2020-21

Ejercicio 1

Gestión de datos

Óscar Gómez Borzdynski y Alejandro Cabana Suárez

## Desarrollo

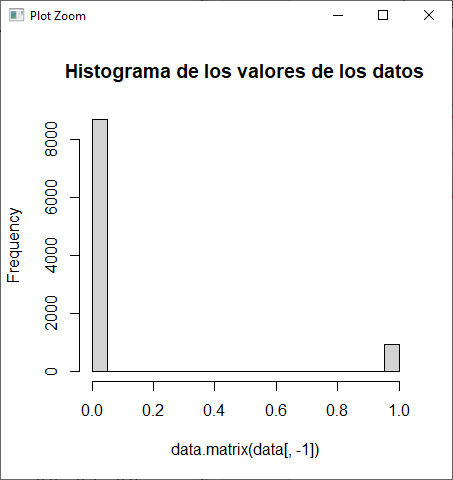
En primer lugar, cargamos los datos desde la url y eliminamos las columnas innecesarias. Hemos decidido eliminar el comentario, el id único, la fecha y las horas de inicio y fin.

# Cargar Datos  
data <- read.csv("http://cardsorting.net/tutorials/25.csv")

# Eliminar columnas innecesarias  
data$Comment <- NULL  
data$Uniqid <- NULL  
data[2:5] <- NULL

Posteriormente, realizamos un histograma de los datos presentes en el dataframe:

# Histograma de todos los valores  
hist(data.matrix(data[,-1]), main="Histograma de los valores de los datos")



Observamos que todos los valores presenten toman valores 0 o 1. Esto nos indica que el conjunto de datos indica la categoría seleccionada con un 1 mientras que en el resto de las categorías toman valor 0. Por otro lado, vemos que hay en torno a un 1 por cada nueve 0s.

Finalmente tratamos de buscar las tarjetas que están relacionadas, por ejemplo, queremos ver si en este experimento podemos establecer una relación entre los distintos tipos de frutas.

Para ello definimos una distancia entre las tarjetas mediante el comando dist de R. Para ello necesitamos trasponer la matriz del dataframe colocando las tarjetas como nombre de las columnas.

# Trasponemos el dataframe  
data <- as.data.frame(t(as.matrix(data)))

# Para poderlo usar con la función dist (que calcula la distancia entre   
# filas), quitamos la primera fila ya que es el nombre de la categoría  
dst <- dist(data[-1,], diag=TRUE)

Ésta matriz es un objeto de R llamado dist, lo que nos dificulta la representación de manera gráfica de la misma, para facilitarnos su tratamiento la transformaremos en una matriz estándar de R.

# Lo ponemos como matriz de datos numérica  
dst <- data.matrix(dst)

La primera representación que realizamos de la matriz de distancias es un mapa de calor, para ello generamos una imagen donde los ejes x e y son las tarjetas y en el eje z se representa la distancias entre las categorías correspondientes.

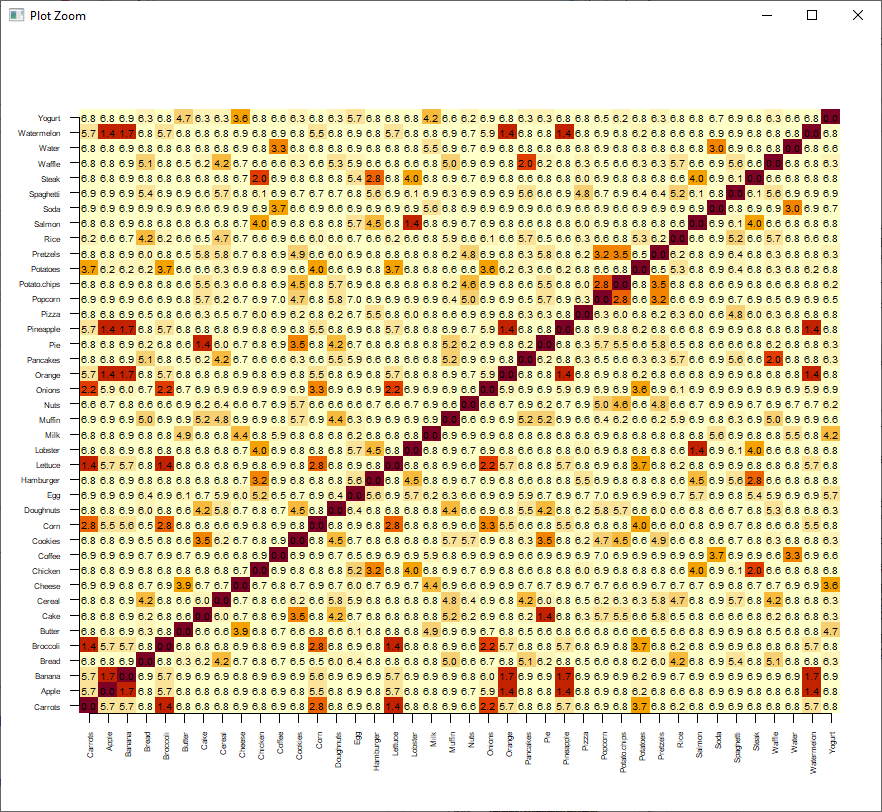
Utilizaremos la negación de la distancia al crear el mapa de calor para que las categorías cercanas tomen colores oscuros, siendo más fáciles de localizar.

