---

title: “Hierarchical Clustering” author: “luis manuel” date: “18/12/2020” output: pdf_document —

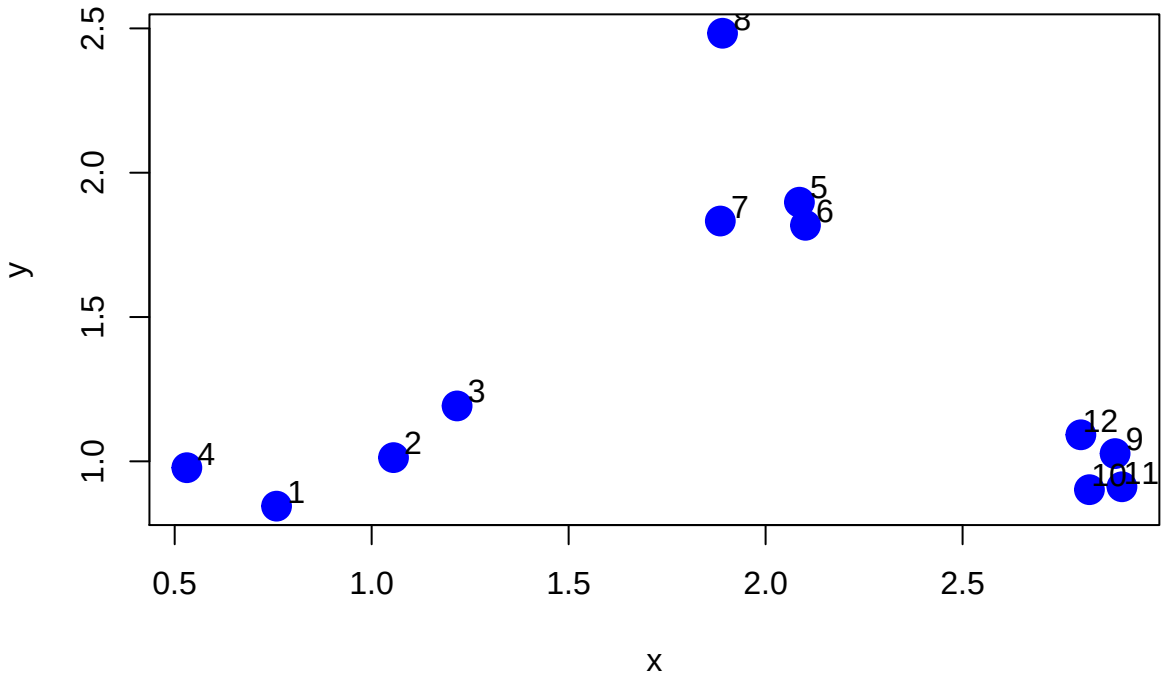
Hierarchical Clustering

En esta lección, aprenderemos sobre la agrupación jerárquica, una forma sencilla de examinar y mostrar rápidamente datos multidimensionales. Esta técnica suele ser más útil en las primeras etapas del análisis cuando intenta comprender los datos, por ejemplo, encontrar algún patrón o relación entre diferentes factores o variables. Como sugiere el nombre, la agrupación jerárquica crea una jerarquía de agrupaciones.

La agrupación en clústeres organiza los puntos de datos que están cerca en grupos. Preguntas tan obvias son “¿Cómo definimos cerca?”, “¿Cómo agrupamos las cosas?” Y “¿Cómo interpretamos la agrupación?” El análisis de conglomerados es un tema muy importante en el análisis de datos.

Para darte una idea de lo que estamos hablando, considera estos puntos aleatorios que generamos. Los usaremos para demostrar la agrupación jerárquica en esta lección. Haremos esto en varios pasos, pero primero tenemos que aclarar nuestros términos y conceptos.

```
set.seed(1234)
 x <- rnorm(12, rep(1:3, each = 4), 0.2)
 y <- rnorm(12, rep(c(1, 2, 1), each = 4), 0.2)
plot(x, y, col = "blue", pch = 19, cex = 2)
text(x + 0.05, y + 0.05, labels = as.character(1:12))
```



La agrupación en clústeres organiza los puntos de datos que están cerca en grupos. Preguntas tan obvias son "¿Cómo definimos cerca?", "¿Cómo agrupamos las cosas?" Y "¿Cómo interpretamos la agrupación?" El análisis de conglomerados es un tema muy importante en el análisis de datos.

Para darte una idea de lo que estamos hablando, considera estos puntos aleatorios que generamos. Los usaremos para demostrar la agrupación jerárquica en esta lección. Haremos esto en varios pasos, pero primero tenemos que aclarar nuestros términos y conceptos.

