Métodos Multivariados II

Centro de Investigación en Matemáticas A.C.

www.cimat.mx

Apéndice D: Análisis de Conglomerados en R

Cuando se aborda el problema de clasificación de datos dado un determinado grupo de mediciones realizadas a observaciones o individuos bajo estudio existen dos perspectivas diferentes:

1) Se desea investigar si existen algunos grupos naturales o clases de individuos a partir de la estructura de las mediciones reportadas, o
2) Clasificar a los individuos según un conjunto de grupos existentes y ya definidos previamente.

El análisis de conglomerados es un término genérico que engloba una amplia gama de métodos numéricos para examinar datos multivariados respecto al primer caso, es decir, de descubrir grupos de observaciones que son homogéneos (comparten características similares de acuerdo a las variables bajo estudio) y están separados de otros grupos.

El análisis de conglomerados se aplica en muchos campos, como las ciencias naturales, las ciencias médicas, la economía, el marketing, etc. En medicina, por ejemplo, descubrir que una muestra de pacientes con mediciones en una variedad de características y síntomas en realidad consiste en un pequeño número de grupos dentro de los cuales estas características son relativamente similares, y entre los cuales son diferentes, podría tener implicaciones importantes tanto en términos de tratamiento futuro y para investigar la etiología de una condición. En marketing, es útil construir y describir los diferentes segmentos de un mercado a partir de una encuesta sobre consumidores potenciales. Por otro lado, una compañía de seguros podría estar interesada en la distinción entre clases de clientes potenciales para que pueda obtener precios óptimos por sus servicios. Más recientemente, las técnicas de análisis de conglomerados se han aplicado a al análisis de imágenes y búsqueda de patrones.

En psicología, el análisis de conglomerados se usa para encontrar tipos de personalidades en el

base de cuestionarios En la arqueología, se aplica para clasificar los objetos de arte en diferentes períodos de tiempo. Las técnicas de agrupamiento esencialmente intentan formalizar lo que los observadores humanos hacen bien en dos o tres dimensiones.

Los individuos que pertenecen a un determinado grupo (conglomerado o clúster) deben ser lo más homogéneos posible entre sí y las diferencias entre los diversos grupos lo más grandes posible. El análisis de conglomerados se puede dividir en dos pasos fundamentales:

1) Elección de una medida de proximidad: *Uno verifica cada par de observaciones (objetos) por la similitud de sus valores. Una medida de similitud (proximidad) se define para medir la "cercanía" de los objetos. Cuanto más "cerca" están, más homogéneos son.*
2) Elección del algoritmo de creación de grupos: *A partir de la proximidad, se establece una estrategia para asignar los objetos o individuos a los grupos de modo que las diferencias entre los grupos se vuelvan grandes y las observaciones en un grupo se vuelvan lo más parecidas posible.*

De forma general existen dos tipos de procedimientos

a) Métodos jerárquicos
b) Métodos no jerárquicos

D.1. Métodos jerárquicos

En una clasificación jerárquica, la lógica de clasificación puede entenderse como una serie de particiones que se hacen a un único *clúster* que contiene a todos los individuos, hasta encontrar *n clústers* que contienen cada uno a un único individuo (técnicas disosiativos) o, empezar de los *n clústers* que contienen a cada uno de los individuos e irlos fusionando sucesivamente hasta tener un único *clúster* de todos los individuos (técnicas aglomerativas). Hay que puntualizar que, una vez realizada la fusión o división según el caso, está es irreversibles, es decir que cuando un algoritmo ha colocado a dos individuos en el mismo grupo, no pueden aparecer posteriormente en diferentes grupos.

Dado que todas las técnicas jerárquicas reducen en última instancia los datos a un único grupo que contiene a todos los individuos o en varios grupos de un solo individuo, el número “correcto” de agrupaciones con las que se desea trabajar después del estudio queda a criterio del investigador. Las clasificaciones jerárquicas pueden representarse mediante un diagrama bidimensional conocido como *dendrograma*, que ilustra las fusiones o divisiones realizadas en cada etapa del análisis.

El ejemplo que se muestra a continuación, describe el procedimiento y análisis a realizar cuando se plantea como objetivo buscar agrupar las observaciones de un estudio.

D.1.1. Caso: Actitud de compradores

Retomando el ejemplo analizado en el material de apoyo de esta unidad, los directivos del centro comercial ubicado en la ciudad buscan identificar segmentos de consumidores a partir de los datos obtenidos a través de una encuesta en relación con la actitud que tienen respecto a la actividad específica de ir de compras para formular, después de su análisis, distintas estrategias de servicio para cada uno de los segmentos identificados que permita aumentar su frecuencia de compra.

Tras diversas dinámicas de grupo, se encuentra que algunas de las motivaciones que llevan a las personas a ir de compras habitualmente son las que se plantean a continuación, y que por lo tanto serán las variables del estudio:

$\mathbf{X_1}$: Es un acto divertido en sí mismo.
$\mathbf{X_2}$: Intento ir poco porque compro compulsivamente y es malo para mi
$\mathbf{X_3}$: Voy con frecuencia porque aprovecho para cenar fuera con mi pareja.
$\mathbf{X_4}$: Me encanta la aventura de encontrar productos a buen precio.
$\mathbf{X_5}$: No me atrae especialmente, voy por obligación o por necesidad.
$\mathbf{X_6}$: Puedes ahorrar mucho dinero si vas a comprar con frecuencia y estás informado.

Para cada una de éstas posibles razones, el investigador pide a un conjunto de consumidores que valoren en qué medida están de acuerdo con la aseveración en una escala de 7 puntos donde 1 significa que se está totalmente en desacuerdo y 7 totalmente de acuerdo.

D.1.1.1. Selección de variables

La primera etapa del cualquiera de las técnicas que conforman el análisis de conglomerados consiste en seleccionar aquellas características (variables de interés) que permitan definir a las agrupaciones que se desean formar. Será a partir de las mediciones de los individuos en estas variables como se calculará la similitud entre pares de observaciones.

Para el caso abordado las mediciones realizadas a 20 personas para las 6 variables descritas anteriormente se encuentran en el archivo *Actitud_Compras.xlsx*. El archivo contiene variables adicionales sobre ciertos rasgos de las personas entrevistadas.

```
 library(readr)
 Gorriones <- read_excel("G:/Mi unidad/CIMAT/EME/MULTIVARIADOS
II/PRED/MATERIAL/CONGLOMERADOS/Actitud_Compras.xlsx")
 View(Actitud_Compras)
```

Recuerde colocar la ruta correcta (sección en color verde) en su computadora donde tiene guardado el archivo antes mencionado.

