10-12-221

GESTIÓN DE DATOS. Práctica 1 Ejercicio sobre Exploración

Carlos Ramos Mateos y Adrián Rubio Pintado
UNIVERSIDAD AUTÓNOMA DE MADRID

Enunciado

1. Crear un código en R (bien documentado internamente, que permita:

a. Leer el dataset desde su origen (a través de la dirección web suministrada).

Elaboramos un código en R que carga el csv desde url.

```
url_path <- "http://cardsorting.net/tutorials/25.csv"

data <- read.csv(url(url_path))
```

b. Realizar las transformaciones que se consideren convenientes para trabajar de manera efectiva con las categorías y las tarjetas. Se deberá obviar toda la información que no sea de utilidad.

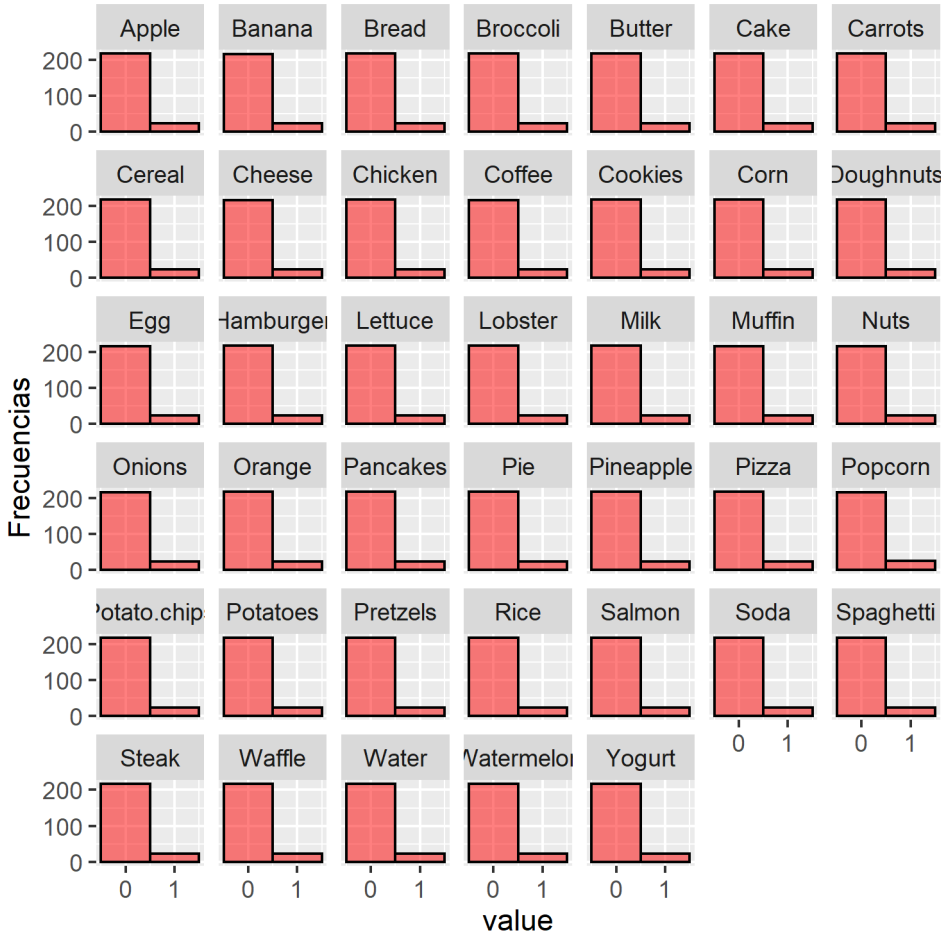
Simplemente hacemos eliminamos las columnas solicitadas del dataset creado.

```
#Eliminamos las columnas Uniqid, Startdate, Starttime, Endtime, QID y Comment

df <- subset(data, select = -c(Category, Uniqid, Startdate, Starttime, Endtime, QID,
Comment))

#Histograma

df %>% gather()
```

c. Representar un histograma, u otro gráfico basado en frecuencias o densidad, para estudiar los datos numéricos que aparecen en el dataset, así como su frecuencia de aparición.

