Actividad 1.9 Conglomerados jerárquicos

Franco Mendoza Muraira A01383399

2023-11-18

Problema 1

```
## Warning: package 'ggplot2' was built under R version 4.2.3
```

```
## Warning: package 'factoextra' was built under R version 4.2.3
```

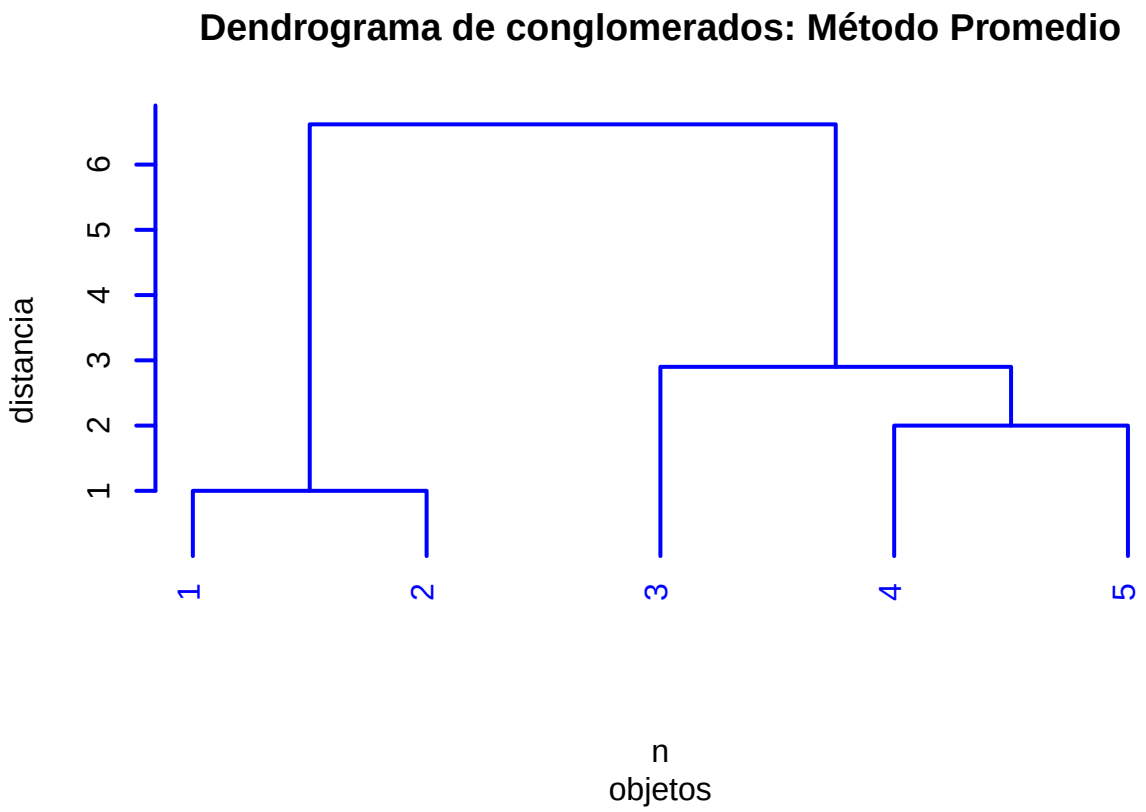
1. Como nos dan sólo para la parte inferior de la matriz (simétrica), puede introducir los datos de varias formas. Por ejemplo, introduzca la matriz en R con 0 en los espacios en blanco.

```
Mpre = matrix(c(0,1,5,8.5,7.2, 0,0, 4.5,7.8,6.7,0,0,0,3.6,2.2,0,0,0,0,2,0,0,0,0 , 0), ncol = 5)
M = Mpre + t(Mpre)
M
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]  0.0  1.0  5.0  8.5  7.2
## [2,]  1.0  0.0  4.5  7.8  6.7
## [3,]  5.0  4.5  0.0  3.6  2.2
## [4,]  8.5  7.8  3.6  0.0  2.0
## [5,]  7.2  6.7  2.2  2.0  0.0
```

