Modelamiento Estadístico y Sistemas Recomendadores: Foro 3

Patricio Águila Márquez

Instrucciones

Considere los datos del archivo 'wholesale.csv', que contienen información de 440 clientes de un distribuidor mayorista. La base de datos contiene información sobre el gasto anual de cada cliente en productos de las siguientes categorías: frescos (Fresh), lácteos (Milk), comestibles (Grocery), congelados (Frozen), detergentes/papel (Detergents_Paper) y rotisería (Delicassen).

En base a este conjunto de datos, realice las siguientes actividades:

Actividades preliminares:

1. Cargue el conjunto de datos en la sesión de trabajo de `R` usando la función `read.table` . Utilizando la función **summary** determine el producto que generó la máxima venta, y el producto que mayor ingreso genera en promedio.

Principales estadísticos de los datos originales:

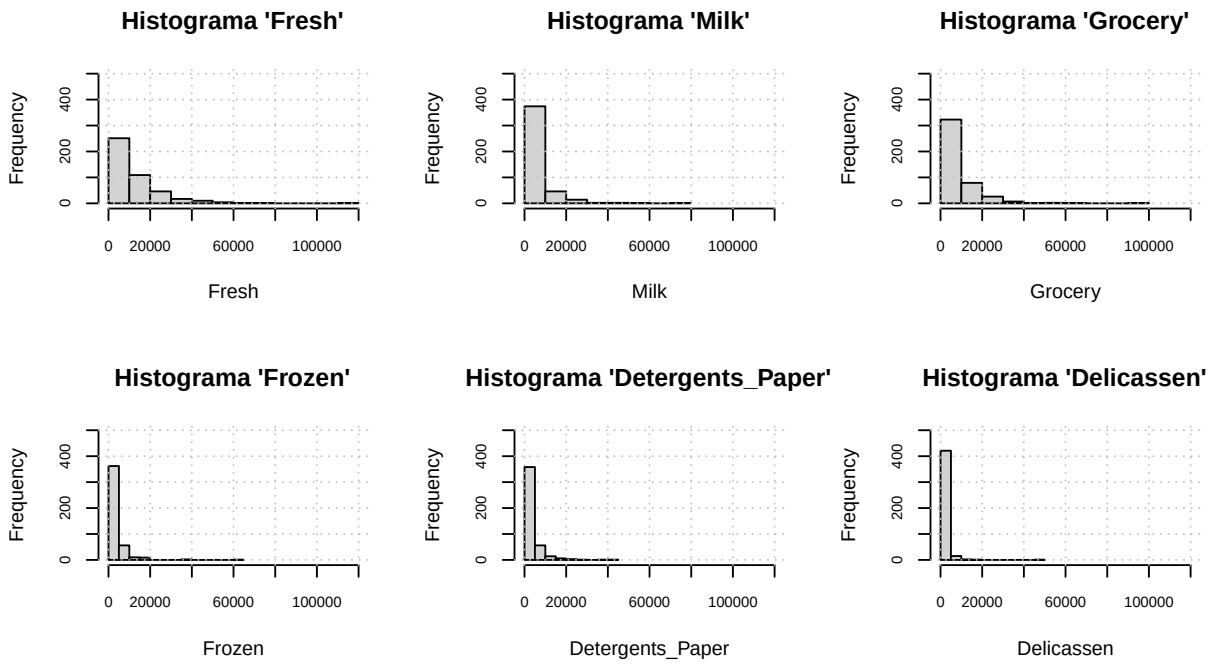
```
##      Fresh             Milk          Grocery          Frozen
##  Min.   :     3   Min.   :   55   Min.   :    3   Min.   :   25.0
##  1st Qu.:  3128   1st Qu.: 1533   1st Qu.: 2153   1st Qu.:  742.2
##  Median :  8504   Median : 3627   Median : 4756   Median : 1526.0
##  Mean   : 12000   Mean   : 5796   Mean   : 7951   Mean   : 3071.9
##  3rd Qu.: 16934   3rd Qu.: 7190   3rd Qu.:10656   3rd Qu.: 3554.2
##  Max.   :112151   Max.   :73498   Max.   :92780   Max.   :60869.0
##  Detergents_Paper    Delicassen
##  Min.   :    3.0   Min.   :    3.0
##  1st Qu.:  256.8   1st Qu.:  408.2
##  Median :  816.5   Median :  965.5
##  Mean   : 2881.5   Mean   : 1524.9
##  3rd Qu.: 3922.0   3rd Qu.: 1820.2
##  Max.   :40827.0   Max.   :47943.0
```

- *La categoría que generó la mayor venta es 'Fresh', con $112.151.*
- *La categoría que mayor ingreso genera en promedio también es 'Fresh', con media de $12.000.*

Luego, considere y responda:

a. ¿Son similares las distribuciones de venta de cada producto?

*Resp: al utilizar los **datos originales**, se observa que la distribución de las ventas en las distintas categorías son similares, concentrando la mayoría de los ingresos al principio de cada histograma.*



b. ¿Cuál es la relación entre las medias y las desviaciones estándar de cada variable? Interprete este resultado.

Resp: la relación entre la media y la desviación estándar nos indica qué tan dispersos se encuentran los datos. Mientras menos distancia hay entre la media y la desviación estándar, existe menos variabilidad en los datos. Por ejemplo, los valores de la variable 'Fresh' están más cerca de la media que aquellos de las otras categorías, lo que implica que en este punto hay una mayor densidad de datos.

```
##           Category  Mean    Sd
## 1            Fresh 12000 12647
## 2          Grocery  7951  9503
## 3             Milk  5796  7380
## 4           Frozen  3072  4855
## 5 Detergents_Paper  2881  4768
## 6       Delicassen  1525  2820
```

c. Adicionalmente, investigue qué producto representa la mayor parte de las ventas, y qué producto la menor. Comente e interprete estos resultados.

Resp: la categoría de productos 'Fresh' representa la mayor parte de las ventas (36,12%), mientras que 'Delicassen' tiene la menor participación (4,59%)

```
##           Category    Suma Porcentaje
## 1            Fresh 5280131      36.12
## 2          Grocery 3498562      23.93
## 3             Milk 2550357      17.44
## 4           Frozen 1351650       9.25
## 5 Detergents_Paper 1267857       8.67
## 6       Delicassen  670943       4.59
```

2. En lo que sigue, haremos análisis de conglomerados sobre los datos. ¿Qué utilidad podría tener este tipo de análisis desde el punto de vista del negocio para el distribuidor mayorista? Responda en el Foro dando ejemplos concretos.