Leemos el dataset utilizando el código en R adjunto a la memoria. Las transformaciones que consideramos necesarios es eliminar la información prescindible, esto es, eliminar las columnas Uniqid, Startdate, Starttime, Endtime, QID y Comment.

Una vez tratados los datos, pintamos las frecuencias para cada columna(tarjetas) mediante un histograma.



Histograma de las variables

Lo primero que podemos apreciar de estos histogramas es que tienen una distribución binaria, donde para todas las variables toman valores entre 0 y 1. Cabe destacar que para todas las variables, los histogramas tienen una distribución similar, más ceros que unos. Esto se debe principalmente a la propia naturaleza del dataset, ya que al ser un card-sorting abierto, los diferentes usuarios han creado diferentes categorías a la hora de clasificar los items.

Usando este histograma, vemos como no es necesario normalizar las distancias, que emplearemos en apartado siguiente para calcular la matriz de distancia. Esto nos permite ver que el método de tarjetas sigue un sistema binario. [1]

d. Crear una matriz de distancia o de similitud de tarjetas. ¿Qué visualización es la más adecuada para esta matriz? Represéntala convenientemente.

Del punto anterior podemos ver claramente que tenemos variables **binarias simétricas**, ya que las dos valores: pertenecia(1) y no pertenencia(0) son igual de relevantes. Por lo que conociendo esto, para el cálculo de la matriz de distancias usaremos la **distancia euclídea**.

Para representar esta matriz de distancias entre pares de variables, decidimos utilizar un mapa de calor por ser la visualización más adecuada para este tipo de modelos, en especial para un problema de card-sorting como el que nos estamos enfrentando.

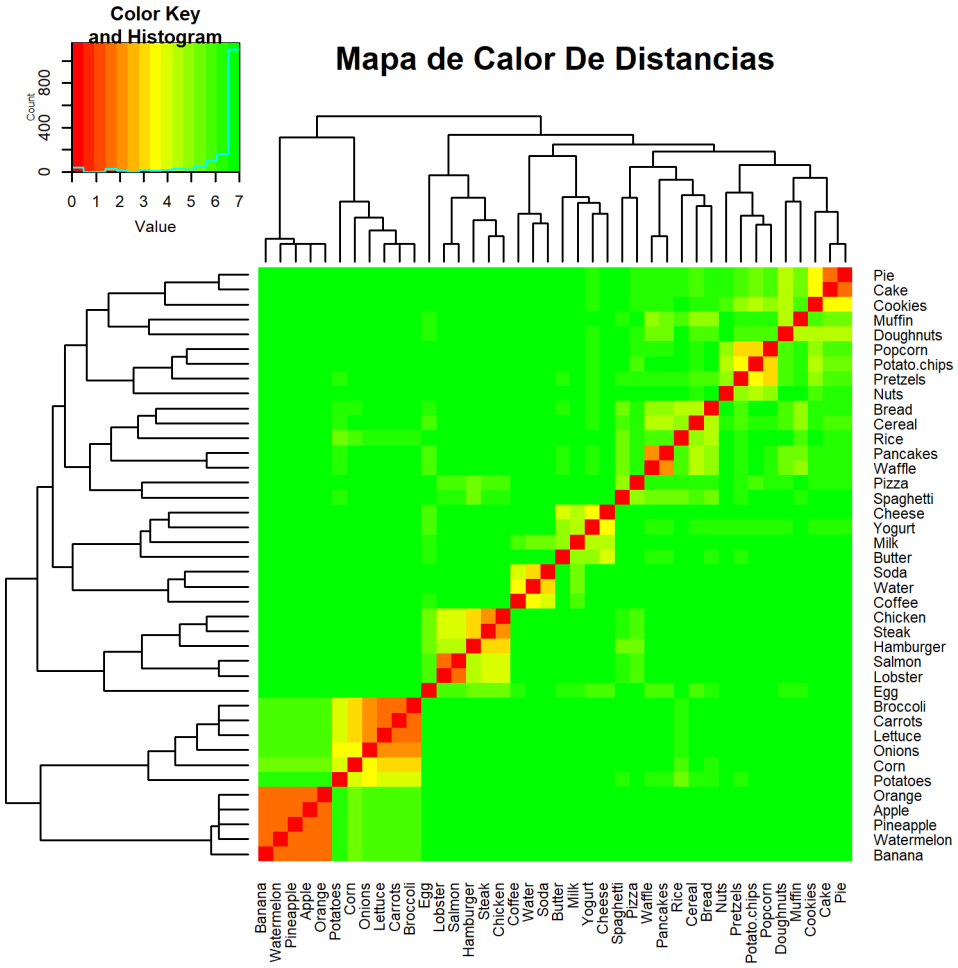
Utilizamos un color verde para representar la distancia 0, un color rojo para representar las distancias con el valor más alto, y colores amarillentos para valores intermedios (usando los

colores intermedios de la escala cromática también a modo de degradado para representar valores intermedios).