D.1.1.2. Elegir la medida de asociación

Los métodos jerárquicos parten de una matriz de distancias o similaridades entre los elementos de la muestra y construyen una jerarquía basada en una distancia. Si todas las variables son continuas, la distancia más utilizada es la distancia euclídea entre las variables estandarizadas.

Es recomendable estandarizar las variables previo cálculo de la distancia ya que el valor resultante presentará un sesgo sobre todo de las variables con valores más grandes, y el resultado del análisis puede cambiar completamente. Al estandarizar, se da a priori un peso semejante a las variables, con independencia de su variabilidad original, lo que puede no ser siempre adecuado. Para el caso de variables cualitativas dicotómicas consultar las medidas de similitud que se mencionan en el material de esta unidad.

En el ejemplo como las variables están dadas en la misma escala no es necesario realizar una estandarización previa. En caso de tener que estandarizar se debe utilizar el comando `scale()`. Las columnas 2 a la 7 contienen las variables consideradas para realizar la agrupación.

```
# Calcular la matriz de distancias inicial usando la distancia
# euclidiana. Revisar en la ayuda los métodos disponibles.
d <- dist(Actitud_Compras[,2:7], method = "euclidean")
round(d,2)
```

```
      1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19
2  8.00
3  2.83 8.25
4  5.57 5.57 6.56
5  8.31 2.65 9.11 6.63
6  1.73 6.86 3.32 4.47 7.21
7  2.24 6.24 3.32 4.69 6.48 1.41
8  2.24 8.77 1.73 6.00 9.49 2.83 3.16
9  6.93 2.83 8.00 5.57 2.24 5.92 5.39 8.31
10 6.93 4.24 7.48 2.24 5.74 5.74 5.57 7.28 4.90
11 7.75 2.00 8.37 6.24 1.73 6.86 6.08 8.89 2.00 5.29
12 2.65 5.92 3.32 3.46 6.78 2.00 2.00 3.16 5.57 4.58 6.08
13 8.06 1.73 7.81 5.83 4.00 7.07 6.32 8.49 4.12 4.36 3.00 5.83
14 6.78 6.00 7.62 1.73 7.14 5.57 5.74 7.00 6.16 2.00 6.93 4.80 6.24
15 3.61 7.00 4.36 4.69 7.62 3.16 3.74 4.24 6.71 6.24 7.14 2.83 7.21 6.24
16 6.93 5.29 7.62 2.24 6.40 5.74 5.57 7.14 5.66 1.41 6.16 4.80 5.39 1.41 6.56
17 3.00 7.42 4.80 4.90 7.21 2.83 2.45 4.00 6.40 6.24 7.00 3.16 7.62 6.08 4.00 5.92
18 7.48 4.90 8.37 4.12 6.40 6.71 6.86 8.43 4.90 3.74 5.48 5.57 5.39 4.69 6.56 4.90 7.68
19 7.48 6.63 7.75 2.65 8.31 6.56 6.71 7.28 7.48 2.83 7.75 5.20 6.40 2.45 6.40 2.83 7.00 4.90
20 8.31 3.00 8.89 7.07 3.16 7.35 6.78 9.49 2.24 5.92 2.65 6.93 4.00 7.42 8.25 6.86 8.25 5.57 8.54
```

D.1.1.3. Elegir la técnica de agrupación

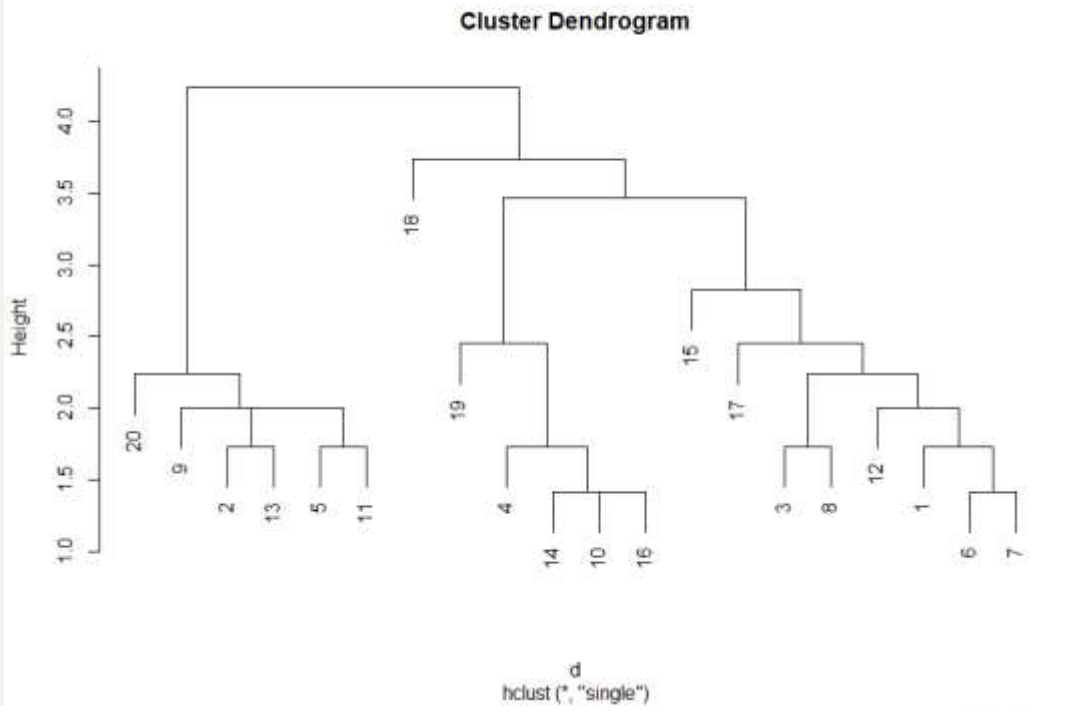
Finalizado el cálculo de la matriz de distancia inicial corresponden seleccionar la técnica de agrupación. En el caso de los métodos aglomerativos como se mencionó anteriormente, se comienza con tantos clústers como individuos se estén analizando en la base de datos y se buscan agrupar aquellos dos elementos que tengan la distancia más próxima de acuerdo con la matriz de distancias hasta llegar a un único grupo que contenga a todas las observaciones. Es de mencionar que después de cada agrupación es necesario recalcular la matriz de distancia siguiendo la formula establecida por cada una de las técnicas de agrupación.

Para el ejemplo se decide comparar la agrupación resultante por el método del vecino más cercano y la técnica de Ward.