La agrupación jerárquica es un enfoque aglomerativo o de abajo hacia arriba. De Wikipedia (http://en.wikipedia.org/wiki/Hierarchical_clustering), aprendemos que en este método, "cada observación comienza en su propio grupo, y los pares de grupos se fusionan a medida que uno asciende en la jerarquía". Esto significa que encontraremos los dos puntos más cercanos y los pondremos juntos en un grupo, luego encontraremos el siguiente par más cercano en la imagen actualizada, y así sucesivamente. Repetiremos este proceso hasta llegar a un punto de parada razonable.

Tenga en cuenta la palabra "razonable". Hay mucha flexibilidad en este campo y la forma de realizar su análisis depende de su problema. Una vez más, nos dice Wikipedia, "uno puede decidir dejar de agrupar cuando los grupos están demasiado separados para fusionarse (criterio de distancia) o cuando hay un número suficientemente pequeño de grupos (criterio de número)".

Primero, ¿cómo definimos cerca? Este es el paso más importante y hay varias posibilidades dependiendo de las preguntas que esté tratando de responder y los datos que tenga. La distancia o la similitud suelen ser las métricas utilizadas.

Es bastante obvio que de las 4 opciones, el par 5 y 6 fueron los más cercanos. Sin embargo, hay varias formas de medir la distancia o la similitud. La distancia euclidiana y la similitud de correlación son medidas continuas, mientras que la distancia de Manhattan es una medida binaria. En esta lección discutiremos brevemente el primero y el último de ellos. Es importante que utilice una medida de distancia que se adapte a su problema.

La distancia euclidiana es lo que aprendiste en álgebra de secundaria. Dados dos puntos en un plano, (x1, y1) y (x2, y2), la distancia euclidiana es la raíz cuadrada de las sumas de los cuadrados de las distancias entre las dos coordenadas x (x1-x2) y las dos coordenadas y -coordenadas (y1-y2). Probablemente reconozca esto como una aplicación del teorema de Pitágoras que da la longitud de la hipotenusa de un triángulo rectángulo.

No debería ser difícil creer que esto se generaliza a más de dos dimensiones como se muestra en la fórmula en la parte inferior de la imagen que se muestra aquí.

La distancia euclidiana es la distancia "en línea recta". Sin embargo, muchas aplicaciones no pueden utilizar la distancia de vuelo de cuervo de manera realista. Los coches, por ejemplo, tienen que seguir las carreteras.

En este caso, podemos usar Manhattan o la distancia de cuadra de la ciudad (también conocida como métrica de taxi). Esta imagen, copiada de http://en.wikipedia.org/wiki/Taxicab_geometry, muestra lo que esto significa.

Desea viajar desde el punto de la parte inferior izquierda hasta el de la parte superior derecha. La distancia más corta es la euclidiana (la línea verde), pero está limitado a la cuadrícula, por lo que debe seguir un camino similar a los que se muestran en rojo, azul o amarillo. Todos tienen la misma longitud (12), que es el número de pequeños segmentos grises cubiertos por sus trayectorias.

Más formalmente, la distancia de Manhattan es la suma de los valores absolutos de las distancias entre cada coordenada, por lo que la distancia entre los puntos (x1, y1) y (x2, y2) es | x1-x2 | + | y1-y2 |. Al igual que con la distancia euclidiana, también se generaliza a más de 2 dimensiones.

Ahora volveremos a nuestros puntos aleatorios. Es posible que haya notado que estos puntos realmente no se ven colocados al azar y, de hecho, no lo están. En realidad, se generaron como 3 grupos distintos. Hemos puesto las coordenadas de estos puntos en un marco de datos para usted, llamado dataFrame.

Usaremos este marco de datos para demostrar una técnica aglomerativa (ascendente) de agrupamiento jerárquico y crear un dendrograma. Esta es una imagen abstracta (o gráfico) que muestra cómo los 12

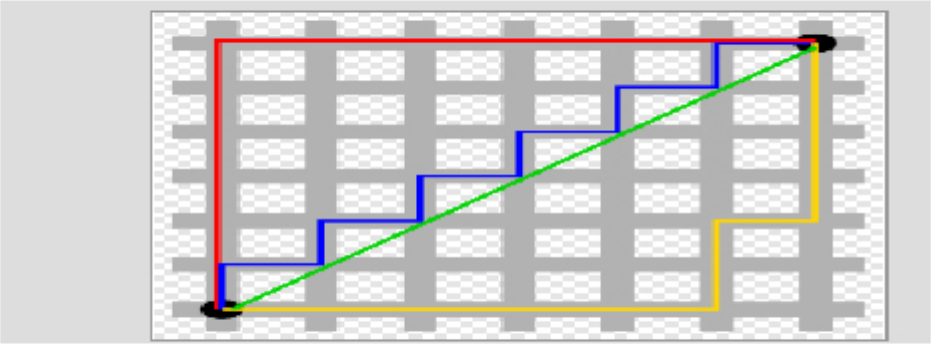


Figure 1: A caption

puntos de nuestro conjunto de datos se agrupan. Dos grupos (inicialmente, estos son puntos) que están cerca están conectados con una línea. Usaremos la distancia euclidiana como nuestra métrica de cercanía. Ahora volveremos a nuestros puntos aleatorios. Es posible que haya notado que estos puntos realmente no se ven colocados al azar y, de hecho, no lo están. En realidad, se generaron como 3 grupos distintos. Hemos puesto las coordenadas de estos puntos en un marco de datos para usted, llamado dataFrame.

