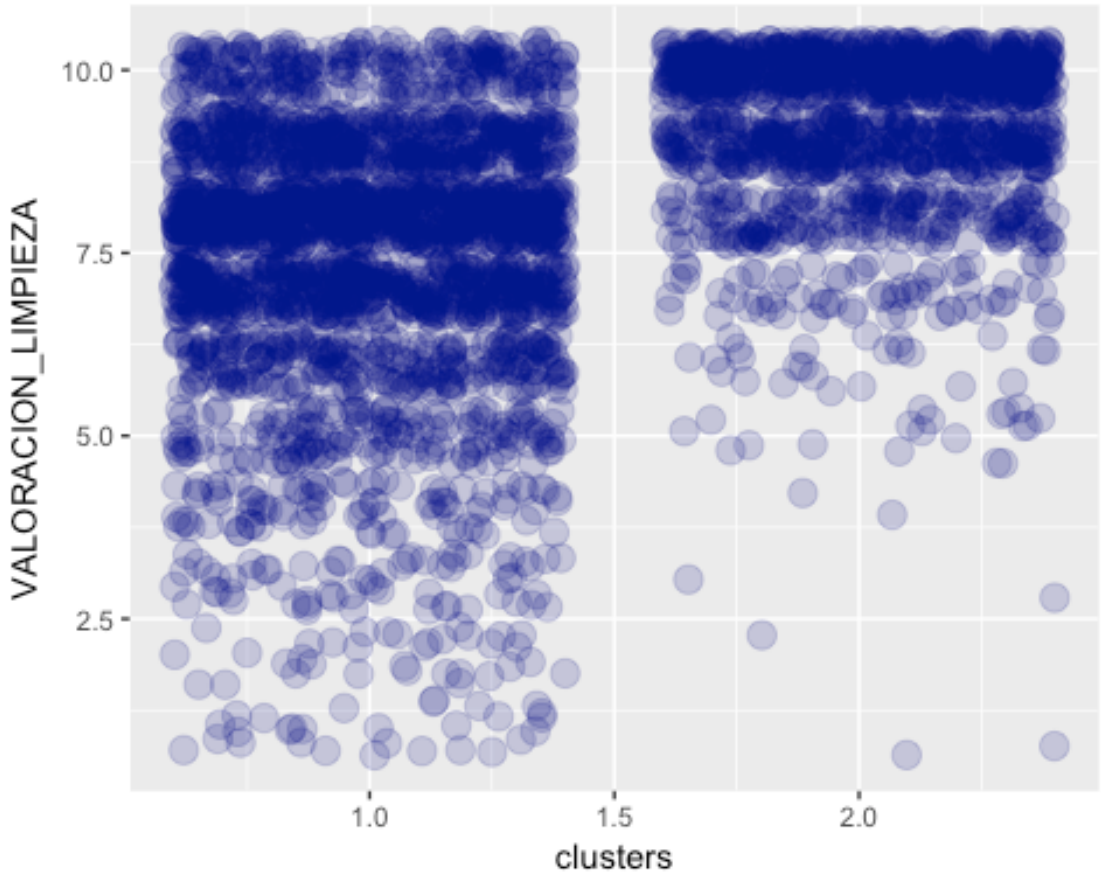
En el CLUSTER 1 valoran más el Alojamiento que en el CLUSTER 2.

```
ggplot(aes(x = clusters, y = VALORACION_GASTRONO_ALOJ), data = resultado)
+
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



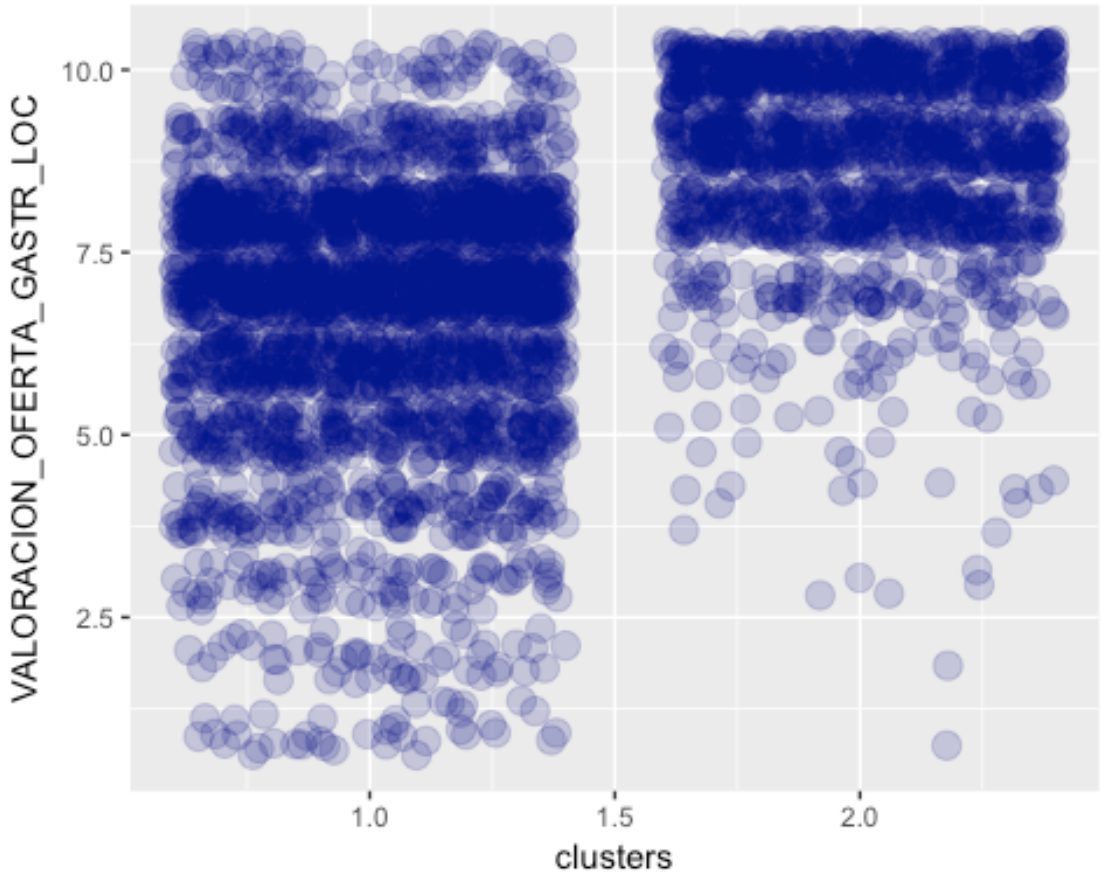