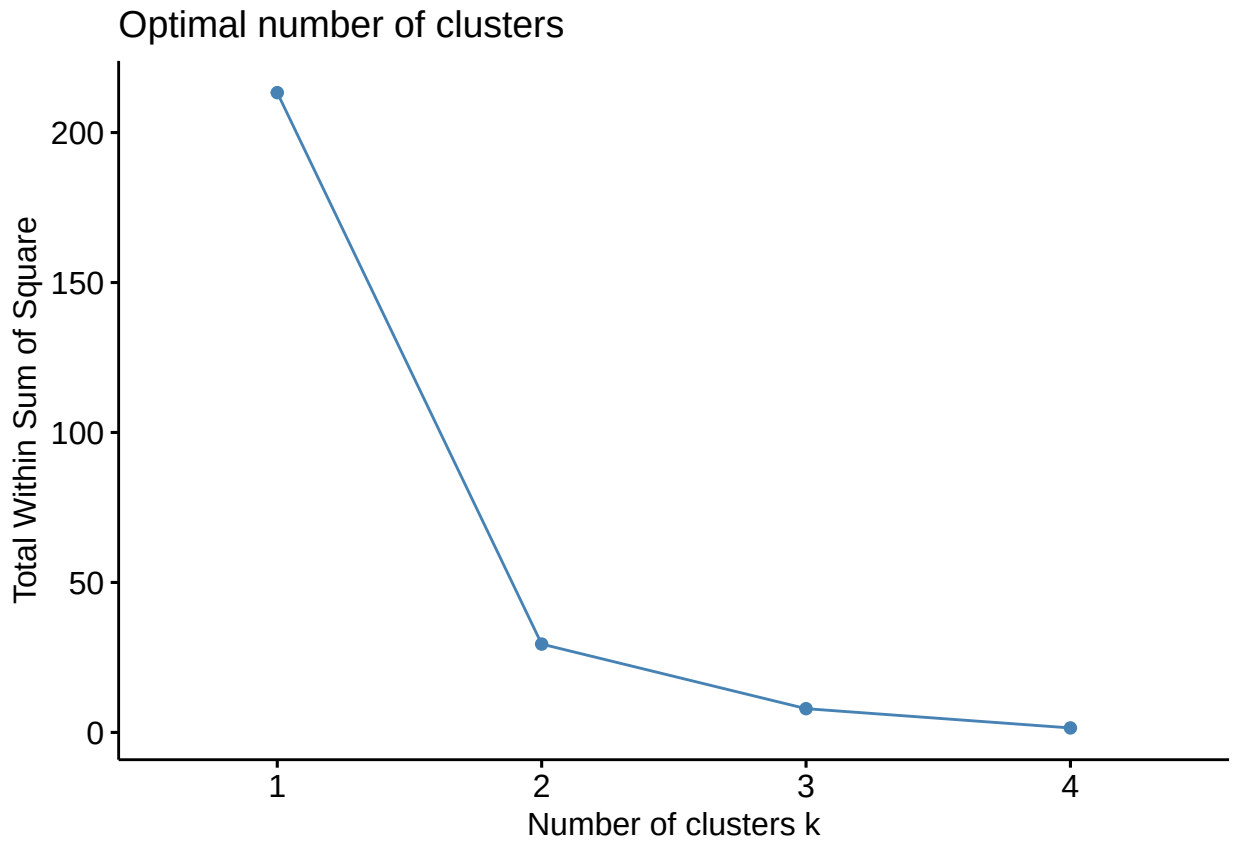
2. Apliquen las funciones as.dist, hclust y plot para explorar los dendrogramas (que se diferencian por las distancias en el eje Y)

```
d = as.dist(M)
J = hclust(d, method = "average")
plot(J, hang = -1, lwd = 2, col = "blue", main = "Dendrograma de conglomerados: Método Promedio", sub =
```



3. Para interpretar sobre el número óptimo de clusters puede ayudar la función fviz_nbclust

```
fviz_nbclust(M, FUNcluster = kmeans, method = "wss", k.max = 4)
```



En esta grafica que nos muestra la suma total cuadrada con cada cluster, podemos ver que el número de clusters que sería mejor podría ser el de 3, ya que nos proporciona buenos cambios en la gráfica, y agregar clusters daría cambios mínimos e innecesarios.

4. Hallar la matriz de ultra distancias (las distancias de los objetos según el dendograma)

```
heights <- J$height

n <- length(heights) + 1
ultra <- matrix(0, nrow = n, ncol = n)

for (i in 1:n) {
  for (j in 1:n) {
    if (i == j) {
      ultra[i, j] <- 0
    } else if (i < j) {
      ultra[i, j] <- heights[i] + heights[j - 1]
      ultra[j, i] <- ultra[i, j]
    }
  }
}
cat("Matriz de ultra distancias:\n\n")
```

```
## Matriz de ultra distancias:
```