Usaremos este marco de datos para demostrar una técnica aglomerativa (ascendente) de agrupamiento jerárquico y crear un dendrograma. Esta es una imagen abstracta (o gráfico) que muestra cómo los 12 puntos de nuestro conjunto de datos se agrupan. Dos grupos (inicialmente, estos son puntos) que están cerca están conectados con una línea. Usaremos la distancia euclidiana como nuestra métrica de cercanía.

Ejecute el comando R dist con el argumento dataFrame para calcular las distancias entre todos los pares de estos puntos. Por defecto dist usa la distancia euclidiana como métrica, pero hay otras métricas disponibles, como Manhattan. Solo use el valor predeterminado.

```
dataFrame<-cbind(x,y)
distxy<-dist(dataFrame)
```

Verá que la salida es una matriz triangular inferior con filas numeradas del 2 al 12 y columnas numeradas del 1 al 11. La entrada (i, j) indica la distancia entre los puntos i y j. Claramente, solo necesita una matriz triangular inferior, ya que la distancia entre los puntos i y j es igual a la que hay entre j e i.

Entonces 0.0815 (las unidades no están especificadas) entre los puntos 5 y 6 es la distancia más corta. Podemos poner estos puntos en un solo grupo y buscar otro par de puntos cercanos. Entonces 10 y 11 son otro par de puntos que estarían en un segundo grupo. Empezaremos a crear nuestro dendrograma ahora. Aquí están la trama original y dos partes iniciales del dendrograma.

Podemos seguir así de la manera obvia y emparejar puntos individuales, pero por suerte, R proporciona una función simple a la que puede llamar y que crea un dendrograma para usted. Se llama hclust () y toma como