En el CLUSTER 2 valoran más la GASTRONOMÍA DEL ALOJAMIENTO que en el CLUSTER 1.

```
ggplot(aes(x = clusters, y = VALORACION_LIMPIEZA), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



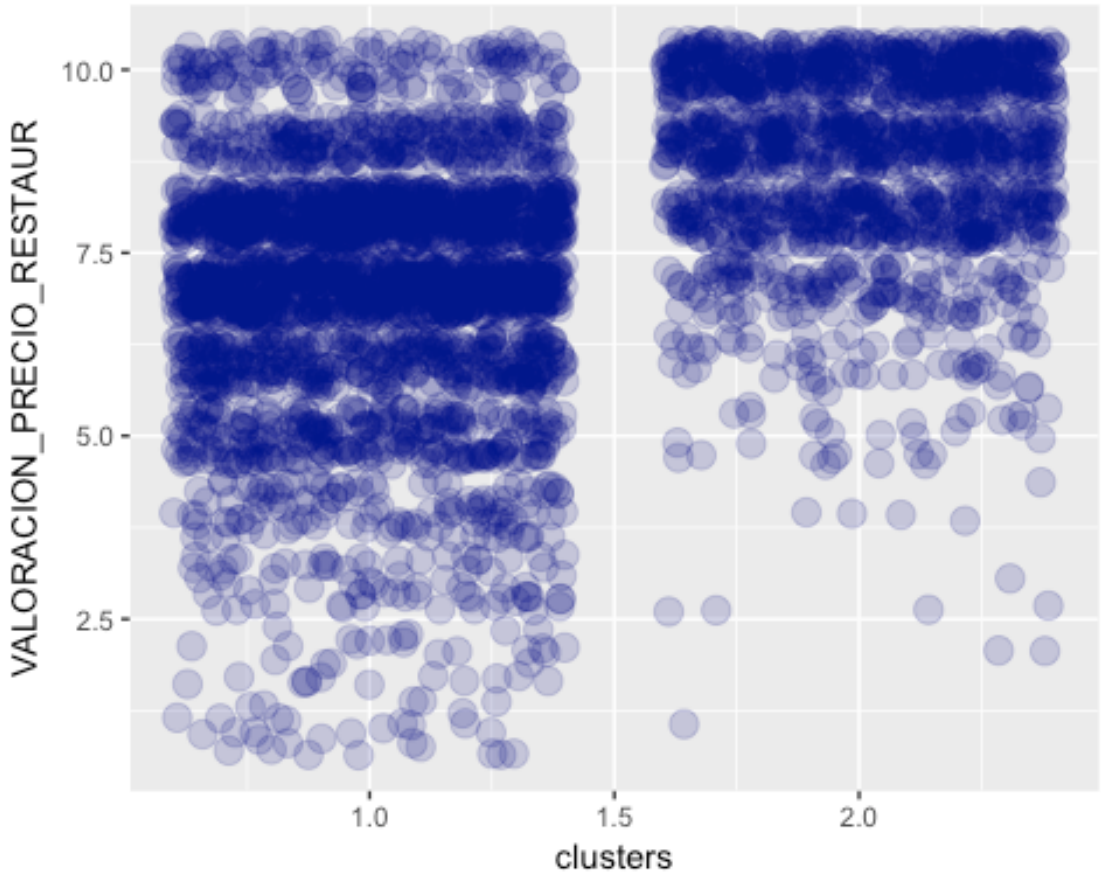
En el CLUSTER 1 valoran más LA LIMPIEZA que en el CLUSTER 2.