- *Resp: el análisis de conglomerados serviría para segmentar a los clientes en base al monto y las categorías de productos comprados. Se podría realizar un análisis estratégico para determinar la conveniencia de conservar un cliente, o bien, deshacerse de él.*

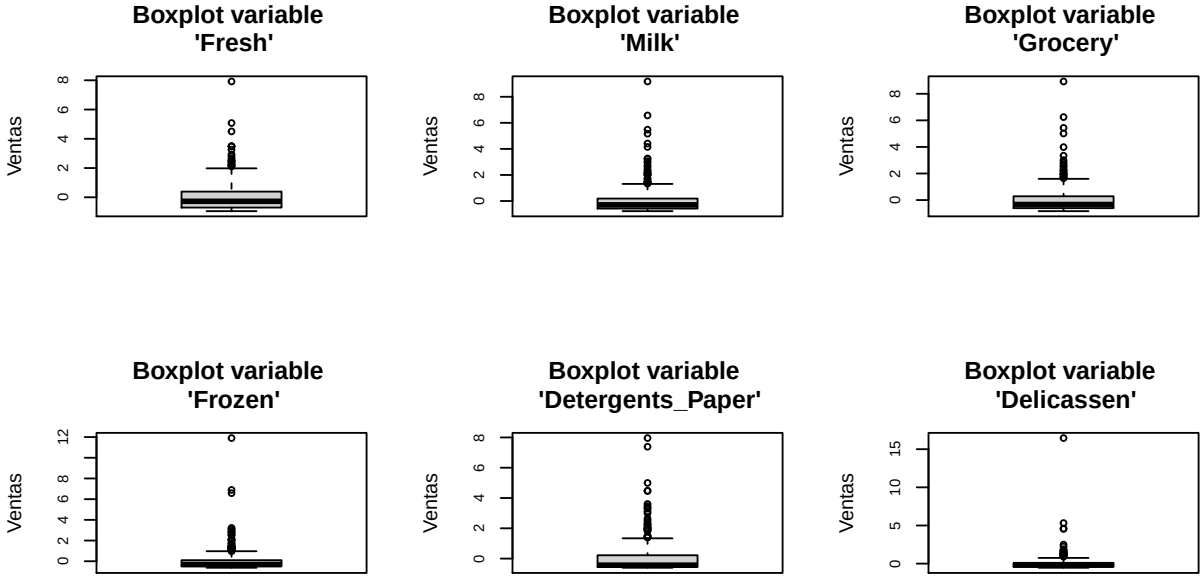
3. Normalice los datos utilizando la función **scale**. Comente sobre el posible beneficio de realizar este pre-procesamiento en análisis de conglomerados.

- *Resp: el escalado de los datos ayuda a evitar que los atributos en rangos numéricos mayores dominen a aquellos en rangos numéricos más pequeños. Sirve también para colocar en una misma escala los valores atípicos u outliers de todas las variables.*

Principales estadísticos de los datos normalizados:

```
##      Fresh              Milk           Grocery           Frozen        
##  Min.   :-0.9486   Min.   :-0.7779   Min.   :-0.8364   Min.   :-0.62763
##  1st Qu.:-0.7015   1st Qu.:-0.5776   1st Qu.:-0.6101   1st Qu.:-0.47988
##  Median :-0.2764   Median :-0.2939   Median :-0.3363   Median :-0.31844
##  Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000
##  3rd Qu.: 0.3901   3rd Qu.: 0.1889   3rd Qu.: 0.2846   3rd Qu.: 0.09935
##  Max.   : 7.9187   Max.   : 9.1732   Max.   : 8.9264   Max.   :11.90545
##  Detergents_Paper    Delicassen      
##  Min.   :-0.6037   Min.   :-0.5396  
##  1st Qu.:-0.5505   1st Qu.:-0.3960  
##  Median :-0.4331   Median :-0.1984  
##  Mean   : 0.0000   Mean   : 0.0000  
##  3rd Qu.: 0.2182   3rd Qu.: 0.1047  
##  Max.   : 7.9586   Max.   :16.4597
```

Visualización de datos atípicos (puntos más allá de los bigotes del 'boxplot'):