# Obtenemos el número de categorías  
dim <- ncol(dst)

# Dibujar una imagen donde el eje "z" es la distancia, de esta manera  
# el color toma valores oscuros para distancias bajas  
image(1:dim, 1:dim, -dst, axes = FALSE, xlab="", ylab="")

# Colocamos los nombres de las categorías en los ejes  
axis(1, 1:dim, row.names(data)[-1], cex.axis = 0.5, las=3)  
axis(2, 1:dim, row.names(data)[-1], cex.axis = 0.5, las=1)

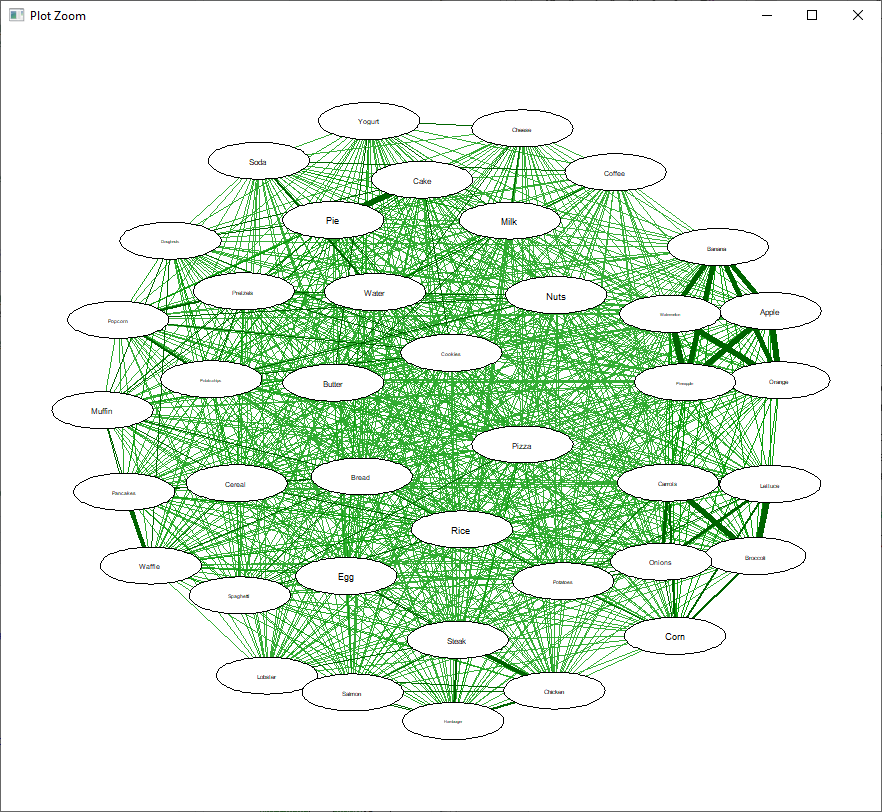
# Colocamos los valores de las distancias en las posiciones correspondientes  
text(expand.grid(1:dim, 1:dim), sprintf("%0.1f", dst), cex=0.6)



En el mapa de calor se visualiza la diagonal con distancia 0 ya que el vector de una categoría presenta distancia 0 respecto a sí mismo. Podemos detectar también cercanía entre algunos elementos como pueden ser el grupo {manzana, plátano, naranja, sandía y piña}.

Finalmente representamos la matriz de distancia como grafo. Para ello utilizamos la librería qgraph. Para tener una visualización más intuitiva mostramos la inversa de la distancia, haciendo que las aristas con menor distancia tomen grosores mayores.