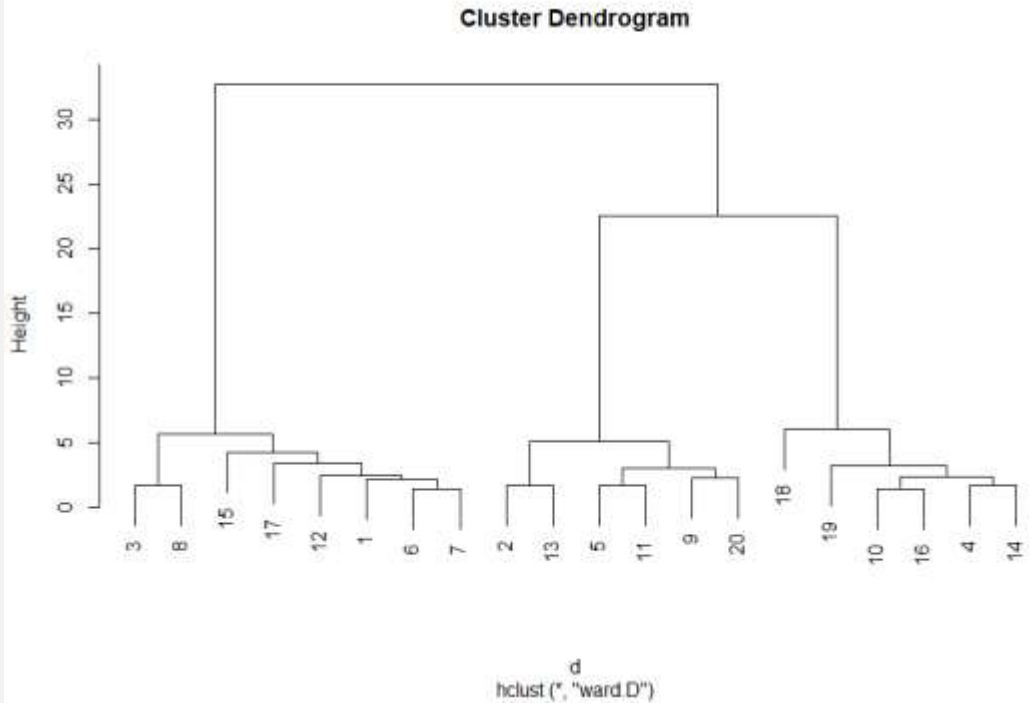
```r
# Agrupacion aplicando el método del vecino mas cercano
Actitud_Compras.cluster1 <- hclust(d, method="single")

# Agrupacion aplicando el método de Ward
Actitud_Compras.cluster2 <- hclust(d, method="ward.D")

# Revisar en la ayuda los métodos disponibles.
```

Para realizar la comparación se generan los *dendogramas* para cada uno de los métodos utilizados.

```r
# Dendograma del vecino mas cercano
plot(Actitud_Compras.cluster1)
```



```
# Dendograma de método de Ward
plot(Actitud_Compras.cluster2)
```



Se aprecia que ambos dendogramas presentan diferencias en la forma en que se fueron realizando las agrupaciones. Mientras que el dendograma generado de aplicar el método de Ward se distinguen con claridad tres agrupaciones de datos a una distancia corta en comparación con la distancia de la agrupación total. El dendograma del método del vecino más cercano no refleja con la misma claridad una definición de un número de grupos a pesar de que la distancia de la agrupación final es más próxima que con Ward.

En ocasiones dada la cantidad de observaciones que se desean agrupar se dificulta la visualización del dendograma, por lo que es necesario dar seguimiento al historial de cómo se fue generando cada agrupación. Para revisar este historial se debe ejecutar la instrucción correspondiente como se muestra a continuación.

```
# Historial de agrupamiento siguiendo el método de Ward
historial <- data.frame(cbind(Actitud_Compras.cluster1$merge,
round(Actitud_Compras.cluster1$height,3)))
colnames(historial) <- c("individuo 1","individuo 2","distancia")
historial
```

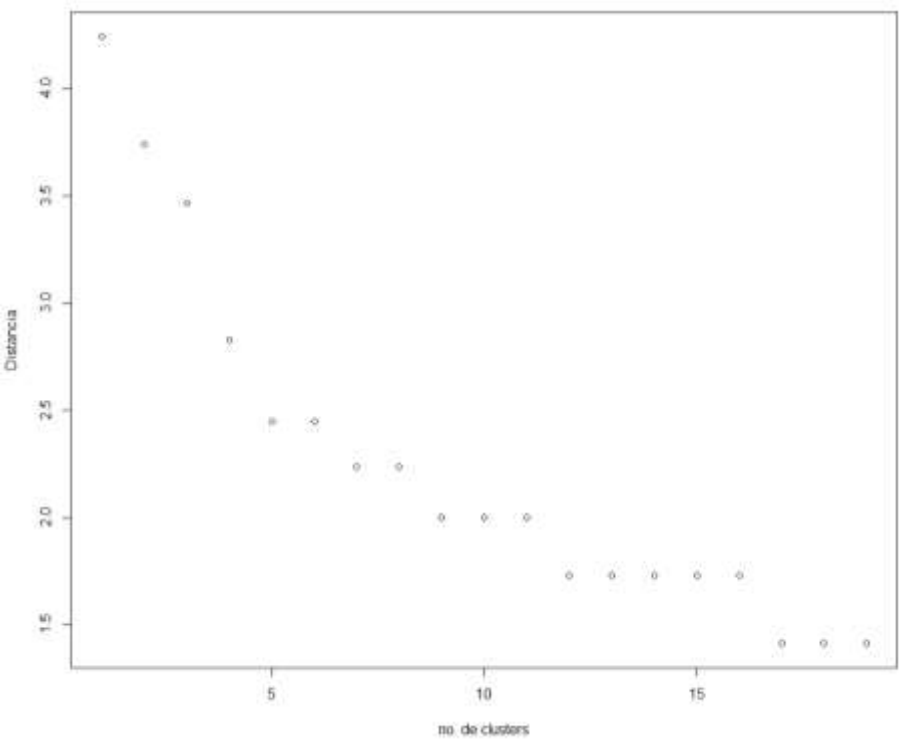
```
   individuo 1 individuo 2 distancia
1           -6          -7     1.414
2          -10         -16     1.414
3           -2         -13     1.732
4           -3          -8     1.732
5           -4         -14     1.732
6           -5         -11     1.732
7           -1           1     2.174
8           -9         -20     2.236
9            2           5     2.370
10         -12           7     2.426
11           6           8     3.038
12         -19           9     3.198
13         -17          10     3.373
14         -15          13     4.215
15           3          11     5.042
16           4          14     5.692
17         -18          12     5.999
18          15          17    22.548
19          16          18    32.690
```

La tabla anterior indica que la primera agrupación que se realizó siguiendo el método de Ward que entre el individuo 6 y el individuo 7 a una distancia de cercanía de 1.414. La segunda agrupación la realiza entre el individuo 10 y el 16 a la misma distancia. La agrupación número siete indica que el individuo 1 fue agrupado con el clúster formado en la primera iteración (individuos 6 y 7) a una distancia de 2.174. En la novena iteración, se conformó un clúster que incluye a los agrupados en la iteración 2 (individuos 10 y 16) y los agrupados en la iteración 5 (individuos 4 y 14) a una distancia de 2.370. Esto puede comprobarse al revisar el dendograma respectivo.