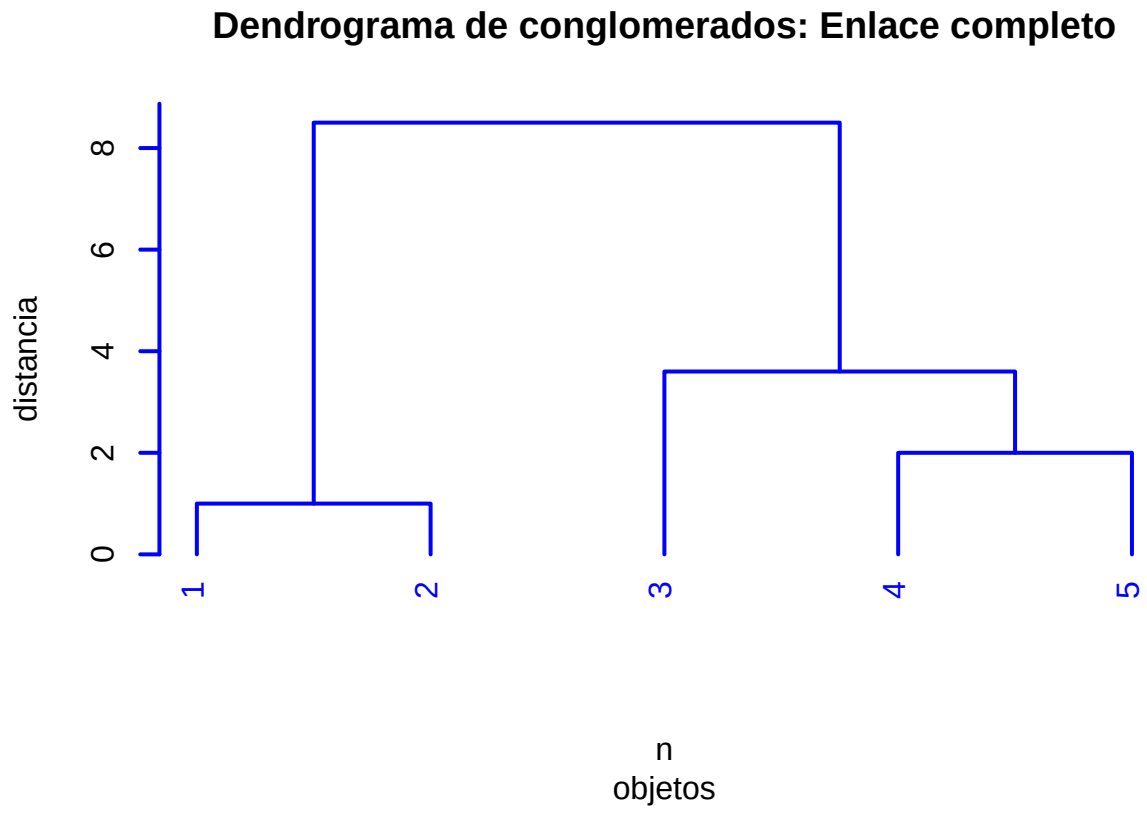
```
ultra
```

```
##          [,1]     [,2]     [,3]     [,4]      [,5]
## [1,] 0.000000 2.000000 3.000000  3.90000  7.616667
## [2,] 2.000000 0.000000 4.000000  4.90000  8.616667
## [3,] 3.000000 4.000000 0.000000  5.80000  9.516667
## [4,] 3.900000 4.900000 5.800000  0.00000 13.233333
## [5,] 7.616667 8.616667 9.516667 13.23333  0.000000
```