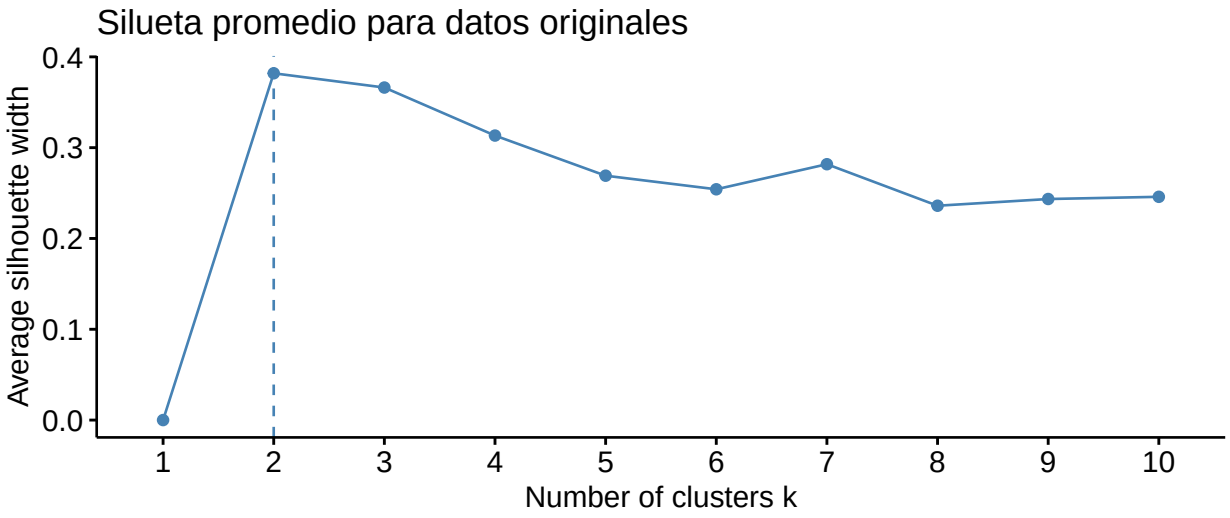


Conglomerados por k-medoids

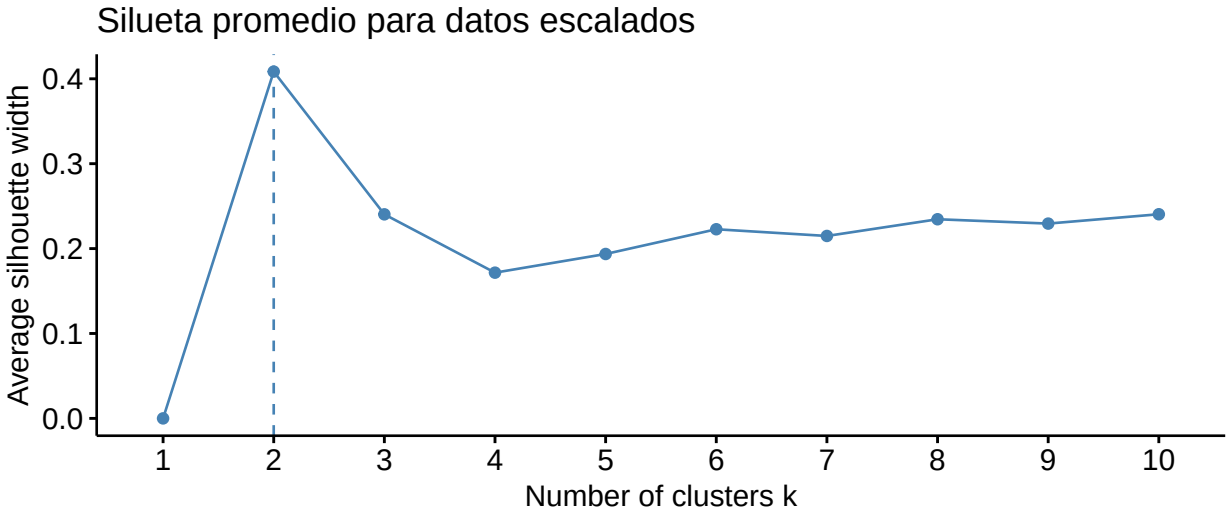
4. Haciendo uso de la función `pam()` , incluida en la librería **cluster**, construya k conglomerados utilizando el método de *k-medoids*, para valores del parámetro k entre 2 y 10, y calcule el ancho de silueta promedio para cada valor de dicho parámetro. En base al ancho de la silueta, determine el número óptimo de conglomerados para agrupar los datos. Investigue la salida de la función `pam()` para poder obtener la silueta promedio. En el siguiente enlace encontrará información relevante sobre ella: https://en.wikipedia.org/wiki/Silhouette_(clustering). Repita el procedimiento para los datos normalizados. Comente sus resultados en el foro.

- *Resp: el número óptimo de clusters es para K=2, valor para el cual se obtiene el máximo ancho de silueta [1]*

```
fviz_nbclust(wholesale_original, pam, method = "silhouette")+
  labs(title = "Silueta promedio para datos originales")
```



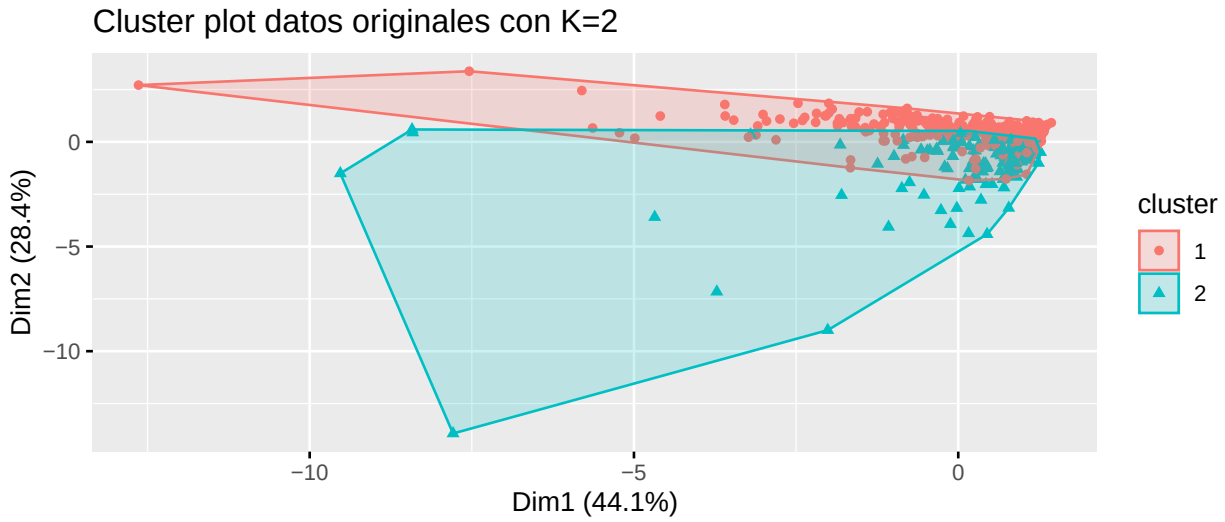

```
fviz_nbclust(wholesale_scaled, pam, method = "silhouette")+
  labs(title = "Silueta promedio para datos escalados")
```



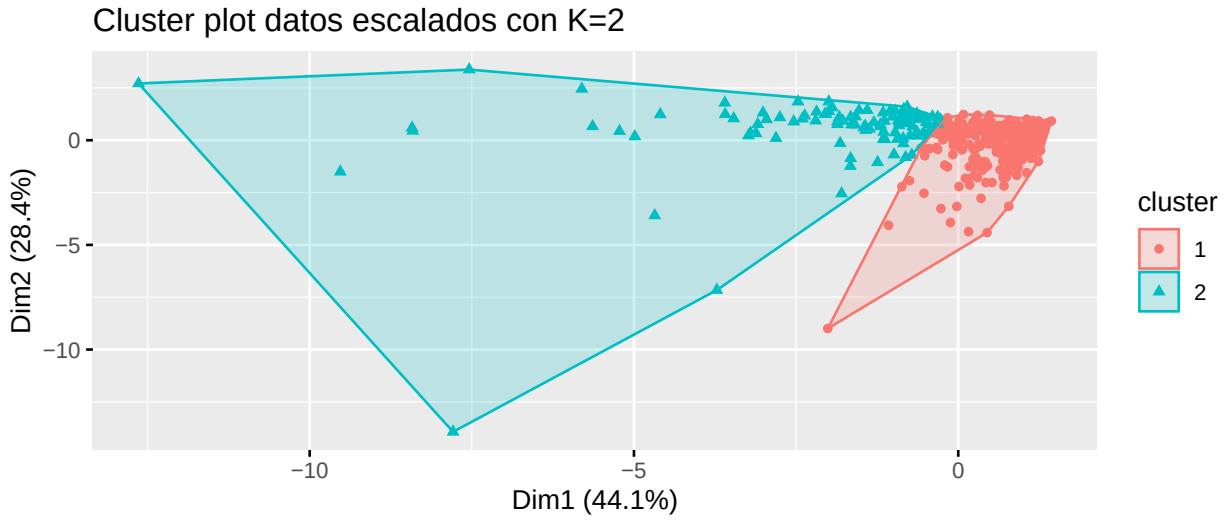
5. Agrupe los datos usando *k-medoids* con el valor de k determinado en el punto anterior, y genere una representación gráfica de los conglomerados generados utilizando la función **fviz_cluster**. Esta función reduce la dimensionalidad de los datos a dos dimensiones utilizando el algoritmo PCA, visto en la clase 1. ¿Qué diferencias puede observar entre los clusters generados por ambos conjuntos de datos? Comente sus resultados en el foro.