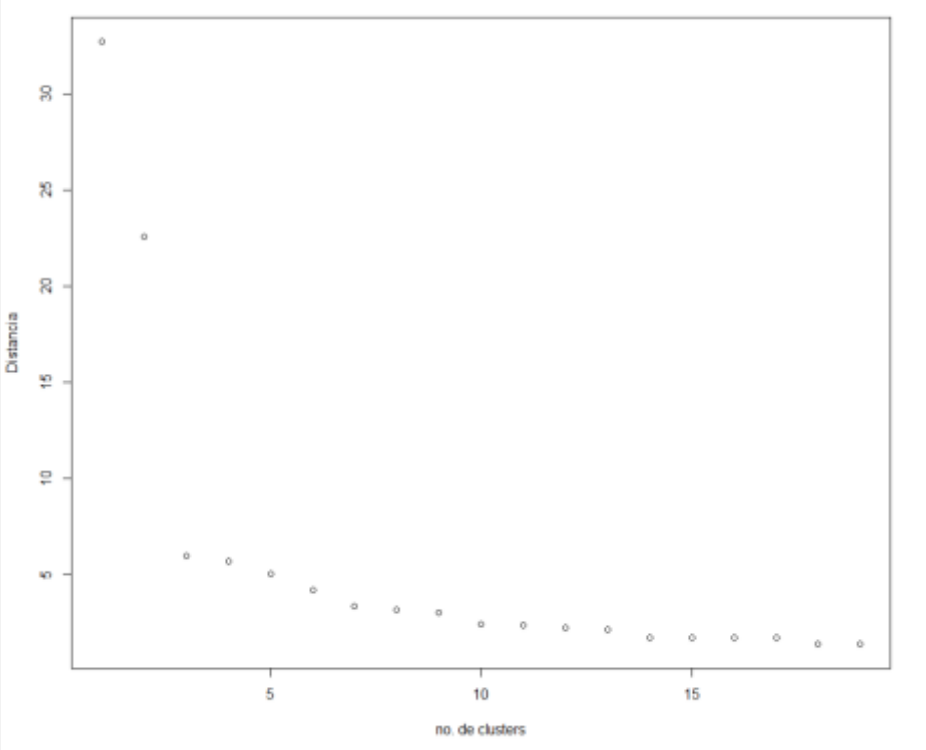
D.1.1.4. Validación e interpretación de resultados

Corresponde ahora determinar el número de grupos que se desea considerar para establecer una campaña de mercadotecnia dirigida. Como ya se mencionó en el material del curso esta es una tarea que depende del conocimiento del contexto del problema abordado y criterio del analista. Un apoyo que en ocasiones puede ser de utilidad cuando no se tiene previamente definido un número de grupos o la experiencia en el tipo de problema abordado para determinar dicha cantidad, es realizar una gráfica donde se represente la distancia de aglomeración y el número de grupos a esa distancia. El número de agrupaciones a considerar será el punto en donde se estabilice la distancia de formación de los grupos, es decir, en donde la curva de describe la gráfica empiece a estabilizarse.

```r
# Utilizando resultados del vecino mas cercano
plot(rev(Actitud_Compras.cluster1$height), ylab = "Distancia", xlab =
"no. de clusters")
```

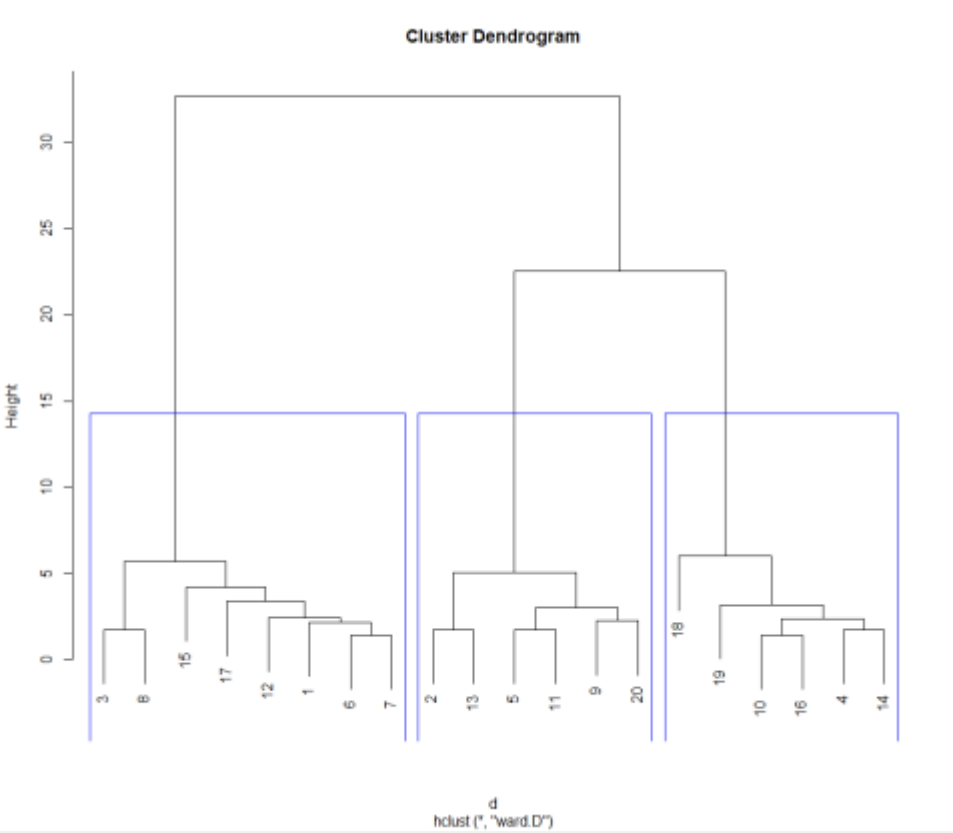



```
# Utilizando resultados del método de Ward
plot(rev(Actitud_Compras.cluster2$height), ylab = "Distancia", xlab =
"no. de clusters")
```