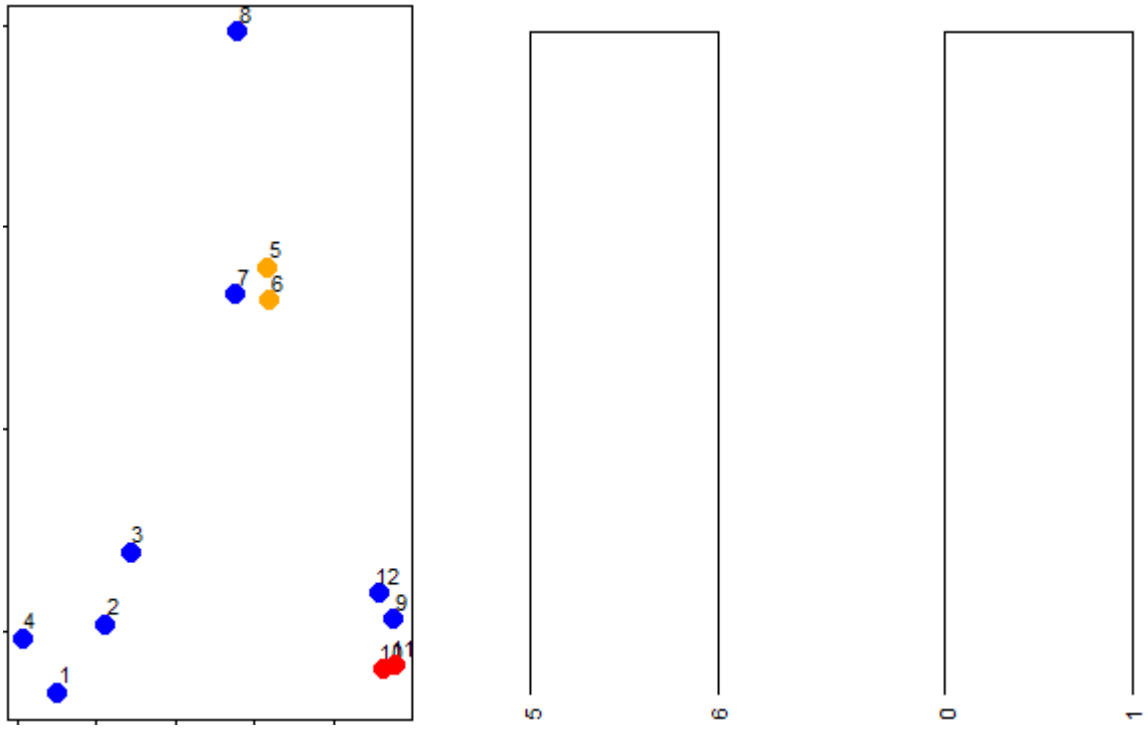
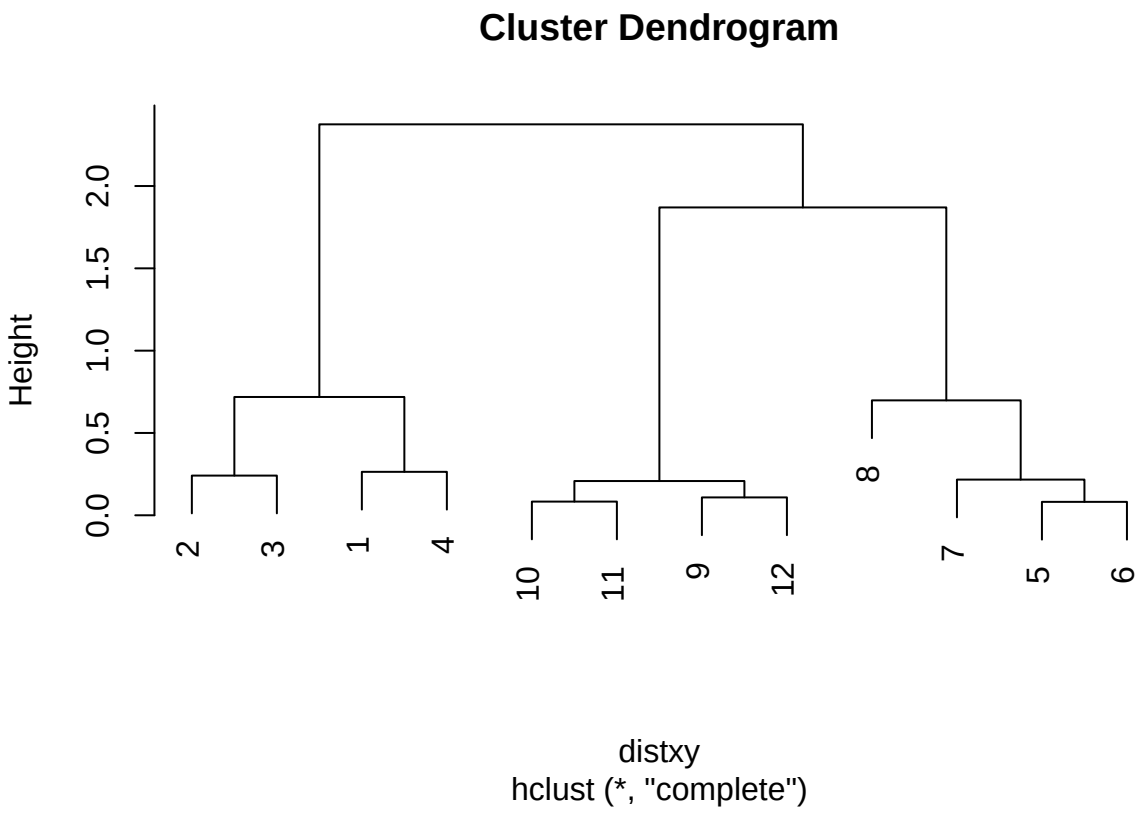


Figure 2: A caption

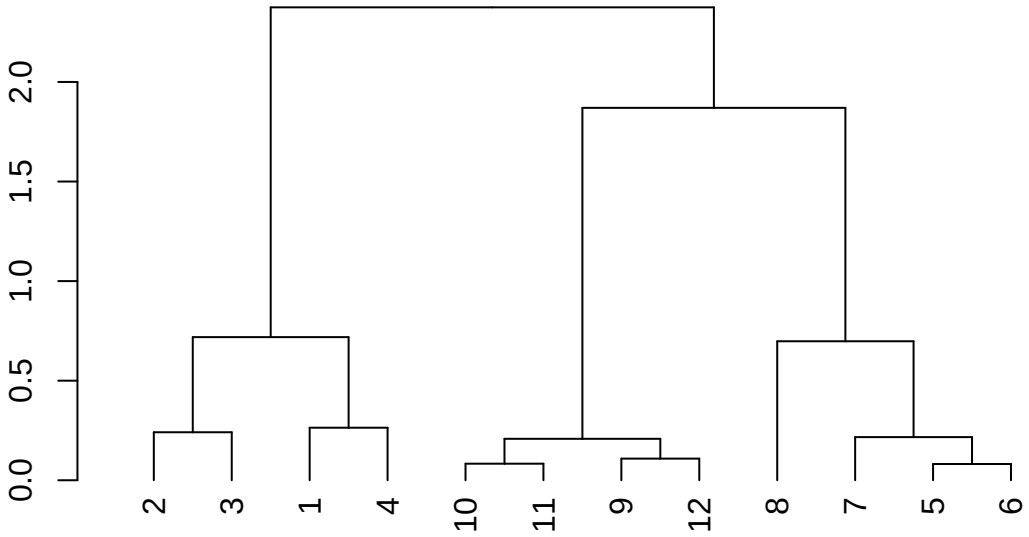
argumento la matriz de distancias por pares que vimos antes. Hemos almacenado esta matriz para usted en una variable llamada distxy. Ejecute hclust ahora con distxy como argumento y ponga el resultado en la variable hc.

```
hc <- hclust(distxy)
plot(hc)
```