- *Resp: se observa una superposición de los puntos cuando estos no están escalados. Al normalizar el conjunto de datos, los conglomerados resultantes tienen muy pocos puntos de intersección, lo cual indica una mejor clusterización*

```
set.seed(1)
pam_original <- pam(wholesale_original,2)
fviz_cluster(pam_original, data = wholesale_original, geom = "point")+
  labs(title = "Cluster plot datos originales con K=2")
```




```
set.seed(1)
pam_scaled <- pam(wholesale_scaled,2)
fviz_cluster(pam_scaled, data = wholesale_scaled, geom = "point")+
  labs(title = "Cluster plot datos escalados con K=2")
```



6. Utilizando los conglomerados generados en el punto anterior: ¿Qué observaciones son utilizadas como representantes de cada grupo? Repita el análisis para los datos normalizados, considerando si las observaciones representantes se mantienen o cambian. Comente su análisis en el foro.

- *Resp: para el conjunto de datos originales se utilizan las observaciones 56 y 90, mientras que para los datos normalizados se usan las observaciones 10 y 322. Se observa que los mejores centros de cada conjunto son aquellos que fueron escalados (observaciones 10 y 322), ya que para cada variable estos puntos están más alejados entre sí respecto a las observaciones del conjunto de datos originales. Por otra parte al escalar los datos cambia el tamaño de los conglomerados, pasando de 313 y 127 observaciones (sin escalar), a 333 y 107 observaciones por conglomerado (con escalado).*

```
# Posición de los k-medoids: Datos Originales
pam_original$id.med
```

```
## [1] 56 90
```

```
# Posición de los k-medoids: Datos Normalizados
pam_scaled$id.med
```

```
## [1] 322  10
```

```
# Información de las observaciones para los datos sin escalar
wholesale_original[c(56,90),1:6]
```

```
##    Fresh Milk Grocery Frozen Detergents_Paper Delicassen
## 56  5264 3683    5005   1057             2024       1130
## 90 24904 3836    5330   3443              454       3178
```

```
# Información de las observaciones para los datos escalados
wholesale_original[c(10,322),1:6]
```

```
##     Fresh  Milk Grocery Frozen Detergents_Paper Delicassen
## 10   6006 11093   18881   1159             7425       2098
## 322  9155  1897    5167   2714              228       1113
```

```
# Cluster info de los k-medoids: Datos Originales
pam_original$clusinfo
```

```
##      size  max_diss  av_diss diameter separation
## [1,]  313 105539.77 10350.61 112373.0    1881.33
## [2,]  127  93099.48 14922.45 107343.7    1881.33
```

```
# Cluster info de los k-medoids: Datos Normalizados
pam_scaled$clusinfo
```

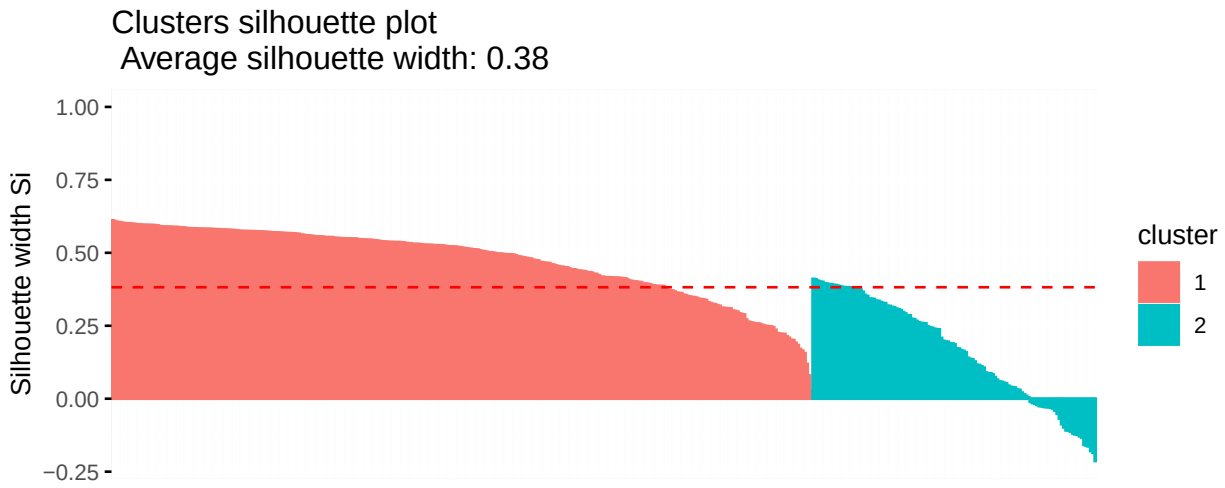
```
##      size max_diss  av_diss diameter separation
## [1,]  333 12.42667 1.259119 13.20710  0.1566052
## [2,]  107 18.58566 2.136609 21.24427  0.1566052
```

7. Cree un gráfico de silueta utilizando la función **fviz_silhouette** y discuta una interpretación para el ancho de silueta promedio total, y dentro de cada uno de los conglomerados. ¿Cómo varía la cantidad de elementos por cluster?, ¿Cómo es la calidad del agrupamiento de los datos?. Repita el análisis para los datos normalizados y comente su análisis en el foro.

- *Resp: el ancho de silueta promedio total, es mayor con los datos escalados (0.41 vs 0.38), lo que indica una mejor calidad de ajuste. Por otra parte, en el análisis dentro de cada conglomerado, en el cluster con menor cantidad de observaciones se aprecian valores negativos, que pueden significar datos muy separados o también entremezclados (tanto en el conjunto original como en el escalado).*

