1 Objetivos

En esta práctica estudiaremos un dataset de *CardSorting* para ver las relaciones entre diversos productos de supermercado. Para ello emplearemos diferentes técnicas de visualización de datos como: un hisograma, un *heatmap*, un dendograma y un grafo de similitudes entre tarjetas.

2 Lectura y limpieza de datos

En primer lugar, utilizamos las funciones *read.csv* y *url* de R para leer los datos del dataset utilizando la dirección `http://cardsorting.net/tutorials/25.csv`. Como únicamente estamos interesados en los datos numéricos, eliminamos las columnas *Uniqid*, *Category*, *Startdate*, *Starttime*, *Endtime*, *QID* y *Comments*.

3 Densidad númerica de los datos

Como primera aproximación a la visualización de estos datos estamos interesados en conocer los valores númericos que toma nuestro dataset, así como la distribución entre los mismos. Para ello nuestra mejor herramienta será un simple histograma sobre el dataset completo.

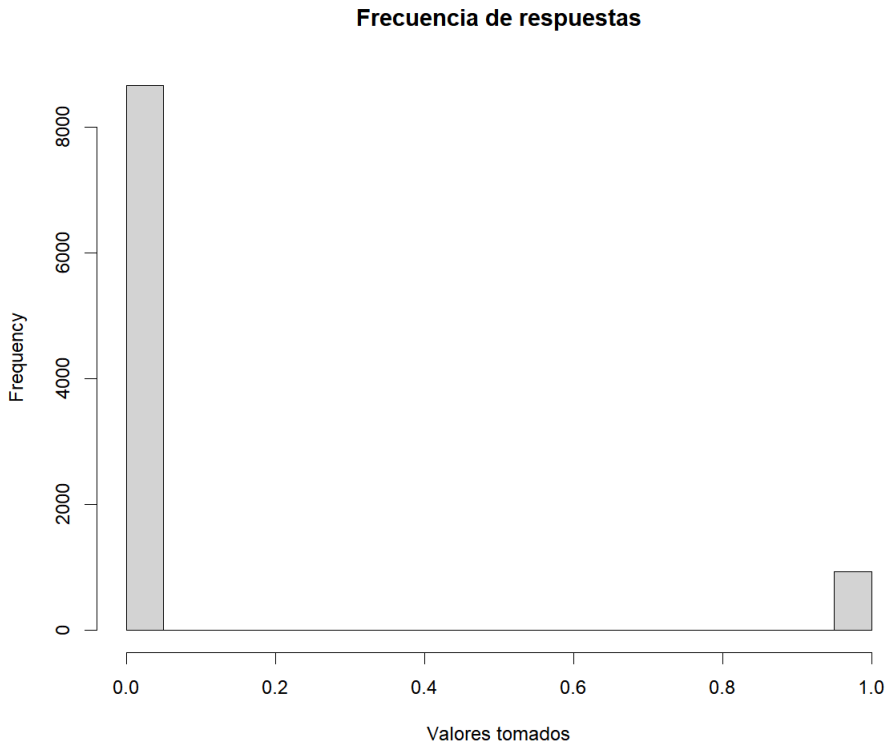

Figure 1: Histograma de nuestro dataset.

Vemos como la mayoría de los valores están en 0, mientras que algunos pocos toman el valor 1. Recordemos la información que codifica este dataset: cada columna es un producto del supermercado (una *card* en nuestro *CardSorting*), mientras que cada fila es una categoría, introducida por el usuario con Id asociado *Uniqid*. Un 1 significa que el usuario ha asociado ese producto con esa categoría, y un 0 que no lo ha hecho.

Por un lado, nuestro análisis númerico dado por el histograma encaja con nuestra intuición: la gran mayoría de los valores serán 0. Por otro lado, puesto que las categorías son introducidas por los usuarios de forma manual, obtendremos muchas categorías similiares pero ligeramente distintas. Por ejemplo, para frutas y verduras obtenemos entre otras las siguientes categorías: *fruits*, *veggies*, *fruits and veggies* y *frutis and veggies*.