Del primer gráfico se observa que la curva de cambio empieza ser más estable al pasar de tener cinco a seis clústers, por lo que para los propósitos del estudio pueden considerarse cinco tipos de grupos de personas. Por otro lado, al revisar el grafico obtenido de aplicar la técnica de Ward la estabilidad se alcanza al pasar de tres a cuatro grupos, por lo que correspondería analizar solo tres grupos. El siguiente paso es determinar que individuos están agrupados en cada uno de los clústers considerados.

```
# Establecer el número de agrupaciones a retener y guardar la observa
# ciones agrupadas en dichos grupos para el causo utilizando el méto-
# do de Ward
cluster <- cutree(Actitud_Compras.cluster2, k=3)
# el valor de k indica el no. de clusters a retener en el estudio

# Se muestra en el dendograma los clusters considerados
plot(Actitud_Compras.cluster2)
rect.hclust(Actitud_Compras.cluster2, k=3, border="blue")
```



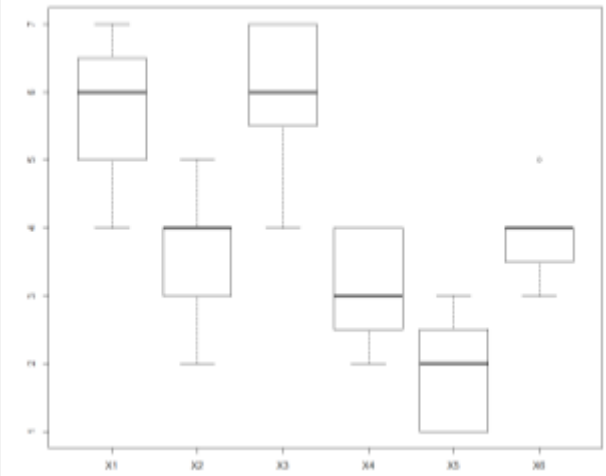

```
# Despliega el total de elementos que contiene cada cluster
table(cluster)

cluster
1 2 3
8 6 6
```

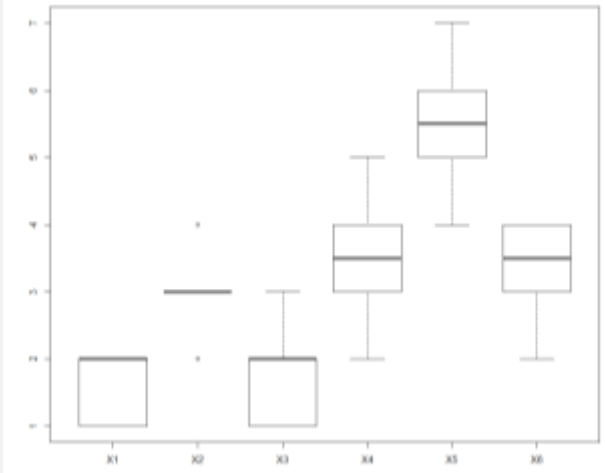
```
# Desplegar cada uno de los casos agrupados en cada cluster
sapply(unique(cluster),function(g)Actitud_Compras$Caso[cluster == g])

[[1]]
[1]  1  3  6  7  8 12 15 17

[[2]]
[1]  2  5  9 11 13 20

[[3]]
[1]  4 10 14 16 18 19


# Otra forma de desplegar las agrupaciones
table(Actitud_Compras$Caso,cluster)

    cluster
     1 2 3
  1  1 0 0
  2  0 1 0
  3  1 0 0
  4  0 0 1
  5  0 1 0
  6  1 0 0
  7  1 0 0
  8  1 0 0
  9  0 1 0
  10 0 0 1
  11 0 1 0
  12 1 0 0
  13 0 1 0
  14 0 0 1
  15 1 0 0
  16 0 0 1
  17 1 0 0
  18 0 0 1
  19 0 0 1
  20 0 1 0
```

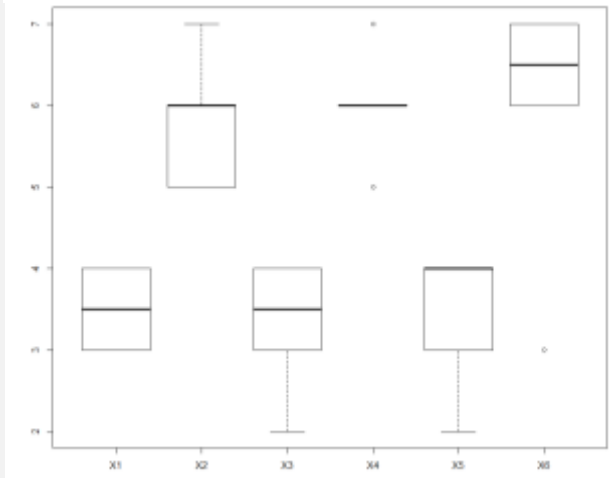
El siguiente paso es caracterizar a los individuos que conforman cada grupo para validar que comparten características similares de acuerdo a las variables bajo estudio y que los grupos son diferenciables entre ellos mismos. Para ellos es necesario emplear herramientas descriptivas tanto para las variables que fueron consideradas en el estudio como aquellas que fueron dejadas a un lado (*sexo* y *edad*). El análisis solo se realiza de los resultados de la agrupación utilizando el método de Ward.

```
# Tabla para desplegar los valores promedio de las variables en cada
# cluster
a3 = aggregate(Actitud_Compras[,2:7],list(cluster),mean)
data.frame(Cluster=a3[,1],Freq=as.vector(table(cluster)),round(a3
[,-1],3))
```

```
  Cluster Freq    X1    X2    X3    X4    X5    X6
1       1    8 5.750 3.625 6.000 3.125 1.875 3.875
2       2    6 1.667 3.000 1.833 3.500 5.500 3.333
3       3    6 3.500 5.833 3.333 6.000 3.500 6.000

# Graficas de boxplot para cada variable en cada cluster
boxplot(Actitud_Compras[c(Actitud_Compras$Caso[cluster == 1]),2:7])
boxplot(Actitud_Compras[c(Actitud_Compras$Caso[cluster == 2]),2:7])
boxplot(Actitud_Compras[c(Actitud_Compras$Caso[cluster == 3]),2:7])
```