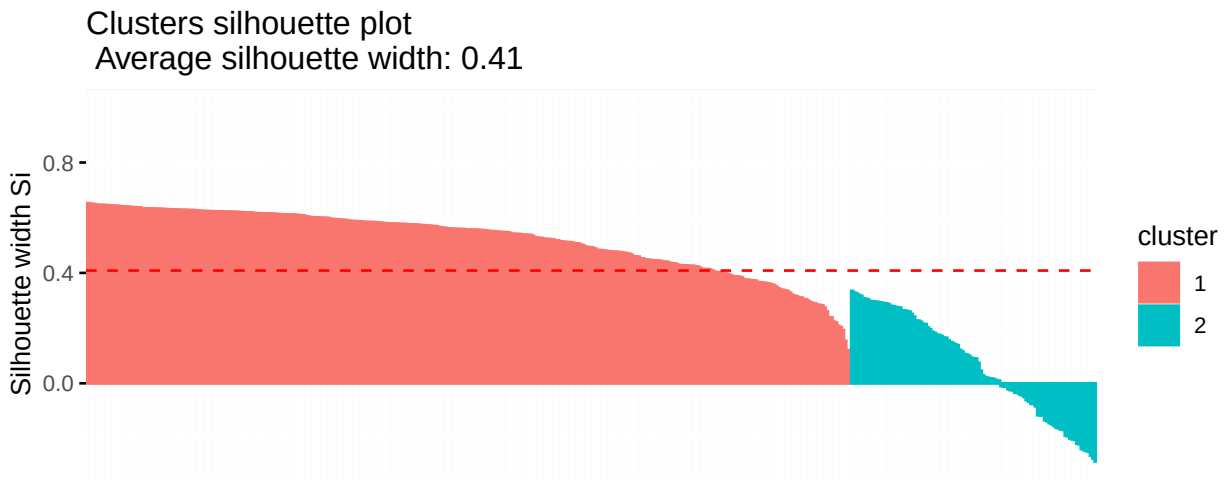
```
# Análisis de silueta: datos originales
sil_original <- silhouette(pam_original)
fviz_silhouette(sil_original)
```

```
##   cluster size ave.sil.width
## 1       1  313          0.47
## 2       2  127          0.16
```



```r
# Análisis de silueta: datos escalados
sil_scaled <- silhouette(pam_scaled)
fviz_silhouette(sil_scaled)
```

```
##   cluster size ave.sil.width
## 1       1  333          0.52
## 2       2  107          0.07
```



* *Como conclusión del método k-medoids, la mejora alcanzada al normalizar los datos es pequeña y, según la visualización de las dos primeras componentes principales, para K=2 se obtiene un cluster que es aproximadamente 3 veces más grande que el segundo, con mayor densidad de puntos y poca dispersión de datos. Al contrario, el segundo cluster posee menor densidad de elementos y una mayor dispersión de datos, con serias sospechas de presencia de valores atípicos, dada la distancia en que se encuentran los puntos más alejados del medoide.*

Conglomerados Jerárquicos

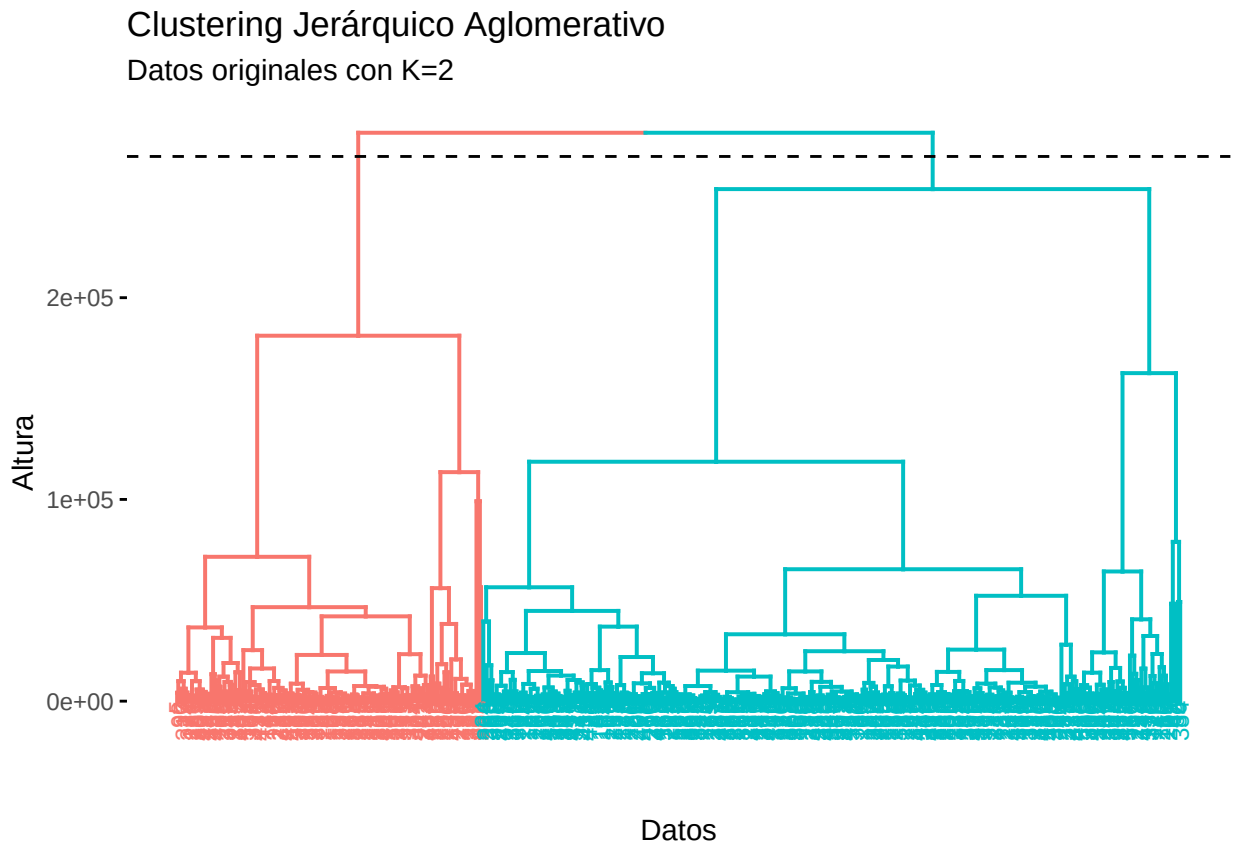
8. Responda en el foro: ¿Qué utilidad podría tener un conglomerado jerárquico en los datos disponibles?.

- *Resp: una utilidad podría ser identificar a los clientes con patrones de compra muy distintos al de la gran mayoría*