4 Similitud entre tarjetas

Tras una primera impresión de la distribución numérica de los datos, procedemos a realizar un estudio sobre la similitud hayada entre las tarjetas de este dataset. Para ello computaremos la matriz de distancias entre las distintas tarjetas, utilizando la distancia euclídea. Cabe destacar que puesto que los valores de nuestro dataset están en $\{0, 1\}$, $x^2 = x$, y la distancia en L_1 será equivalente a la distancia euclídea de L_2. A pesar de ello, sería interesante utilizar distintas distancias para comparar los resultados entre ambas.

Tras computar la matriz de distancias hacemos uso de la función *heatmap.2* de R para representar un mapa de calor. Valores más bajos (rojos) representarán distancias más bajas. Es decir, tarjetas más relacionadas. En contraparte, valores más altos (amarillo claro y blanco) representarán distancias altas. Es decir, tarjetas poco relacionadas.

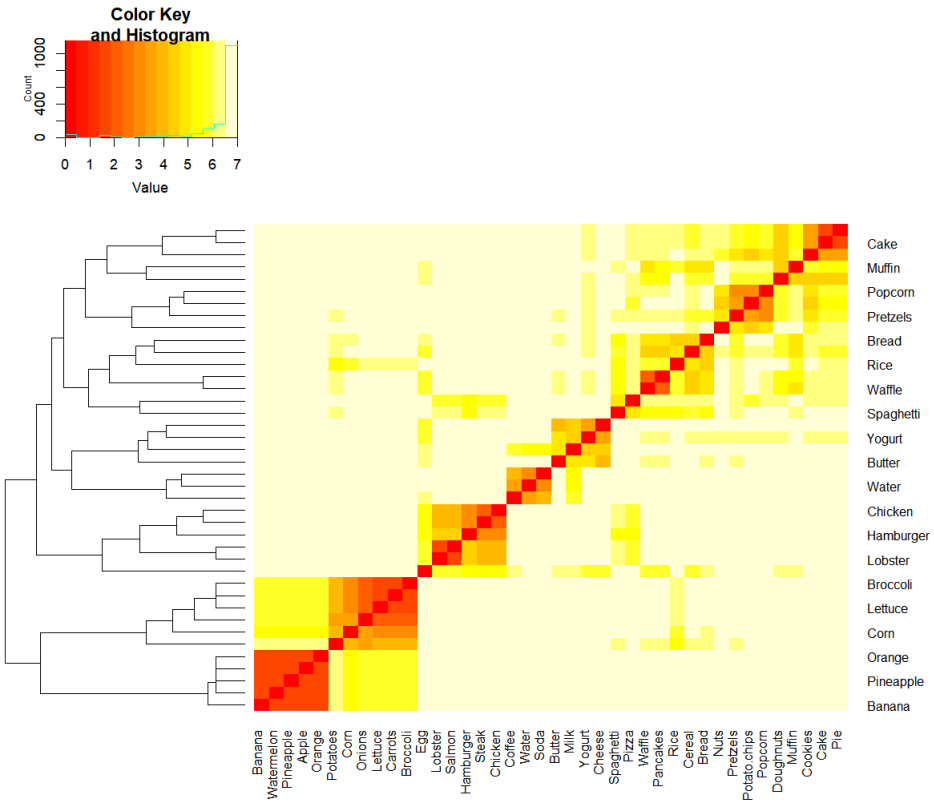

Figure 2: Heatmap y dendograma de las distancias entre tarjetas.

Vemos un bloque de tarjetas muy relacionado entre si en la parte inferior izquierda que corresponde a las frutas, y uno cercano también muy relacionado que corresponde a las verduras: *Banana*, *Watermelon*, *Pineapple*, *Apple*, *Orange*, *Potatoes*, *Lettuce*, *Carrots* y *Broccoli*.

Vemos también un conjunto de tarjetas relacionadas en la esquina superior derecha. Destaca en particular *Pie* y *Cake*. Finalmente, en la zona central pegada a las verduras hay otra pareja que destaca: *Salmon* y *Lobster*. Todo este tipo de relaciones tienen sentido semántica y culturalmente, lo que revela que nuestro análisis sobre la relación entre tarjetas está siendo acertado.

A la izquierda del *heatmap* podemos ver un dendograma con las diviones en clusters y subclusters. Esto nos indica lo cercanos que están los elementos entre si de forma visualmente distintas, pero las conclusiones que podemos obtener al respecto son equivalentes a las ya comentadas.