Utilizar un **dendrograma** es recomendable como método de distancias para un problema con el tipo de datos de card-sorting, ya que funcionan bien para la clasificación en clúster jerárquica. Este dendrograma nos permite apreciar claramente las relaciones de agrupación entre los datos. También entre grupos de datos, pudiendo así observar las sucesivas subdivisiones que hace y hacernos una idea sobre los criterios de agrupación de los mismos. Los dendrogamas graficados a continuación, en el eje X y el eje Y de la matriz de distancias, son el mismo, data la naturaleza de la matriz de distancias.

Si observamos por ejemplo el dendogama, vemos tarjetas como “Apple”, “Banana” o “Orange” en el mismo primer grupo del dendogama. Esto nos permite ver el criterio de agrupación de los datos, y ver que estas tarjetas son fuertemente similares. Vemos por ejemplo otro caso similar, las tarjetas “Onions” y “Potatoes”, forman otro grupo fuertemente similar. En cambio, tarjetas como “Banana” y “Cookies”, vemos que guardan una disimilitud fuerte, de acuerdo con el criterio de “número de tarjetas clasificadas en cada categoría”.

Obtenemos el siguiente mapa de calor, junto con su dendrogama pertinente utilizando la matriz de **distancias euclídeas** como función de disimilitud:

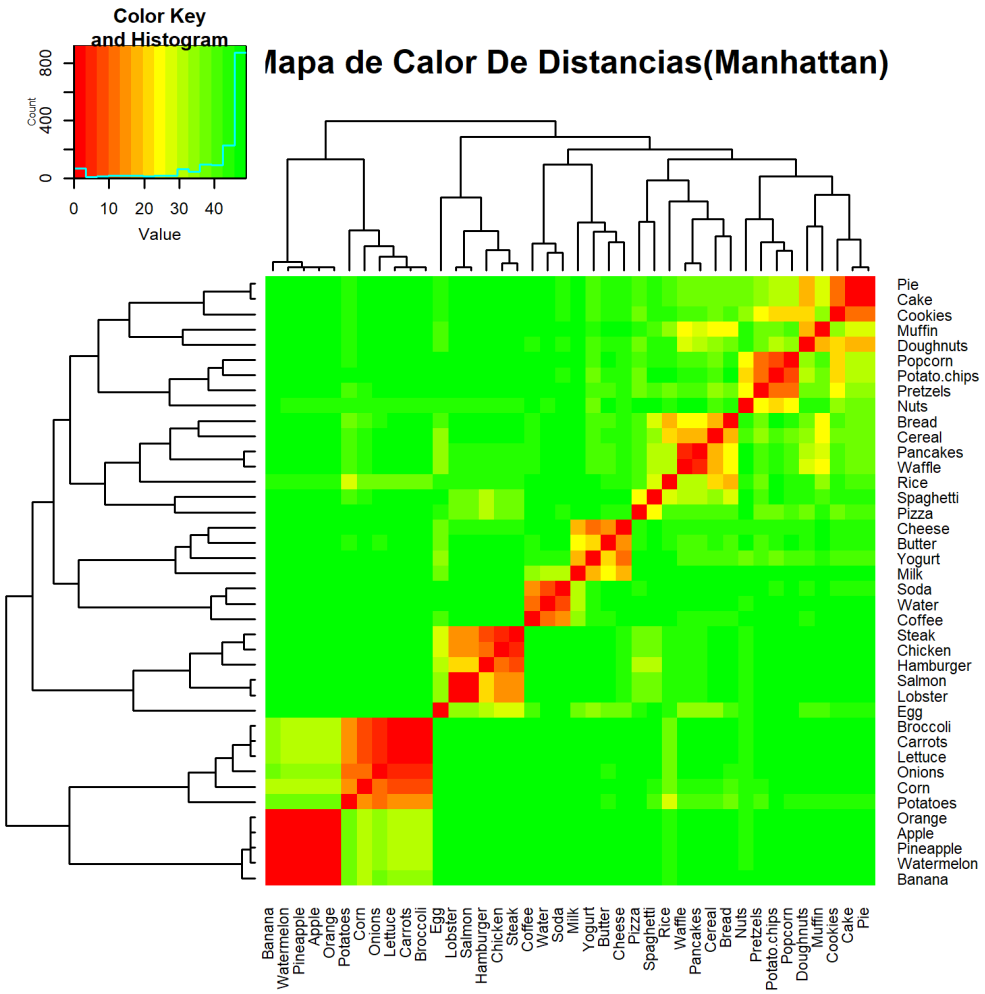


Mapa de calor y dendrograma con matriz de distancias euclídea.

Vemos como en la leyenda de colores se dibuja una línea azul, que representa la distribución de distancias entre las tarjetas. Vemos como la gran mayoría de ellas se encuentran en la zona verde, ya que la mayoría de las tarjetas son distantes (color predominante es el verde en el mapa

de calor), es decir, que las tarjetas guardan poca información semántica entre ellas, de acuerdo con la categorización de los usuarios.

También podemos utilizar la distancia Manhattan como medida de disimilitud, ya que como la euclídea pertenece a la familia de métrica Minkowski.

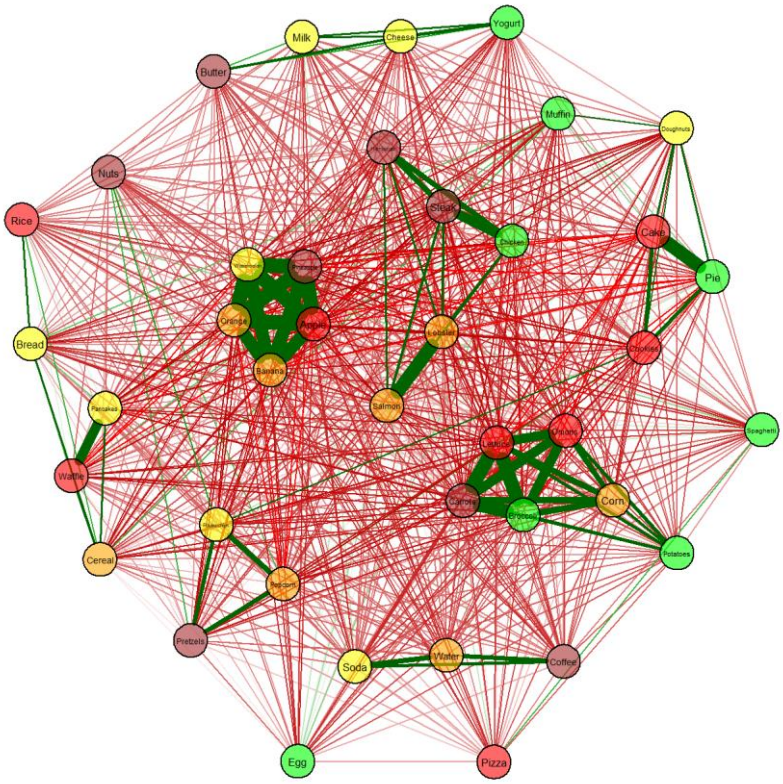


Mapa de calor y dendrograma con matriz de distancias Manhattan

Los **dendrogramas entre ambas distancias** son muy similares en cuanto a agrupación de categorías. Sin embargo, existen diferencias topológicas y estadísticas en cuanto al método de distancia y el criterio de vinculación aplicado, obteniendo diferentes topologías, ramas y alturas. Es decir, en la distancia euclídea se producen más relaciones entre clústeres a un mayor nivel jerárquico, mientras que en manhattan estos agrupamientos se producen a un nivel más bajo ya que sus distancias relativas son menores.