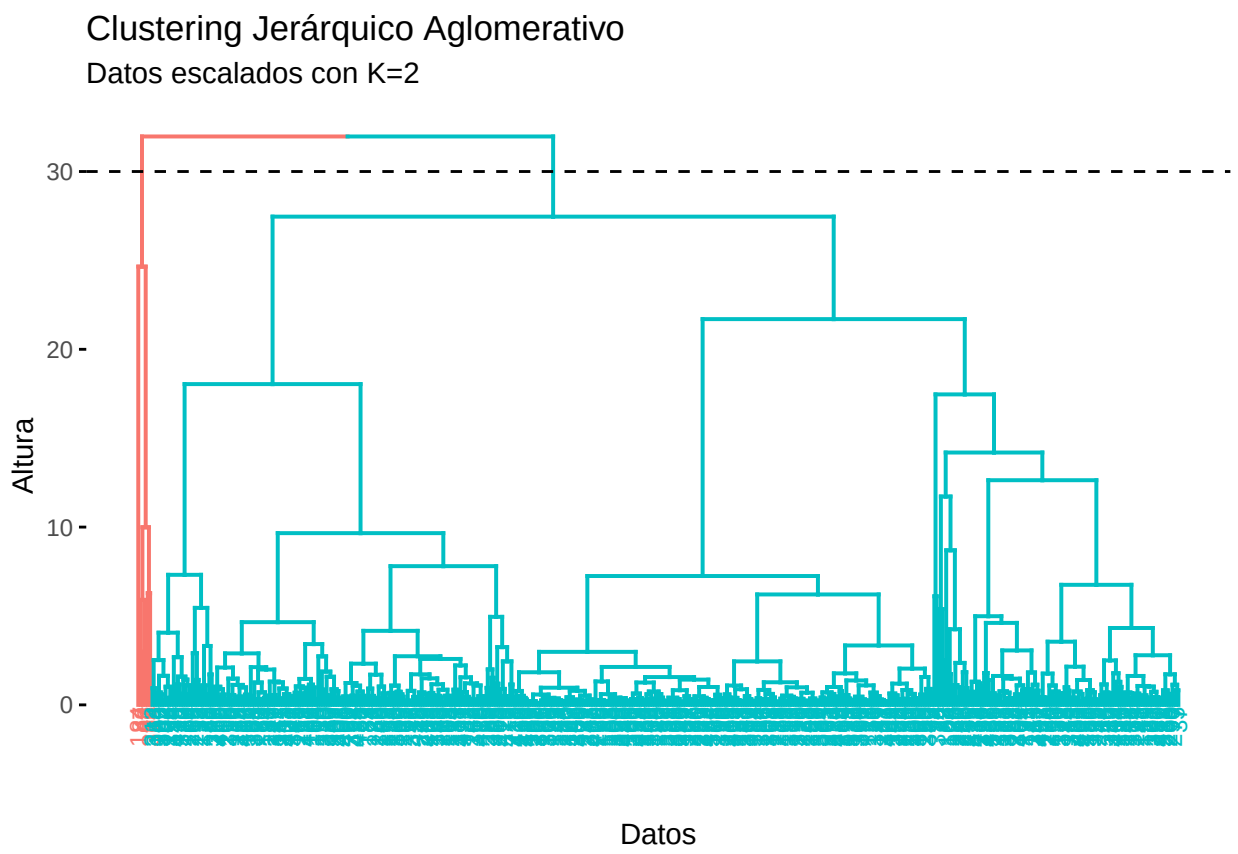
9. Utilizando la función `hclust()` realice un análisis de conglomerados jerárquico aglomerativo. Grafique el dendograma obtenido para cada conjunto de datos y comente sus resultados en el foro. En el siguiente enlace encontrará información sobre este tipo de gráfico: https://en.wikipedia.org/wiki/Dendrogram.

- *Resp: llama la atención la diferencia de tamaño de los clusters entre los dendogramas con datos originales vs los normalizados para un mismo valor de K (en este caso igual a 2). En el dendograma con datos escalados se aprecia un grupo reducido de observaciones con pocos niveles de altura respecto al otro cluster, que agrupa a la mayoría de los datos y que contiene mayores niveles de sub-división.*

```
# Conglomerado Jerárquico Aglomerativo: datos originales
clustering_original <- hclust(dist(wholesale_original), method='ward.D2')
fviz_dend(x = clustering_original, k = 2, cex = 0.6) +
  geom_hline(yintercept = 270000, linetype = "dashed") +
  labs(title = "Clustering Jerárquico Aglomerativo",
       subtitle = "Datos originales con K=2", y="Altura", x="Datos")
```



```r
# Conglomerado Jerárquico Aglomerativo: datos escalados
clustering_scaled <- hclust(dist(wholesale_scaled), method='ward.D2')
fviz_dend(x = clustering_scaled, k = 2, cex = 0.6) +
  geom_hline(yintercept = 30, linetype = "dashed") +
  labs(title = "Clustering Jerárquico Aglomerativo",
       subtitle = "Datos escalados con K=2", y="Altura", x="Datos")
```



10. Si se quedase con la misma cantidad de grupos que en la parte 1, ¿Se parece este agrupamiento al realizado con el método de *k-medoids*? Justifique su respuesta en el foro. Para obtener la agrupación en base a un número determinado de conglomerados puede utilizar la función `cutree()`. Repita el análisis para los datos normalizados y comente en el foro.

- *Resp: el agrupamiento entre k-medoids y el modelo jerárquico sí es parecido para los datos originales. Sin embargo, para el conjunto de datos normalizados es muy distinto, ya que el cluster más pequeño tiene 6 elementos en el modelo jerárquico, mientras que en k-medoids tenía 107 observaciones.*

```
# Reducción: datos originales
clustering_info_original <- cutree(clustering_original, k=2)
# Informacion del cluster cortado: datos originales
table(clustering_info_original)
```

```
## clustering_info_original
##   1   2
## 306 134
```

```
# Reducción: datos normalizados
clustering_info_scaled <- cutree(clustering_scaled, k=2)
# Informacion del cluster cortado: datos normalizados
table(clustering_info_scaled)
```

```
## clustering_info_scaled
##   1   2
## 434   6
```

- *Resp: el resultado anterior, nos lleva a realizar un análisis de segundo orden, con la finalidad de poder identificar los principales estadísticos de los clusters 1 y 2.*

```
wholesale_original$cluster <- as.matrix(factor(clustering_info_scaled))
# Cluster 1
summary(wholesale_original[wholesale_original$cluster==1,])
```

```
##      Fresh             Milk          Grocery          Frozen
##  Min.   :     3   Min.   :   55   Min.   :    3   Min.   :   25.0
##  1st Qu.:  3098   1st Qu.: 1516   1st Qu.: 2146   1st Qu.:  738.8
##  Median :  8258   Median : 3608   Median : 4725   Median : 1526.0
##  Mean   : 11786   Mean   : 5274   Mean   : 7307   Mean   : 2999.8
##  3rd Qu.: 16725   3rd Qu.: 7092   3rd Qu.:10391   3rd Qu.: 3543.5
##  Max.   :112151   Max.   :36423   Max.   :45828   Max.   :60869.0
##  Detergents_Paper    Delicassen          cluster.V1
##  Min.   :    3.0   Min.   :    3.0   Length:434
##  1st Qu.:  256.2   1st Qu.:  405.2   Class :character
##  Median :  811.0   Median :  960.5   Mode  :character
##  Mean   : 2575.5   Mean   : 1404.3
##  3rd Qu.: 3866.2   3rd Qu.: 1783.0
##  Max.   :24231.0   Max.   :16523.0
```

```
# Cluster 2
summary(wholesale_original[wholesale_original$cluster==2,])
```

```
##      Fresh            Milk          Grocery          Frozen
##  Min.   : 8565   Min.   : 4980   Min.   :20170   Min.   :  131.0
##  1st Qu.:17819   1st Qu.:39764   1st Qu.:37978   1st Qu.:  996.8
##  Median :29434   Median :45074   Median :57585   Median : 2140.0
##  Mean   :27477   Mean   :43542   Mean   :54589   Mean   : 8285.7
##  3rd Qu.:36621   3rd Qu.:52244   3rd Qu.:65373   3rd Qu.: 6650.0
##  Max.   :44466   Max.   :73498   Max.   :92780   Max.   :36534.0
##  Detergents_Paper   Delicassen        cluster.V1
##  Min.   :  239    Min.   :  903   Length:6
##  1st Qu.:21095    1st Qu.: 1416   Class :character
##  Median :25436    Median : 2480   Mode  :character
##  Mean   :25018    Mean   :10248
##  3rd Qu.:35252    3rd Qu.: 5585
##  Max.   :40827    Max.   :47943
```