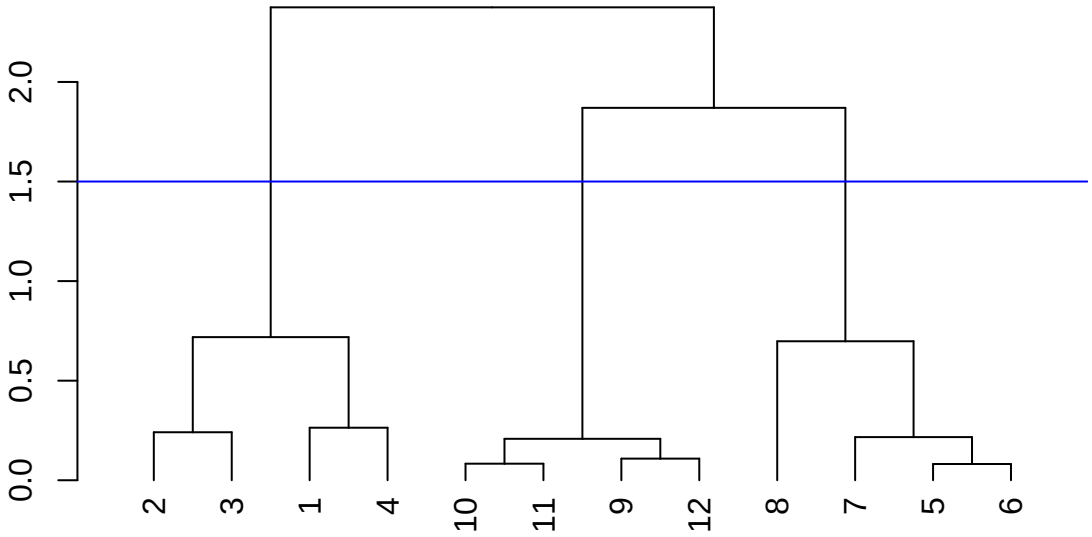
Bonita trama, ¿verdad? La trama de R convenientemente etiquetó todo para ti. Los puntos que vimos son las hojas en la parte inferior del gráfico, 5 y 6 están conectados, al igual que 10 y 11. Además, vemos que las 3 agrupaciones originales de puntos están más juntas como hojas en la imagen. Eso es reconfortante. Ahora llame a plot de nuevo, esta vez con el argumento as.dendrogram (hc).

```
plot(as.dendrogram(hc))
```



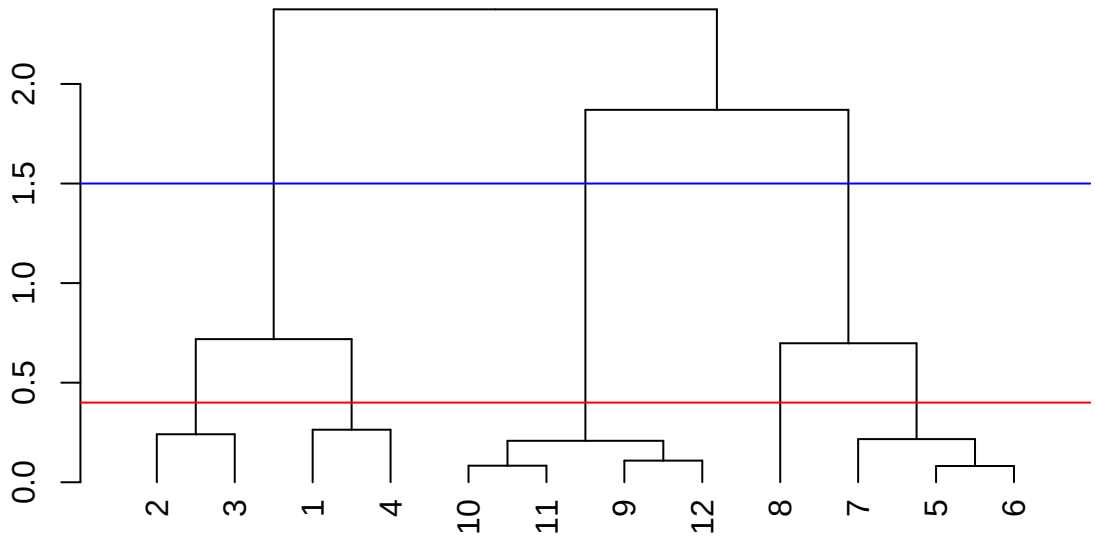
Lo esencial es lo mismo, pero faltan las etiquetas y las hojas (puntos originales) están todas impresas al mismo nivel. Observe que las alturas verticales de las líneas y el etiquetado de la escala en el borde izquierdo dan una indicación de la distancia. Use el comando R abline para dibujar una línea azul horizontal en 1.5 en este gráfico. Recuerde que esto requiere 2 argumentos, h = 1.5 y col = "blue".

```
plot(as.dendrogram(hc))
abline(h=1.5,col="blue")
```



