

Gestión de Datos: Ejercicios 1

Exploración gráfica de datos

13 de diciembre de 2021

María Barroso Honrubia
Gloria del Valle Cano

Exploración gráfica de datos

En este ejercicio hemos explorado gráficamente un dataset que se obtuvo mediante un *Card Sorting* de alimentos. Buscaremos información relevante sobre la tipología y el rango de los datos numéricos y las tarjetas más relacionadas entre sí.

Pregunta 1

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

Para ello se ha utilizado la función `read.csv` de R (ver más en el código adjunto).

Pregunta 2

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

Debido a que en el estudio del dataset son de importancia solamente los datos de clasificación de categorías y las tarjetas, se eliminan las columnas `Uniqid`, `Startdate`, `Starttime`, `Endtime`, `QID` y `Comment`, a través de un `subset`.

Pregunta 3

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.

Para visualizar la información, separamos los datos numéricos de la columna `Category` y con los datos restantes numéricos realizamos un *bar plot*. Observamos en la Figura 1 que el conjunto está formado por dos valores numéricos, 0 y 1, siendo mucho mayor la proporción de 0s.

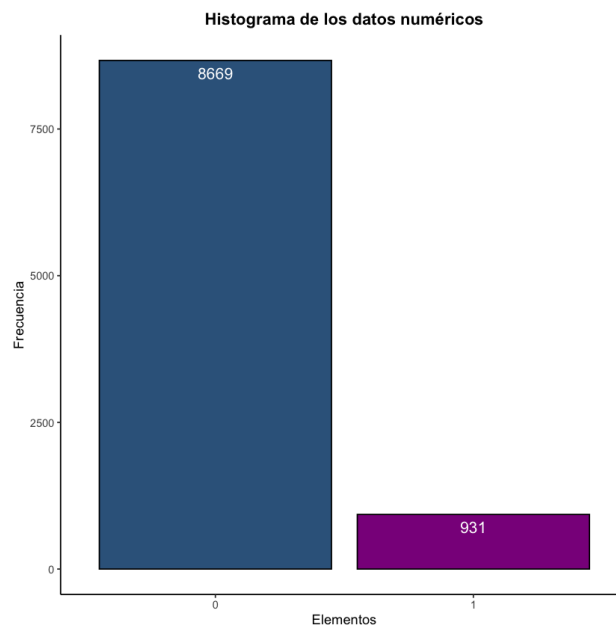


Figura 1: Histograma de los datos numéricos

Asimismo, vemos que tiene sentido al haber un 1 por cada alimento que los usuarios hayan declarado pertenecer a una categoría determinada, dentro de una matriz donde cada uno de los alimentos no tienen por qué pertenecer a muchas de las categorías, ya que las categorías son muy distinguibles entre sí.

Pregunta 4

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

Para crear la matriz se ha necesitado utilizar la métrica euclídea con la función `dist` de R y la opción *euclidean*. La mejor visualización para esta matriz de distancias es un mapa de calor (*heatmap*), para lo que se puede usar la función `gplots::heatmap.2` la cual nos ofrece la ordenación por similitud las filas de la matriz y así obtenemos una representación visual de los clusters de similitud. Este resultado se ofrece en la Figura 2.