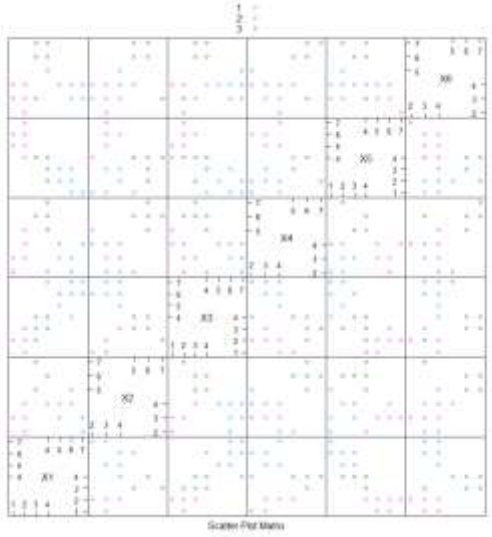


```
# Frecuencia de las variables sexo y edad en cada grupo
table(cluster,Actitud_Compras$sexo)

cluster Hombre Mujer
      1      0     8
      2      5     1
      3      2     4

table(cluster,Actitud_Compras$edad)

cluster 20-30 31-40 41-mas
      1     5     2      1
      2     1     5      0
      3     0     2      4
```

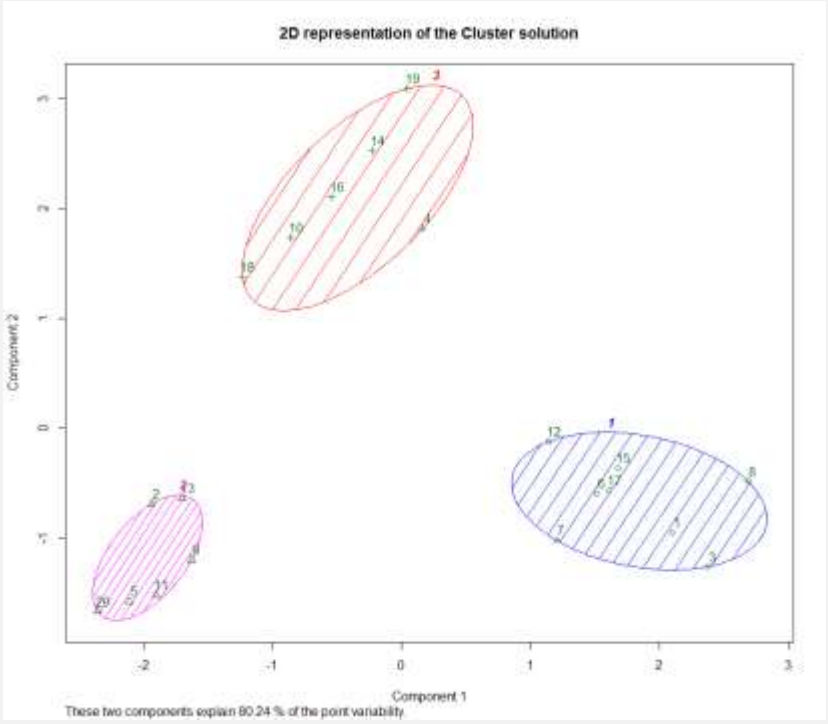
A partir de los datos anteriores observa una buena definición de las características de cada uno de los tres clústers. El grupo 1 puede caracterizarse como un grupo de mujeres mayoritariamente jóvenes que no consideran comprar como una obligación y lo hacen porque es un acto divertido y/o aprovechan para cenar con su pareja. Caso contrario al grupo 2 que está conformado mayoritariamente por hombres de entre 31 y 40 años que van al centro comercial por obligación o necesidad y no lo encuentran divertido. Por su lado el grupo 3 son las personas de mayor edad que asisten porque les encanta la aventura de encontrar productos a buen precio y porque consideran que ahorran dinero al saber cómo evolucionan los precios por ir con frecuencia.

Dos representaciones gráficas que pueden ayudar a determinar si la agrupación es conveniente son las siguientes:

```
# Se debe instalar y ejecutar el paquete lattice
install.packages("lattice")
library(lattice)

# Muestra la matriz de diagramas de dispersión entre las variables
# incluidas en el análisis
splom(Actitud_Compras[,2:7],groups=cluster,auto.key=TRUE)
```



```
# Visualizacion en 2D de las agrupaciones con el paquete cluster
install.packages("cluster")
library(cluster)
clusplot(Actitud_Compras[,2:7], cluster, main='2D representation of
the Cluster solution', color=TRUE, shade=TRUE, labels=2, lines=0)
```



Como se observa en la matriz de gráficos de dispersión hay casos en donde es fácil separar a las observaciones y otros donde hay traslapes. Al analizar el último gráfico, si se proyecta en un gráfico biplot el análisis realizado, se distinguen muy claramente los tres grupos de individuos resultantes. Por lo que se puede concluir que es muy razonable considerar estos tres clústers.

A partir de esta caracterización, el analista y/o los directivos deben decidir si está conforme con los resultados. En caso afirmativo, se procedería a establecer las estrategias específicas para cada perfil, en caso contrario, se debe repetir el análisis anterior para una nueva cantidad de grupos y comparar si fue mejor que el escenario anterior.

D.2. Métodos no jerárquicos

Como se pudo experimentar en con el caso anterior, los métodos jerárquicos parten de un funcionamiento de agrupar conglomerados hasta llegar a uno solo o dividir un único conglomerado en varios, pero manteniendo una secuencia dada. En el caso de los métodos no jerárquicos, no contemplan esa funcionalidad. Los métodos no jerárquicos, tiene como función realizar solo una partición de los objetos en K clústers o conglomerados sin seguir una secuencia como los

métodos jerárquicos. Lo crucial en el uso de estas técnicas, viene dado por la elección *a priori* del número de clústers por parte del analista. En esto radica la diferencia fundamental de su uso, ya que por lo general se busca comprar, en la mayoría de las ocasiones, una suposición sobre las relaciones que se establecen entre los individuos del estudio.

Otras diferencias importantes son que el proceso de asignación de los objetos se hace mediante una optimización del criterio de selección y, no se trabaja con una matriz de distancias o de similitudes sino con los datos originales.

El método de *K*-medias (*K-mean*) es uno de los más empleados. Esta técnica trata. Mediante la aplicación de un algoritmo, asignar a cada uno de las observaciones a uno de los K conglomerados (grupos o clústers) previamente fijados que tenga el centroide más próximo. En este método propuesto por MacQueen en el año 1972, la clave radica en que el centroide se calcula a partir de los miembros del clúster tras cada asignación y no al final de cada ciclo.

D.2.1. Caso: Actitud de compradores

Se continúa trabajando con el caso referente al estudio de los motivos por los que la gente va a comprar.