Vemos que esta línea azul interseca 3 líneas verticales y esto nos dice que usar la distancia 1.5 (unidades no especificadas) nos da 3 grupos (1 a 4), (9 a 12) y (5 a 8). A esto lo llamamos un "corte" de nuestro dendrograma. Ahora corte el dendrogama dibujando una línea horizontal roja en .4.

```
plot(as.dendrogram(hc))
abline(h=1.5,col="blue")
abline(h=.4,col="red")
```



vemos que al cortar a .4 tenemos 5 grupos, lo que indica que esta distancia es lo suficientemente pequeña como para romper nuestra agrupación original de puntos.

Entonces, la cantidad de clústeres en sus datos depende de dónde dibuje la línea. (Dijimos que hay mucha flexibilidad aquí). Ahora que hemos visto la práctica, volvamos a algo de "teoría". Observe que las dos agrupaciones originales, 5 a 8 y 9 a 12, están conectadas con una línea horizontal cerca de la parte superior de la pantalla. Probablemente se esté preguntando cómo se miden las distancias entre grupos de puntos.

Hay varias formas de hacer esto. Solo mencionaremos dos. El primero se llama enlace completo y dice que si está tratando de medir una distancia entre dos grupos, tome la mayor distancia entre los pares de puntos en esos dos grupos. Obviamente, estos pares contienen un punto de cada grupo.

Entonces, si estuviéramos midiendo la distancia entre los dos grupos de puntos (1 a 4) y (5 a 8), usando el enlace completo como métrica, usaríamos la distancia entre los puntos 4 y 8 como medida, ya que esta es la distancia más grande. entre los pares de esos grupos.

La distancia entre los dos grupos de puntos (9 a 12) y (5 a 8), utilizando el vínculo completo como métrica, es la distancia entre los puntos 11 y 8, ya que esta es la distancia más grande entre los pares de esos grupos.

Por suerte, la distancia entre los dos grupos de puntos (del 9 al 12) y (del 1 al 4), utilizando el enlace completo como métrica, es la distancia entre los puntos 11 y 4.

Hemos creado el marco de datos dFsm para usted que contiene estos 3 puntos, 4, 8 y 11. Ejecute dist en dFsm para ver cuál es la distancia más pequeña entre estos 3 puntos.

```
 dFsm<-matrix(c(0.5308605,1.8906736, 2.9045615,0.9779429,2.4831670, 0.9118904),nrow = 3,ncol = 2)
dist(dFsm)
```

```
##          1        2
```