```
ggplot(aes(x = clusters, y = VALORACION_OFERTA_GASTR_LOC), data =
resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



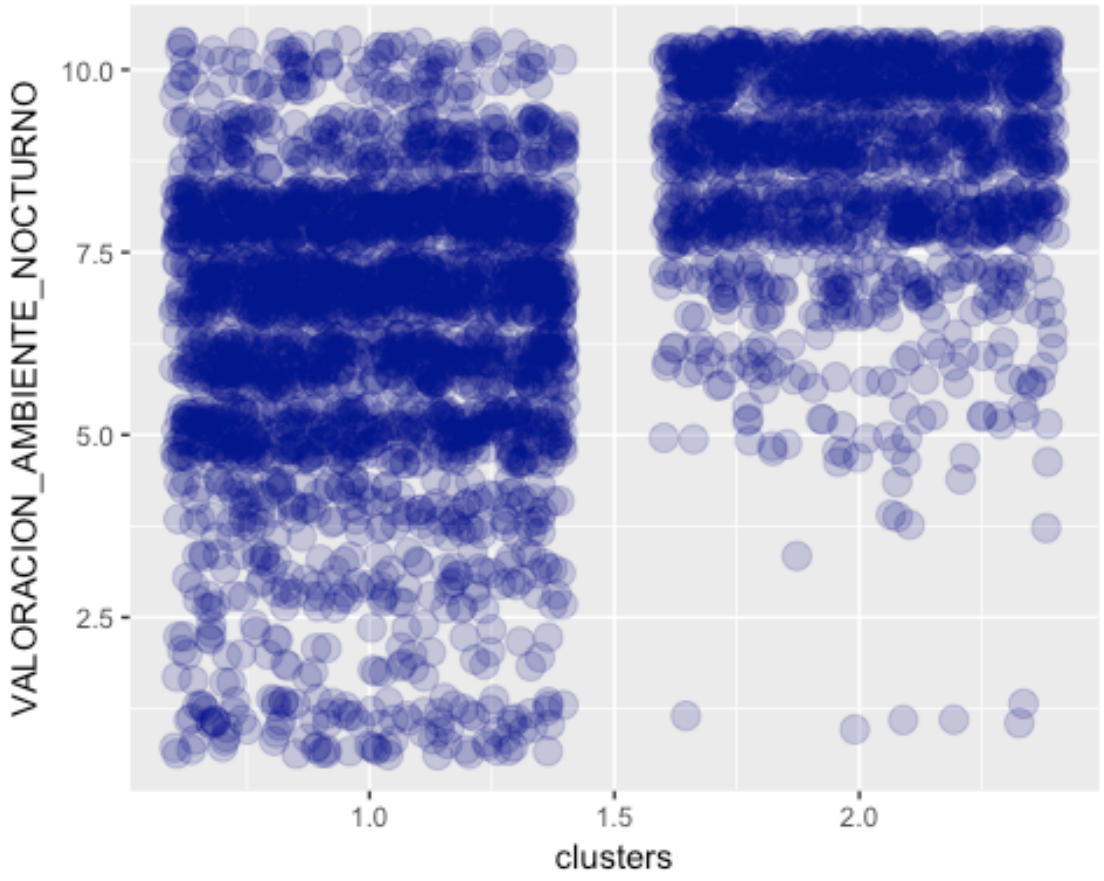
En el CLUSTER 2 valoran más LA OFERTA GASTRONÓMICA LOCAL que en el CLUSTER 1.