- *Resp: podemos concluir que el cluster 2, conformado por 6 clientes, es un conglomerado con una participación sobre el total de ventas cercano al 7%. Con estos antecedentes, se podría diseñar una estrategia para la fidelización de estos clientes estableciendo, por ejemplo, un contrato de ventas a largo plazo para grandes volúmenes de productos a un precio competitivo.*

```
##           Category Porcentaje_sobre_total_ventas
## 1 Detergents_Paper                          11.8
## 2             Milk                          10.2
## 3          Grocery                           9.4
## 4       Delicassen                           9.2
## 5           Frozen                           3.7
## 6            Fresh                           3.1
```

11. Calcule el número de observaciones para cada uno de los k grupos, en base al conglomerado jerárquico aglomerativo obtenido en el punto anterior, con k entre 2 y 10. Repita el análisis para los datos normalizados. Comente los resultados en el foro.

```
# Número de observaciones en cada cluster para los datos originales
# Con valores de K entre 2 y 10
counts_original
```

```
## $`2`
##
##   1   2
## 306 134
##
## $`3`
##
##   1   2   3
## 261 134  45
##
## $`4`
##
##   1   2   3   4
## 261 111  45  23
##
## $`5`
##
##   1   2   3   4   5
## 261 111  40  23   5
##
## $`6`
##
##   1   2   3   4   5   6
##  86 175 111  40  23   5
##
## $`7`
##
##   1   2   3   4   5   6   7
##  86 175 111  40  20   5   3
##
## $`8`
##
##   1   2   3   4   5   6   7   8
##  86 175 111  40  20   5   1   2
##
## $`9`
##
##   1   2   3   4   5   6   7   8   9
##  86 175 111  40  20   3   2   1   2
##
## $`10`
##
##   1   2   3   4   5   6   7   8   9  10
##  86 175  29  40  82  20   3   2   1   2
```

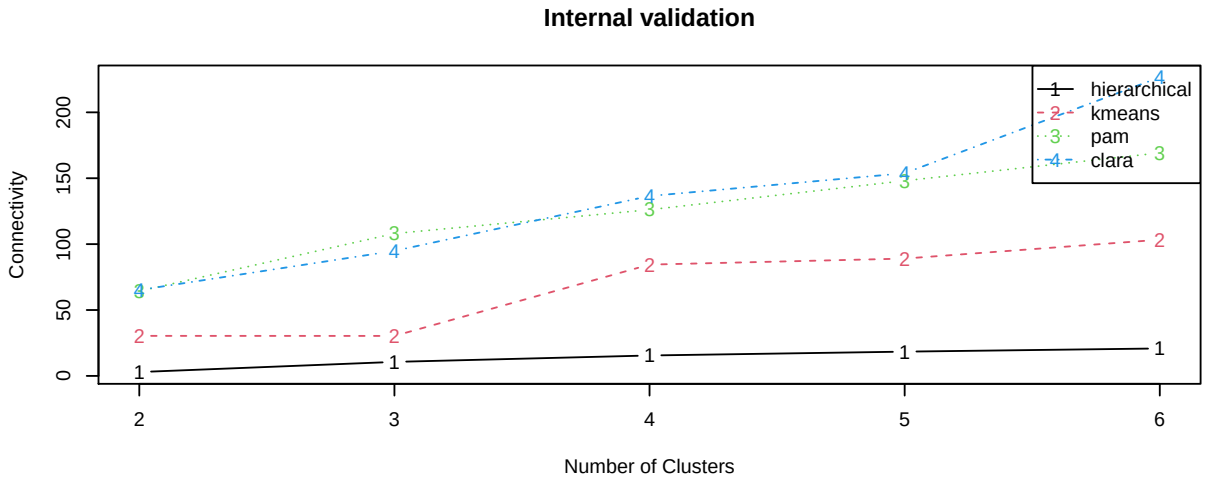
```r
# Número de observaciones en cada cluster para los datos normalizados
# Con valores de K entre 2 y 10
counts_scaled
```

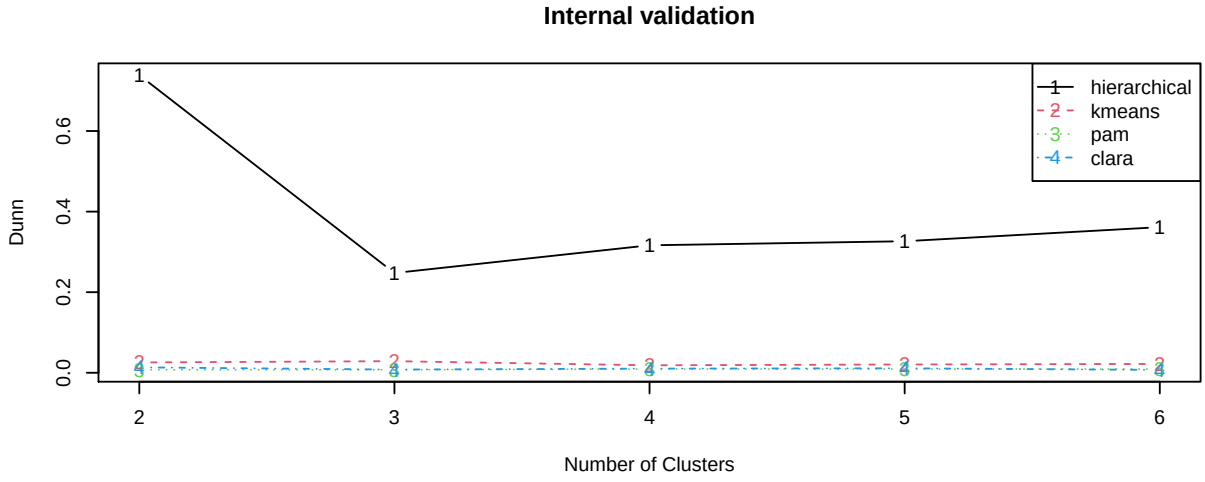
```
## $`2`
##
##   1   2
## 434   6
##
## $`3`
##
##   1   2   3
## 153 281   6
##
## $`4`
##
##   1   2   3   4
## 153 281   5   1
##
## $`5`
##
##   1   2   3   4   5
## 153 177 104   5   1
##
## $`6`
##
##   1   2   3   4   5   6
## 127 177 104  26   5   1
##
## $`7`
##
##   1   2   3   4   5   6   7
## 127 177 102  26   5   2   1
##
## $`8`
##
##   1   2   3   4   5   6   7   8
## 127 177  89  13  26   5   2   1
##
## $`9`
##
##   1   2   3   4   5   6   7   8   9
## 127 177  59  30  13  26   5   2   1
##
## $`10`
##
##   1   2   3   4   5   6   7   8   9  10
## 127 177  59  30   3  26  10   5   2   1
```