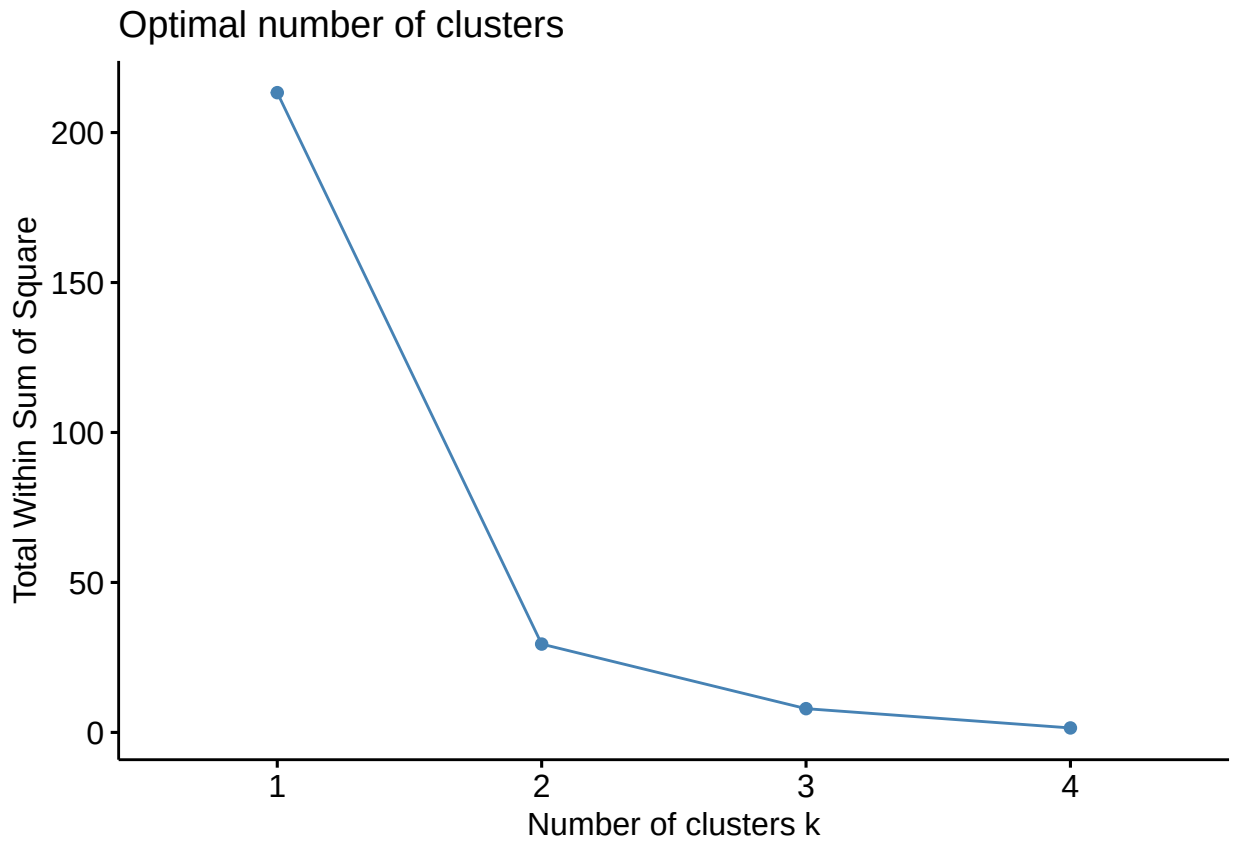
5. Elige otro metodo de agrupación y elabora el dendograma, ¿qué disferencias encuentras entre ambos?

Usamos el método de enlace completo, busca clusters similares.

```
d = as.dist(M)
J = hclust(d, method = "complete")
plot(J, hang = -1, lwd = 2, col = "blue", main = "Dendrograma de conglomerados: Enlace completo", sub =
```




```
fviz_nbclust(M, FUNcluster = kmeans, method = "wss", k.max = 4)
```



Con este método de enlace completo, podemos ver que el número óptimo ce clusters seguiría siendo de 3, ya que más que eso hay cambio casi nulo, y con 2 clusters se puede ver que tiene mucho por ganar todavía en la gráfica.

```
heights <- J$height

n <- length(heights) + 1
ultra <- matrix(0, nrow = n, ncol = n)

for (i in 1:n) {
  for (j in 1:n) {
    if (i == j) {
      ultra[i, j] <- 0
    } else if (i < j) {
      ultra[i, j] <- heights[i] + heights[j - 1]
      ultra[j, i] <- ultra[i, j]
    }
  }
}
cat("Matriz de ultra distancias:\n\n")
```

```
## Matriz de ultra distancias:
```

```
ultra
```

```
##      [,1] [,2] [,3] [,4] [,5]
```

```
## [1,]  0.0  2.0  3.0  4.6  9.5
## [2,]  2.0  0.0  4.0  5.6 10.5
## [3,]  3.0  4.0  0.0  7.2 12.1
## [4,]  4.6  5.6  7.2  0.0 17.0
## [5,]  9.5 10.5 12.1 17.0  0.0
```

También podemos ver que aunque las gráficas sean muy similares, la matriz de ultra distancias aún tiene cambios.

Diferencias en los métodos.

- En el primer método con enlace promedio, las distancias ultramétricas tienden a ser más pequeñas. Por ejemplo, las distancias entre los elementos 1, 2, y 3 son 2.0, 3.0, y 3.9 respectivamente. Estos valores son menores que las distancias correspondientes en el método de enlace completo.
- En el segundo método con enlace completo, las distancias ultramétricas tienden a ser mayores en comparación con el método de enlace promedio. Aquí, las distancias entre los elementos tienden a ser más grandes. Por ejemplo, las distancias entre 1, 2, y 3 son 2.0, 4.0, y 5.8 respectivamente, valores más grandes que en el primer método.

Estas diferencias se deben a cómo se calcula la distancia entre los clusters en cada método de agrupación jerárquica. El método de enlace promedio calcula las distancias entre clusters considerando el promedio de todas las distancias entre pares de puntos en los clusters individuales. Mientras que el método de enlace completo considera la máxima distancia entre pares de puntos en los clusters individuales.

Podemos ver que aún con estas diferencias obtuvimos resultados muy similares, y llegamos a las mismas conclusiones, pero en otras aplicaciones, o con datos más grandes esto puede cambiar dependiendo del método por lo que es bueno identificar el uso que se le va a dar a los datos para elegir correctamente dependiendo de lo que se busque.