```
ggplot(aes(x = clusters, y = VALORACION_PRECIO_RESTAUR), data =
resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



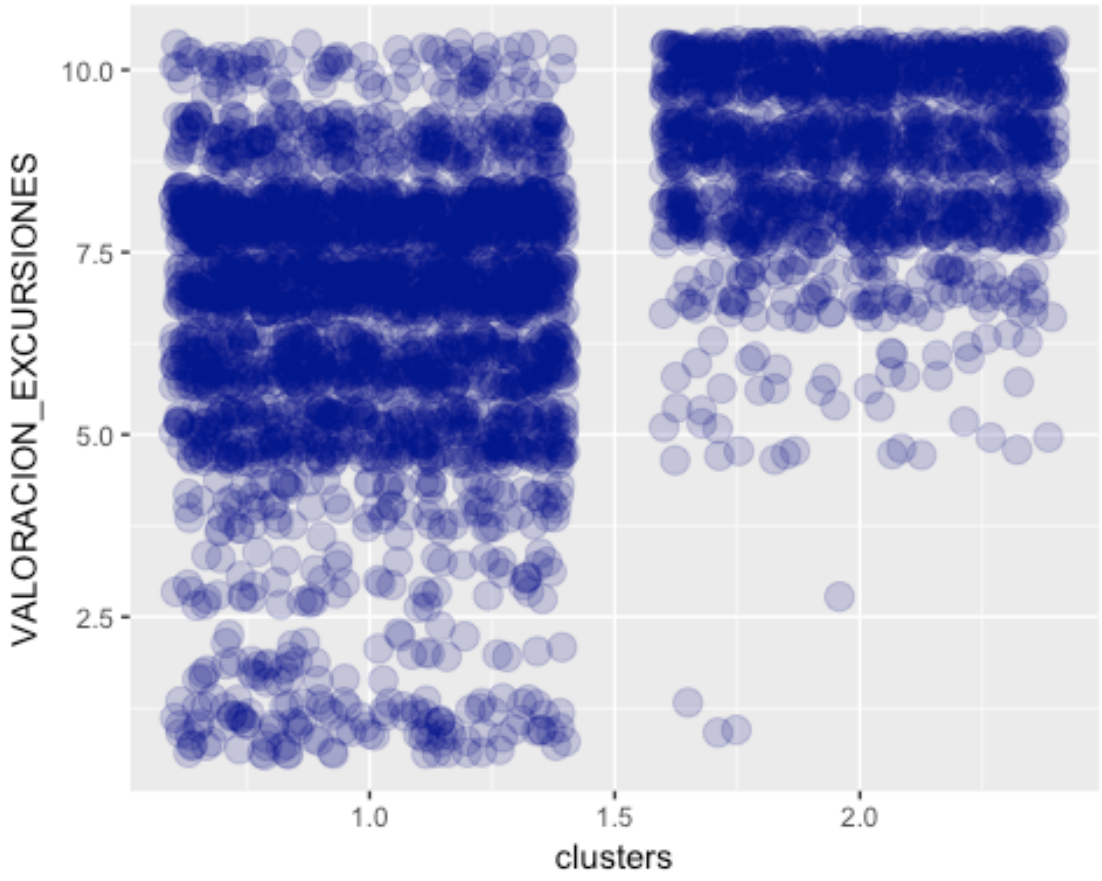
En el CLUSTER 2 valoran más EL PRECIO DE LOS RESTAURANTES que en el CLUSTER 1.

```
ggplot(aes(x = clusters, y = VALORACION_AMBIENTE_NOCTURNO), data =
resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



En el CLUSTER 2 valoran más el AMBIENTE NOCTURNO que en el CLUSTER 1.

```
ggplot(aes(x = clusters, y = VALORACION_EXCURSIONES), data = resultado) +
   geom_point(size = 4, alpha=0.2 , position="jitter",  colour = 'dark
blue')
```



En el CLUSTER 2 valoran más LAS EXCURSIONES que en el CLUSTER 1.

CONCLUSIÓN FINAL:

Por lo tanto a la vista de los resultados expresados anteriormente podemos decir lo siguiente: - No existen dos perfiles demográficos claramente definidos y segmentados, pero podemos decir que demográficamente hay dos grupos que se diferencias PRINCIPALMENTE por las siguientes características: · CLUSTER 1: En este cluster predominan los hombres y las impresiones generales suelen ser más elevadas que en el CLUSTER 2. · CLUSTER 2: En este cluster no podemos decir que predominen las mujeres, ya que es un cluster bastante equilibrado en cuanto a proporción de sexos, pero si podemos decir que la impresión media tiende a ser inferior en comparación con e cluster 1.

- Dentro de cada cluster hemos escogido una serie de variables para comprobar si existe una segmentación de opiniones entre clusters, comprobando positivamente que es así: · A diferencia del CLUSTER 1, en el CLUSTER 2 se valoran más los siguientes aspectos: TRATO ALOJAMIENTO, ALOJAMIENTO, GASTRONOMIA ALOJAMIENTO, LIMPIEZA, GASTRONOMIA LOCAL, PRECIOS RESTAURANTES, AMBIENTE NOCTURNO y EXCURSIONES.

Por lo tanto concluimos el informe explicando que a la vista del análisis, el cluster 1 (formado por una mayoría de hombres) tiende a dar mejores puntuaciones a los

alojamientos, y tiende a ser menos exigente según se ha visto en sus valoraciones de diferentes características en comparación con el cluster 2.