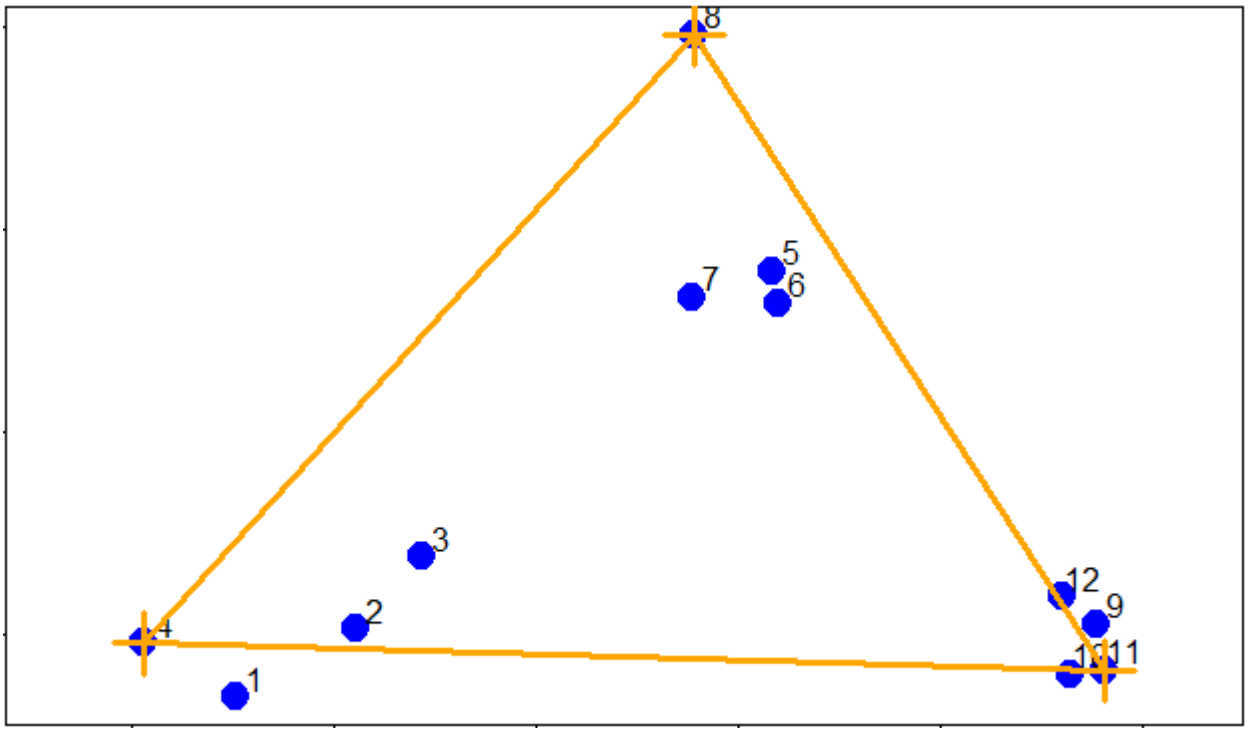


Figure 3: A caption

```
## 2 2.028495
## 3 2.374620 1.869994
```

Vemos que la distancia más pequeña está entre los puntos 2 y 3 en este conjunto reducido, (estos son en realidad los puntos 8 y 11 en el conjunto original), lo que indica que los dos grupos que estos puntos representan ((5 a 8) y (9 a 12) respectivamente) se unirían (a una distancia de 1.869) antes de conectarse con el tercer grupo (1 a 4). Esto es consistente con el dendrograma que trazamos.

La segunda forma de medir una distancia entre dos grupos que mencionaremos se llama enlace promedio. Primero calcula un punto "promedio" en cada grupo (considérelo como el centro de gravedad del grupo). Para ello, calcula las coordenadas xey medias (promedio) de los puntos del grupo.

Luego calcula las distancias entre el promedio de cada clúster para calcular la distancia entre clústeres. Ahora mire el clúster jerárquico que creamos antes

```
hc
```

```
##
## Call:
## hclust(d = distxy)
##
## Cluster method   : complete
## Distance         : euclidean
## Number of objects: 12
```

En nuestro conjunto simple de datos, los vínculos promedio y completo no son tan diferentes, pero en conjuntos de datos más complicados, el tipo de vínculo que usa podría afectar la forma en que se agrupan sus datos. Es una buena idea experimentar con diferentes métodos de vinculación para ver las diferentes

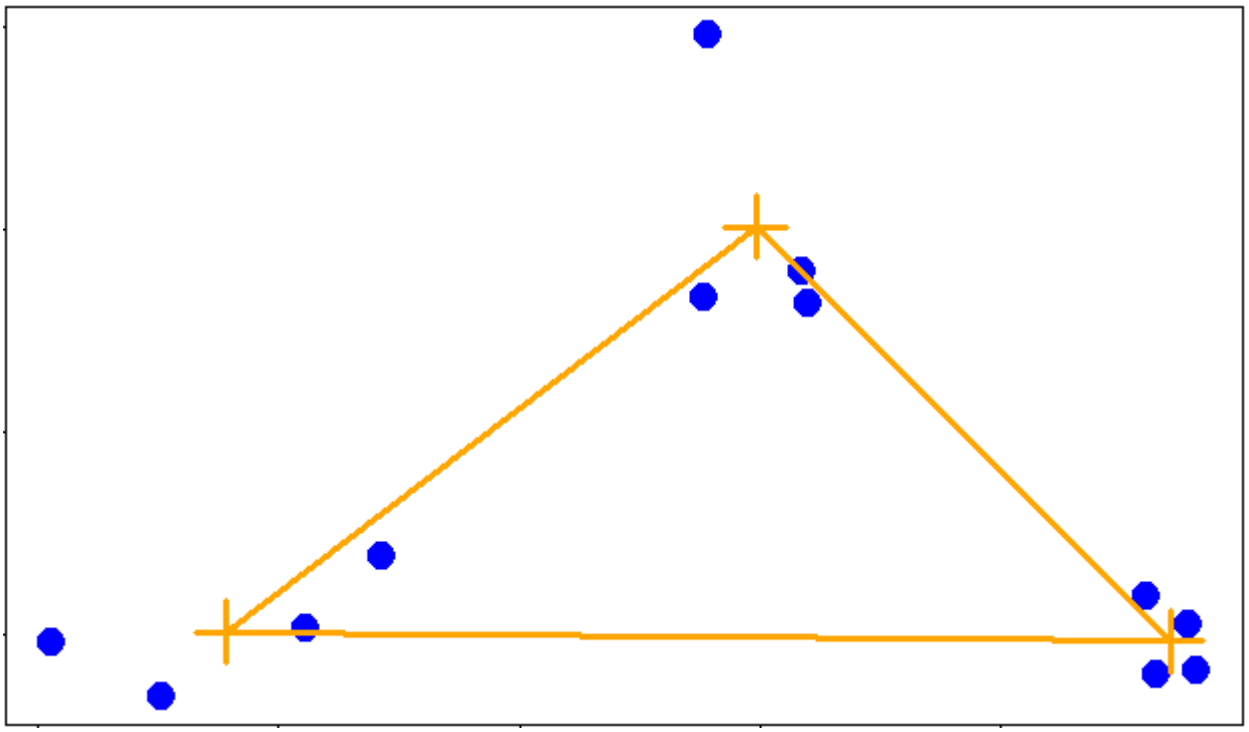


Figure 4: A caption

formas en que se agrupan sus datos. Esto le ayudará a determinar la mejor manera de continuar con su análisis.