# Hacemos el grafo de distancias. Utilizamos 1/dst porque cuanto menor   
# distancia mayor tamaño de arista.  
dst <- 1/dst  
qgraph(dst, layout='spring', labels = colnames(dst))



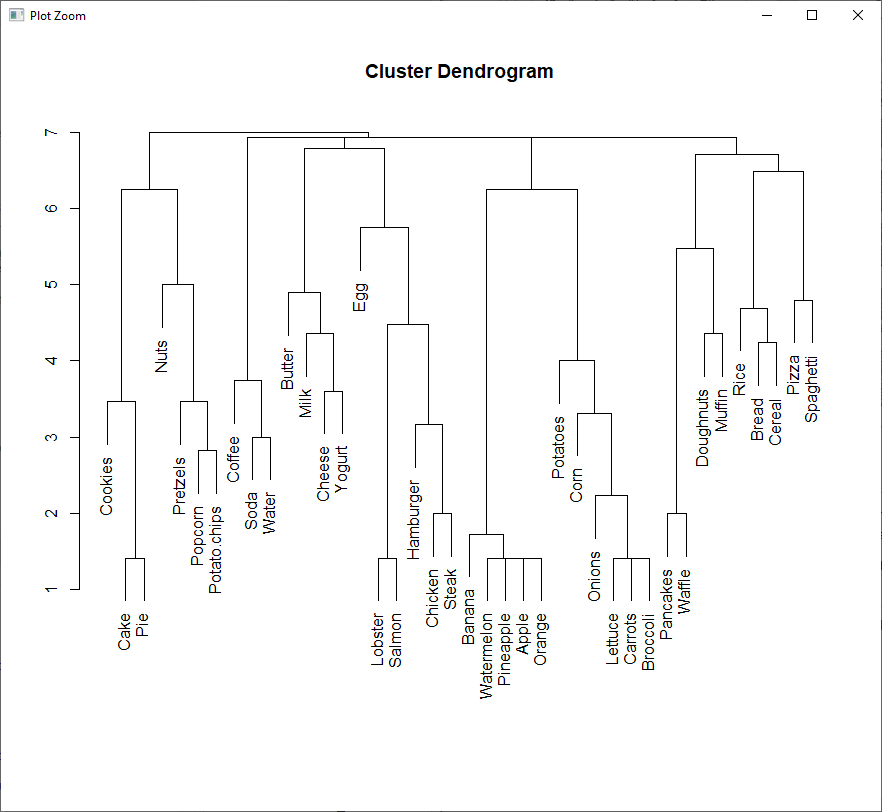
En este grafo se pueden visualizar con gran claridad la relación entre las cinco frutas comentadas previamente. Apreciamos también la relación entre los siguientes grupos:

* {pastel y tarta}, dulces de formato similar.
* {gofres y tortitas}, dulces que habitualmente se toman en desayunos o meriendas.
* {filete y pollo}, elementos cárnicos.
* {zanahoria, cebolla, brócoli y lechuga}, verduras varias.

Todas estas relaciones parecen tener un significado ya que todos los elementos pertenecen a un grupo de alimentos diferentes.

Pese a que no se nos proponía, hemos decidido realizar un dendrograma partiendo desde la matriz de distancias. Para ello utilizamos el comando hclust de R.

# Dendrograma desde la matriz de distancias  
plot(hclust(dist(data[-1,], diag=TRUE)))



En este gráfico podemos ver las relaciones previas así como otras que pasaron inadvertidas como el conjunto {salmón y langosta}.

## Conclusiones

Partiendo de la misma información, en nuestro caso una matriz de distancias, hemos generado diversas gráficas que sintetizan la información presente de manera más comprensible para un humano. Cada representación tiene sus características positivas y negativas.

En el caso del mapa de calor es muy visual para buscar cambios grandes en la distancia entre categorías. En el caso de tener una zona con distancias grandes (tonos claros) con un solo elemento oscuro, podremos diferenciarlo rápidamente. También permite buscar la relación uno a muchos de una categoría.

El grafo es muy útil para buscar relaciones entre varios elementos como conjunto, como se aprecia en el conjunto de las frutas. Esta relación es muy compleja de ver en el mapa de calor.

El dendrograma nos permite visualizar prácticamente todas las relaciones de una manera sencilla. En caso de que dos elementos estén en la misma rama significará que pertenecen a la misma categoría. Estas ramificaciones también ayudan a crear una jerarquía entre los elementos, pudiendo decir que una tarta está mas cerca de una nuez que de una patata, que se encuentra en una rama completamente distinta.