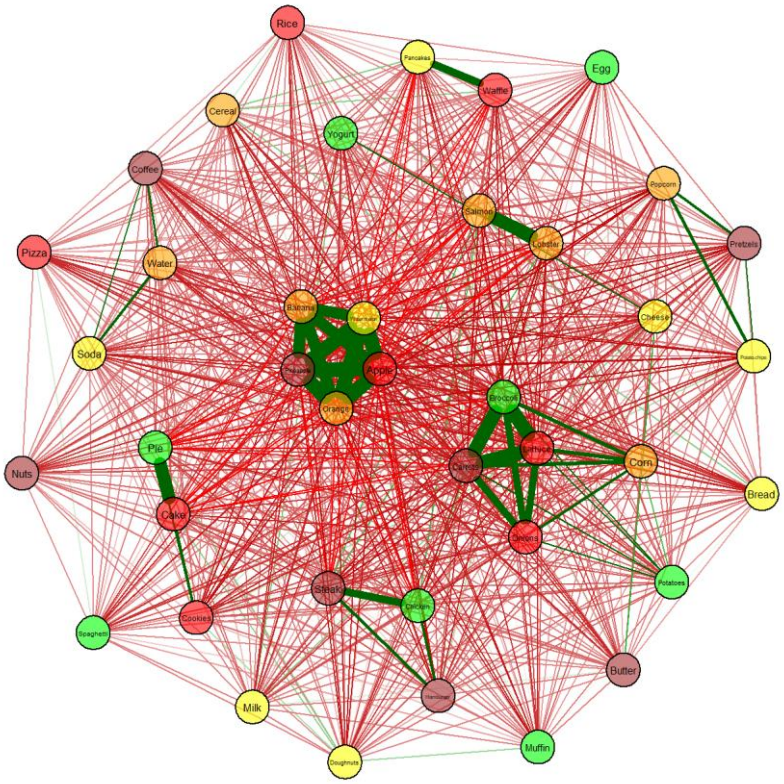
e. Representar gráficamente las relaciones entre las tarjetas a través de un grafo, utilizando para ello la librería qgraph de R, de forma que las tarjetas más relacionadas se distingan de manera visual.

Utilizando esta misma matriz de distancias, utilizamos la función qgraph para mostrar esta información en forma de grafo, hemos implementado dos paletas de colores: la primera nos permite separar los diferentes grupos(marrón-rojo-naranja-amarillo-verde) y otras nos muestran el valor que tienen las relaciones entre los pares (rojo muestra mucha distancia, verde muestra cercanía, es decir, correlación).



Grafo con matriz de distancia euclídeas

Con ambos grafos podemos validar las conclusiones extraídas del dendrograma y del mapa de calor, lo cual es lógico ya que estamos partiendo de la misma matriz de distancias, simplemente es una forma diferente de visualizar esta misma información.



Grafo con matriz de distancia manhattan

f. Finalmente, ¿cuáles son las tarjetas que están más relacionadas? ¿Tiene esta relación sentido a nivel semántico (en función de los ítems de dominio que representan)?

Las tarjetas más relacionadas son las pertencientes al grupo de **“Apple”, “Watermelon”, “Pineapple” “Banana” y “Orange”** en el mismo primer gran grupo del dendrogama. Esto lo podemos apreciar de manera más marcada con la distancia Manhattan con la matriz de distancias. Semánticamente si guardan relación, ya que todas ellas son frutas.

El segundo grupo de tarjetas más relacionadas es el de “Onions” y “Potatoes”, que forman otro grupo fuertemente similar. A nivel semántico sí guardan relación, ya que ambos alimentos son verduras. Es decir, que dentro del dominio que representan, sí guardan relación semántica muy fuerte. También podemos ver otros grupos menores pero que si tienen cierta correlación entre sí como el de bebidas (Coffe, Soda, Water), el de lácteos (Cheese, Butter, Yogurt, Milk), bollería (Pie, Cake, Cookies) o Snack (Popcorn, chips, Pretzles, Nuts). En definitiva, podemos ver que hay relaciones fuertes entre los productos que irían dentro de una misma sección de un supermercado.

En cambio, tarjetas como “Banana” y “Cookies”, vemos que guardan una disimilitud fuerte, de acuerdo con el criterio de “número de tarjetas clasificadas en cada categoría”.

Bibliografía

[1] José A. Macías - Enhancing Card Sorting Dendrograms through the Holistic Analysis of Distance Methods and Linkage Criteria - 2021