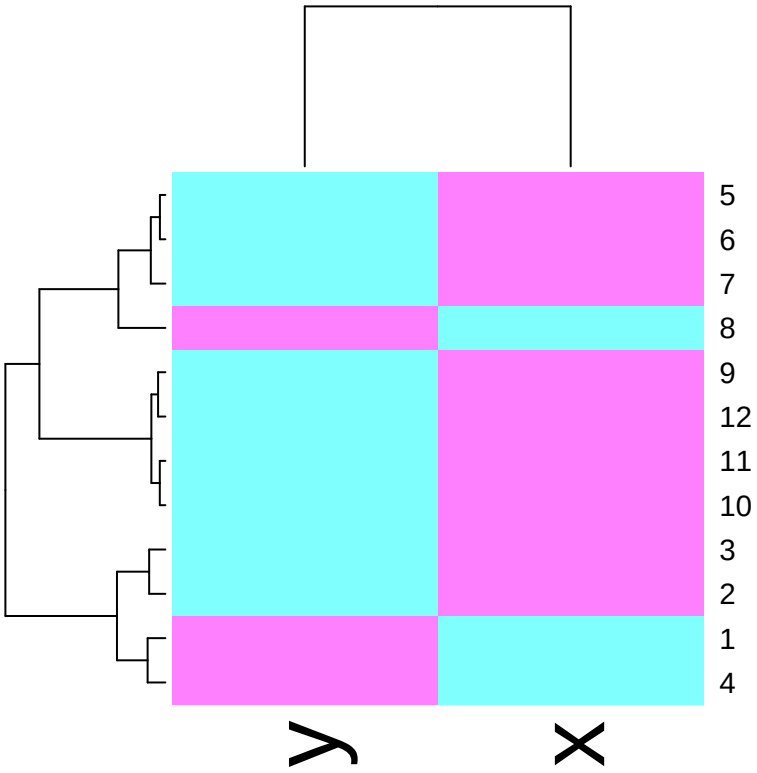
El último método de visualización de datos que mencionaremos en esta lección se refiere a los mapas de calor. Wikipedia (http://en.wikipedia.org/wiki/Heat_map) nos dice que un mapa de calor es "una representación gráfica de datos donde los valores individuales contenidos en una matriz se representan como colores. ... Los mapas de calor se originan en pantallas 2D de los valores en una matriz de datos. Los valores más grandes fueron representados por pequeños cuadrados grises oscuros o negros (píxeles) y los valores más pequeños por cuadrados más claros".

Probablemente haya visto muchos ejemplos de mapas de calor, por ejemplo, radares meteorológicos y pantallas de salinidad del océano. De Wikipedia (http://en.wikipedia.org/wiki/Heat_map) aprendemos que los mapas de calor se utilizan a menudo en biología molecular "para representar el nivel de expresión de muchos genes en una serie de muestras comparables (por ejemplo, células en diferentes estados , muestras de diferentes pacientes) ya que se obtienen de microarrays de ADN".

No diremos demasiado sobre este tema, pero existe un tutorial conciso muy agradable sobre cómo crear mapas de calor en R en http://sebastianraschka.com/Articles/heatmaps_in_r.html#clustering.

R proporciona una función útil para producir mapas de calor. Se llama mapa de calor. Hemos colocado los datos de puntos que hemos estado usando a lo largo de esta lección en una matriz. Llame al mapa de calor ahora con 2 argumentos. El primero es dataMatrix y el segundo es col set igual a cm. Colores (25). Este último es opcional, pero nos gustan más los colores que los predeterminados.

```
dataMatrix<-cbind(x,y)
heatmap(dataMatrix,col=cm.colors(25))
```



Vemos una especie de exhibición interesante. Este es un mapa de calor muy simple, simple porque los datos no son muy complejos. Las filas y columnas se agrupan como se muestra por colores. Las filas superiores (etiquetadas como 5, 6 y 7) parecen estar en el mismo grupo (mismos colores), mientras que el 8 está al lado de ellas pero con colores diferentes. Esto coincide con el dendrograma que se muestra en el borde izquierdo. De manera similar, 9, 12, 11 y 10 están agrupados (en filas) junto con 3 y 2. Estos son seguidos por 1 y 4 que están en un grupo separado. Los datos de las columnas se tratan independientemente de las filas, pero también se agrupan.

He