5 Relaciones entre tarjetas

De cara a visualizar las relaciones entre tarjetas de forma aún más directa podemos utilizar un grafo ponderado, donde los nodos representarán tarjetas y los lados, las relaciones entre las mismas. Grafos más resaltados (con mayor ponderación) revelarán una mayor relación entre las tarjetas.

Para computar este tipo de gráfico necesitamos una matriz de similitudes en vez de una matriz de distancias como la que hemos utilizado hasta ahora. Podemos utilizar directamente la inversa elemento a elemento de esta matriz: definiremos la similitud entre dos elementos como el inverso de su distancia:

$$\text{Sim}(a, b) = \frac{1}{\text{dist}(a, b)}$$

Obtenemos el siguiente gráfico como resultado:

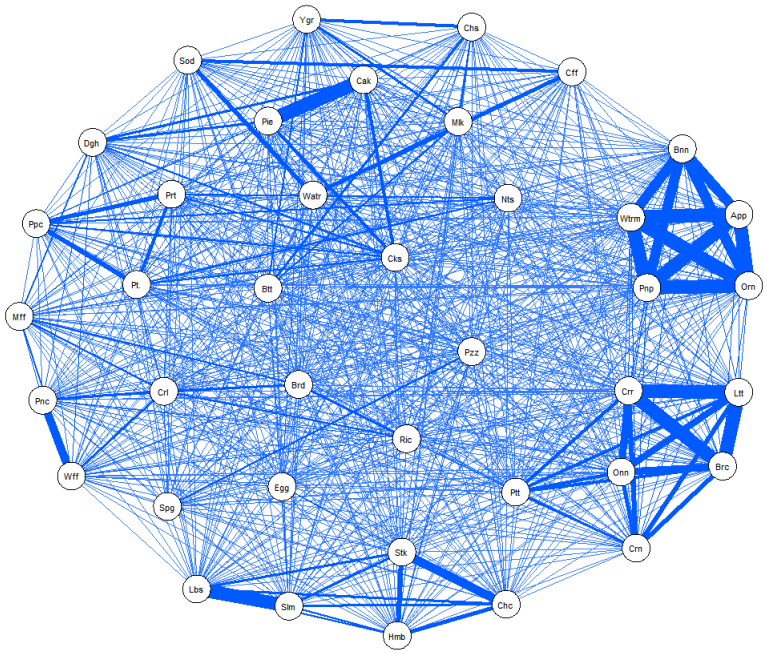

Figure 3: Grafo ponderado de las similitud entre tarjetas

En este nuevo gráfico es sencillo apreciar los mismos resultados que en el *heatmap* y en el dendograma: hay dos clusteres de elementos relacionados claramente, las frutas y las verduras, ligeramente serarados. Por otro lado encontramos fuertes relaciones entre las parejas (*Salmon*, *Lobster*) (*Pie*, *Cake*).

6 Tarjetas más relacionadas

Aunque hemos ido comentando este concepto a lo largo de toda la práctica podemos realizar una análisis numérico para verificar que los valores observados en los gráficos están asociados a las mayores similitudes, y verificar cuáles están cercanos pero no llegan a ser exactamente las más relacionadas entre las anteriormente mencionadas. Para ello, computamos la mínima distincia en la matriz de distancias, obteniendo $1.4142 \approx \sqrt{2}$. Es decir, las tarjetas más relacionadas se diferencian en únicamente dos valores.

Finalmente, mostramos una tabla con todas las parejas que alcanzan esta distancia mínima en nuestra matriz:

Item 1	Item 2
Broccoli	Carrots
Lettuce	Carrots
Orange	Apple
Pineapple	Apple
Watermelon	Apple
Carrots	Broccoli
Lettuce	Broccoli
Pie	Cake
Carrots	Lettuce
Broccoli	Lettuce
Salmon	Lobster
Apple	Orange
Pineapple	Orange
Watermelon	Orange
Cake	Pie
Apple	Pineapple
Orange	Pineapple
Watermelon	Pineapple
Lobster	Salmon
Apple	Watermelon
Orange	Watermelon
Pineapple	Watermelon

En esta tabla observamos las tarjetas previamente comentadas. En particular, las parejas (*Salmon*, *Lobster*) (*Pie*, *Cake*) están presentes, así como la mayoría de frutas y verduras. Si nos fijamos con atención podemos apreciar como *Onion* no está presente en esta tabla, pues no está tan íntimamente relacionada con el resto de verduras.