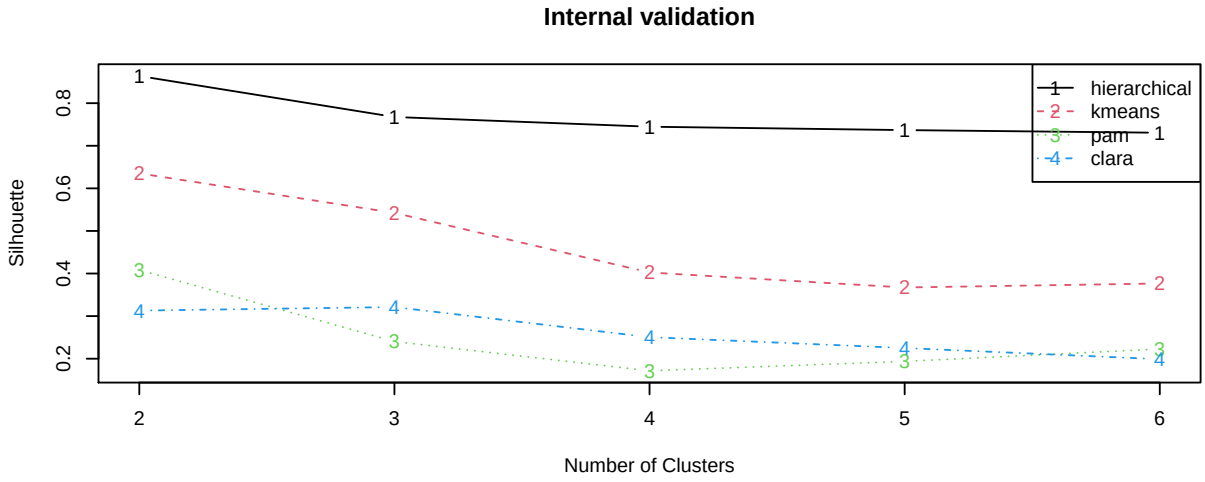
- Se evidencia que en los datos normalizados, el segundo cluster en K=2 contiene 6 observaciones que corresponden a valores atípicos. Al aumentar los valores de K, este conglomerado se sub-dividió una sola vez, lo que indica lo 'raro' de este conjunto de clientes.
- En resumen, el pre-procesamiento de datos es necesario para una mejor interpretación de las observaciones.
- El modelo de agrupación jerárquico aglomerativo nos permitió identificar los valores atípicos que explicaban un 7% del total de las ventas (6 datos de 440). La gran desventaja de este modelo es el costo computacional, ya que utiliza muchos recursos cuando los set de datos son muy grandes.
- Con k-medoids se gana en velocidad, pero se pierde en asertividad.
- Recomiendo siempre comparar los algoritmos de agrupación para trabajar con el modelo más robusto, siempre y cuando el costo computacional no sea elevado.

Comparación de modelos:

```
##
## Clustering Methods:
##  hierarchical kmeans pam clara
##
## Cluster sizes:
##  2 3 4 5 6
##
## Validation Measures:
##                                  2        3        4        5        6
##
## hierarchical Connectivity    2.9290  10.6298  15.4877  18.4167  20.8274
##              Dunn            0.7387   0.2476   0.3164   0.3264   0.3614
##              Silhouette      0.8638   0.7676   0.7447   0.7368   0.7306
## kmeans       Connectivity   30.3750  30.3452  84.4512  89.0345 103.3405
##              Dunn            0.0257   0.0288   0.0185   0.0206   0.0219
##              Silhouette      0.6352   0.5424   0.4025   0.3672   0.3766
## pam          Connectivity   64.2956 108.0433 126.0992 148.0893 169.2020
##              Dunn            0.0074   0.0074   0.0093   0.0093   0.0093
##              Silhouette      0.4085   0.2403   0.1716   0.1936   0.2228
## clara        Connectivity   65.1706  94.8671 136.4944 153.8063 226.6075
##              Dunn            0.0134   0.0079   0.0103   0.0111   0.0071
##              Silhouette      0.3129   0.3213   0.2505   0.2251   0.1989
##
## Optimal Scores:
##
##              Score  Method       Clusters
## Connectivity 2.9290 hierarchical 2
## Dunn         0.7387 hierarchical 2
## Silhouette   0.8638 hierarchical 2
```


Internal validation
Connectivity
0
50
100
150
200
2
3
4
5
6
Number of Clusters
hierarchical
kmeans
pam
clara


Internal validation
Dunn
0.0
0.2
0.4
0.6
2
3
4
5
6
Number of Clusters
hierarchical
kmeans
pam
clara


Internal validation
Silhouette
0.2
0.4
0.6
0.8
2
3
4
5
6
Number of Clusters
hierarchical
kmeans
pam
clara

En donde:

- **'hierarchical'** corresponde al método de agrupación por jerarquía.
- **'kmeans'** corresponde al método de agrupación K-means.
- **'pam'** corresponde al método de agrupación K-medoides.
- **'clara'** corresponde al método de agrupación que combina la idea de K-medoides con el 'resampling' (remuestreo) para que pueda aplicarse a grandes volúmenes de datos.
- **'Connectivity', 'Dunn' y 'Silhouette'** son medidas de validación interna [2]

Según lo anterior, el mejor método de agrupación para los datos analizados sería el jerárquico con k=2, ya que tiene el valor más bajo para 'Connectivity', y los valores más altos para 'Dunn' y 'Silhouette' [2].

[1][Clustering, https://rpubs.com/Joaquin_AR/310338, Número óptimo de clusters]

[2][Documentation 'clValid' Package, https://www.rdocumentation.org/packages/clValid/versions/0.6-9/topics/clValid, Internal measures]