D.2.1.1. Elegir la técnica de agrupación

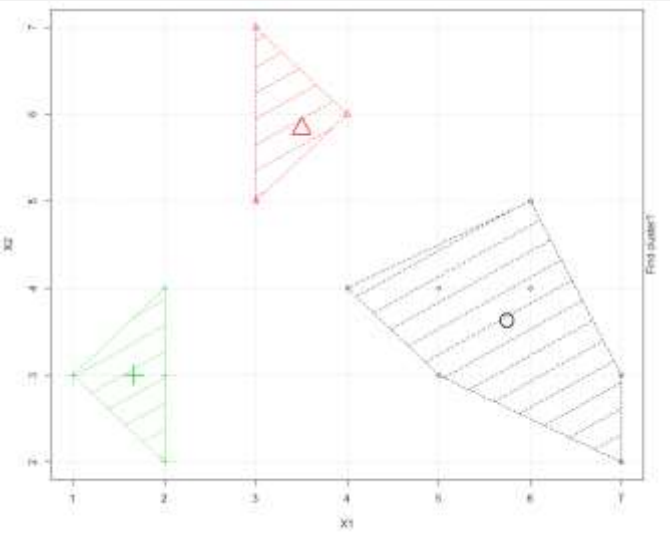
Se aplica la técnica de *K*-medias considerando tres grupos para realizar una comparación con el resultado obtenido de la aplicación de un método jerárquico.

```
# Agrupación de los datos aplicando el algoritmo de k-medias. El
# valor de K indica el no. de clústers en que se agruparan los datos
Actitud_Compras.k.means <- kmeans(Actitud_Compras[,2:7], 3)

# Muestra el valor los centroides de cada grupo en cada una de las
# variables analizadas
round(Actitud_Compras.k.means$centers,3)
```

```
     X1    X2    X3    X4    X5    X6
1 5.333 4.000 5.833 2.833 2.167 4.000
2 2.583 4.417 2.583 4.750 4.500 4.667
3 7.000 2.500 6.500 4.000 1.000 3.500
```

A diferencia de los métodos jerárquicos, no es posible realizar un dendograma en donde se muestre el historial de las agrupaciones. Sin embargo, existe la posibilidad a través del comando `kmeans.ani()` de darse una idea de la forma en que la técnica fue realizando la agrupación de los individuos mediante una animación. Esta animación solo muestra el gráfico bidemensional para las dos primeras variables consideradas en el análisis.

```
# Debe de instalarse el paquete animation
install.packages("animation")
library(animation)
cl<- kmeans.ani(Actitud_Compras[,2:7], 3)
```



D.2.1.2. Validación e interpretación de resultados

Al igual que en caso jerárquico es necesario comprobar la coherencia entre los miembros de cada uno de los clústers. En primer lugar, es necesario saber cómo quedaron agrupados los individuos para posteriormente hacer un análisis descriptivo de cada grupo.

```
# Para los tres grupos
data.frame(Cluster=c(1,2,3),Frecuencia=as.vector(Actitud_Compras.k
.means$size))

  Cluster Frecuencia
1       1          6
2       2         12
3       3          2

# Despliega los cada uno de los casos
data.frame(Observacion=Actitud_Compras$Caso[o],Cluster=Actitud_
Compras.k.means$cluster[o])

   Observacion Cluster
1            1       1
2            6       1
3            7       1
4           12       1
5           15       1
6           17       1
7            2       2
8            4       2
9            5       2
10           9       2
11          10       2
12          11       2
13          13       2
```

```
14          14       2
15          16       2
16          18       2
17          19       2
18          20       2
19           3       3
20           8       3
```

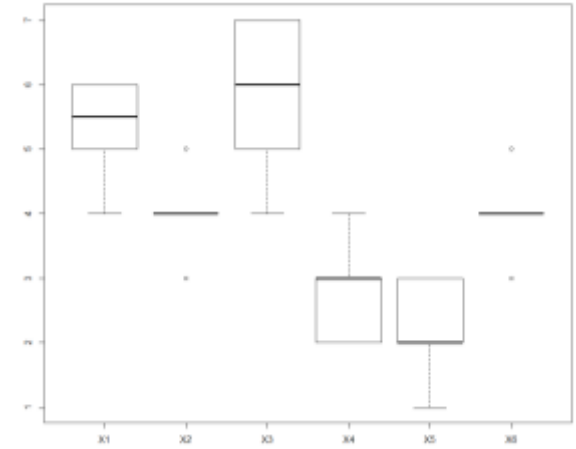
```
# Otra forma de desplegar las agrupaciones
o=order(Actitud_Compras.k.means$cluster)
table(Actitud_Compras$Caso[o],Actitud_Compras.k.means$cluster[o])
```

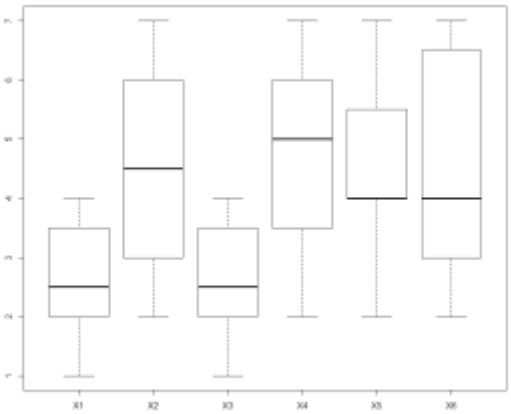
```
    1 2 3
 1  1 0 0
 2  0 1 0
 3  0 0 1
 4  0 1 0
 5  0 1 0
 6  1 0 0
 7  1 0 0
 8  0 0 1
 9  0 1 0
 10 0 1 0
 11 0 1 0
 12 1 0 0
 13 0 1 0
 14 0 1 0
 15 1 0 0
 16 0 1 0
 17 1 0 0
 18 0 1 0
 19 0 1 0
 20 0 1 0
```

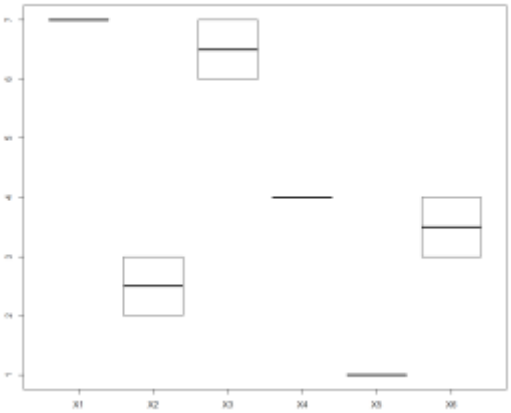
```
# Tabla para desplegar los valores promedio de las variables en cada
# cluster
data.frame(Cluster=a3[,1],Freq=as.vector(table(Actitud_Compras.k.
means$cluster)),round(a3[,-1],3))
```

```
  Cluster Freq    X1    X2    X3    X4    X5    X6
1       1    6 5.333 4.000 5.833 2.833 2.167 4.000
2       2   12 2.583 4.417 2.583 4.750 4.500 4.667
3       3    2 7.000 2.500 6.500 4.000 1.000 3.500
```

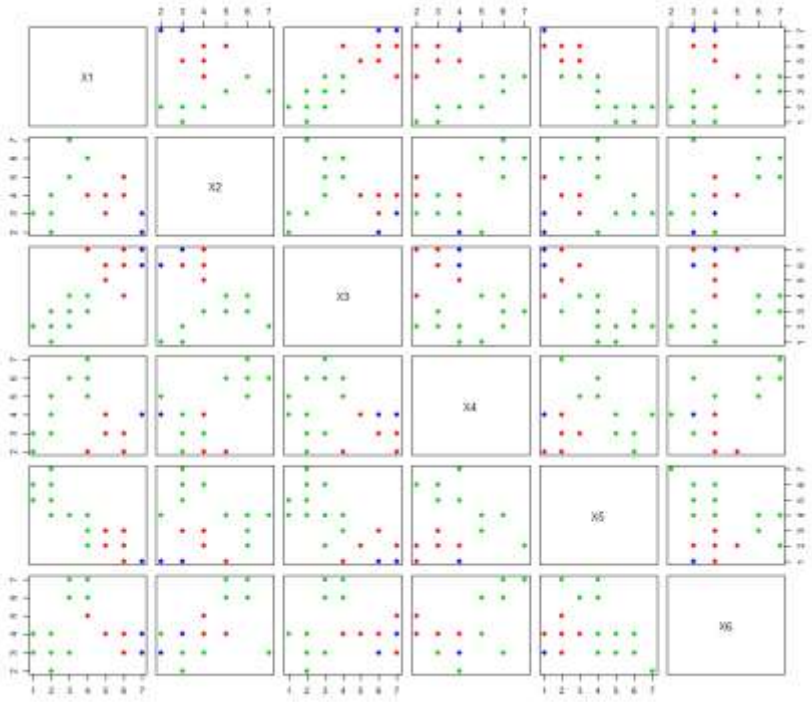
```
# Graficas de boxplot para cada variable en cada cluster
boxplot(Actitud_Compras[c(Actitud_Compras$Caso[Actitud_Compras.k.
means$cluster == 1]),2:7])
boxplot(Actitud_Compras[c(Actitud_Compras$Caso[Actitud_Compras.k.
means$cluster == 2]),2:7])
boxplot(Actitud_Compras[c(Actitud_Compras$Caso[Actitud_Compras.k.
means$cluster == 3]),2:7])
```



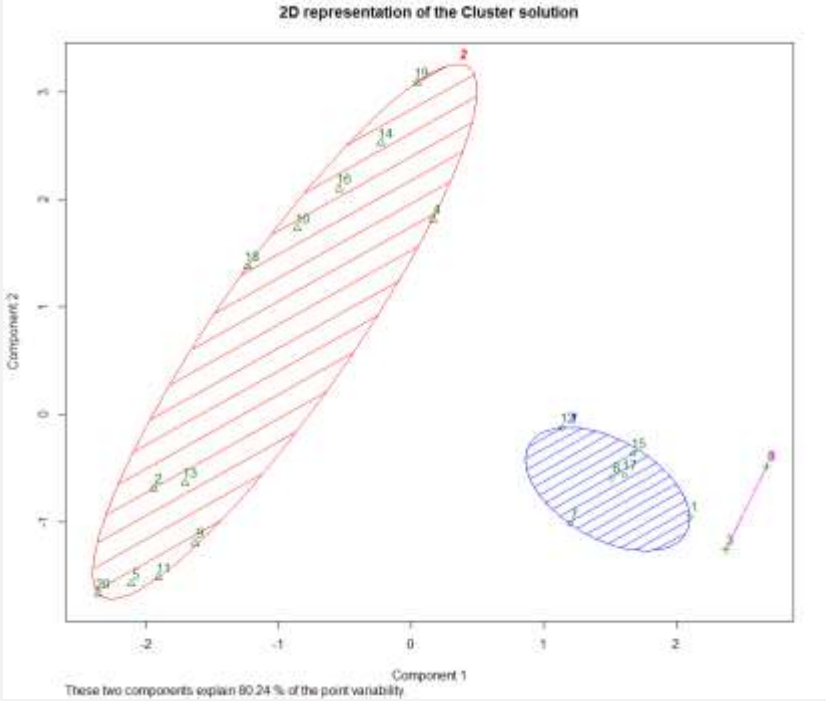




```
# Frecuencia de las variables sexo y edad en cada grupo
table(Actitud_Compras.k.means$cluster,Actitud_Compras$sexo)

    Hombre Mujer
  1      0     6
  2      7     5
  3      0     2


table(Actitud_Compras.k.means$cluster,Actitud_Compras$edad)

    20-30 31-40 41-mas
  1     4     1      1
  2     1     7      4
  3     1     1      0


# Muestra la matriz de diagramas de dispersión entre las variables
# incluidas en el análisis
plot(Actitud_Compras[,2:7],col =(Actitud_Compras.k.means$cluster +1),
pch=20, cex=2)
```




```
library(cluster)
clusplot(Actitud_Compras[,2:7], Actitud_Compras.k.means$cluster,
main='2D representation of the Cluster solution',color=TRUE, shade=
TRUE,labels=2, lines=0)
```



De los resultados reportados más arriba, se aprecia que la asociación obtenida, al revisar el último gráfico, también es muy buena ya que se distinguen muy bien los tres grupos al proyectar los valores en el plano. A diferencia del caso jerárquico, no resulta tan claro distinguir las características que definen a cada uno de los grupos. Al revisar los boxplot de cada uno de los grupos no se alcanza a distinguir una diferenciación clara entre los individuos del segundo grupo con aquellos del primero y tercer grupo. Es recomendable repetir el estudio considerando dos o cuatro grupos y verificar si hay claridad en cuanto a las características de los individuos que conforman cada clúster.

Como conclusión, el caso que se estuvo trabajando hacer evidente que esta técnica multivariada requiere de habilidad para determinar un número adecuado de agrupaciones que permitan distinguir cada una de ellas pero que guarden cohesión los individuos que conforman el grupo. Al mismo tiempo se realza la necesidad de aplicar al menos dos técnicas del análisis de conglomerados para comparar los resultados, evaluar las agrupaciones resultantes y seleccionar aquel que mejor cumpla con el objetivo planteado por el que se efectuó el análisis.