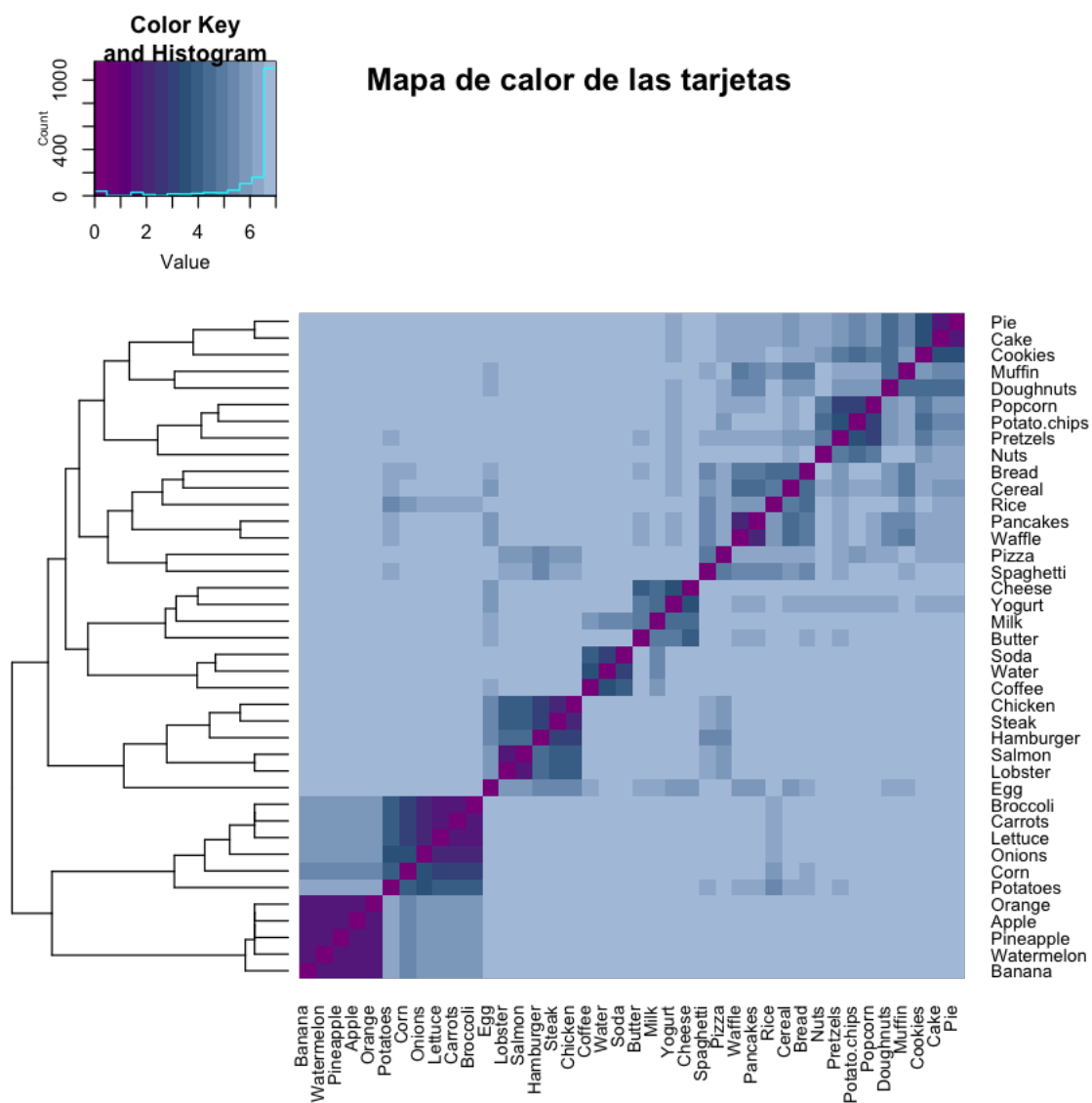


Figura 2: Mapa de calor de las tarjetas

Es destacable mencionar que se ven con mucha más fuerza la similitud entre tarjetas a medida que el color es más intenso, en concreto, hay varios grupos que se ven en azul más oscuro y morado. Además, el dendograma facilita la visualización de las conexiones entre variables mucho más correlacionadas.

Pregunta 5

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

Para poder estudiar un poco más allá las relaciones entre tarjetas, creamos el grafo con la librería mencionada, escogiendo como métrica la inversa de las distancias. El grafo resultante se puede apreciar en la Figura 3.

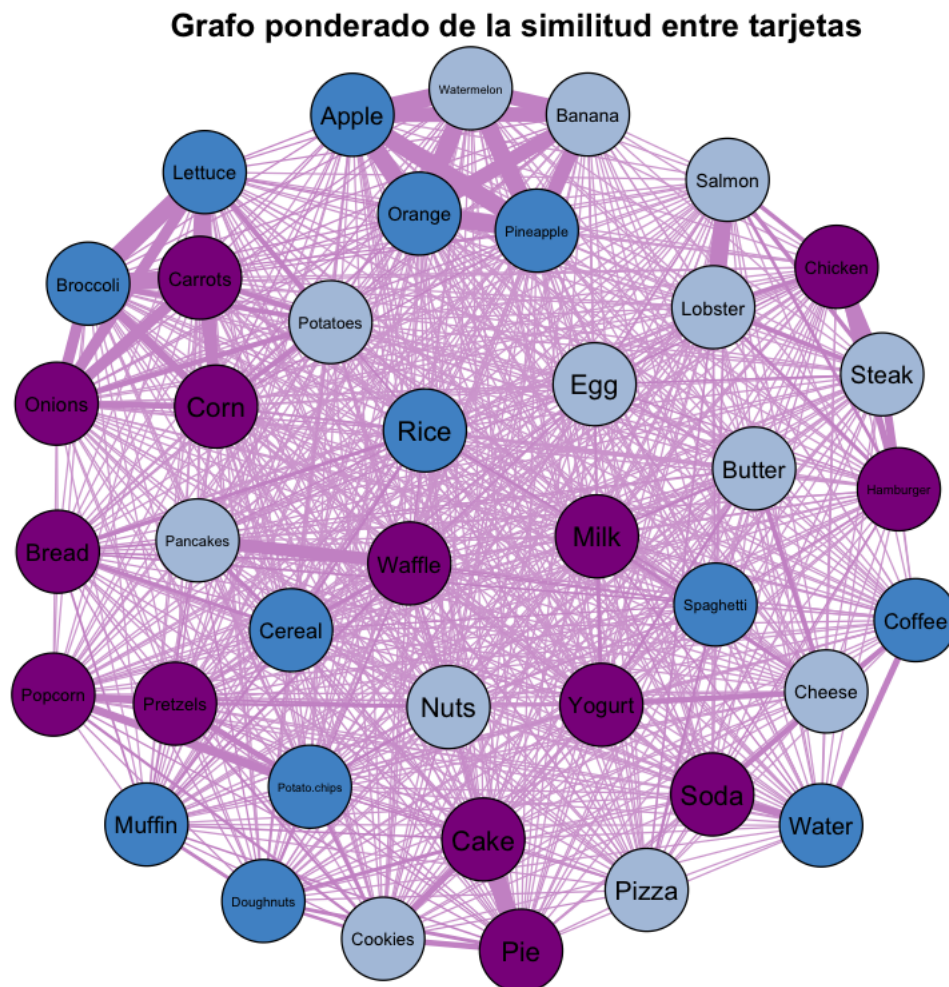


Figura 3: Grafo ponderado de la similitud entre tarjetas

En la representación del grafo, las aristas mas gruesas representan una relación más fuerte, por tanto más similar.

Pregunta 6

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)?

A través del grafo podemos observar fácilmente las tarjetas similares. Para filtrar un poco, consultamos con `which` cuáles son las que tienen la mínima distancia (ver Figura 4).

	row	col
Broccoli	5	1
Lettuce	17	1
Orange	23	2
Pineapple	26	2
Watermelon	39	2
Carrots	1	5
Lettuce	17	5
Pie	25	7
Carrots	1	17
Broccoli	5	17
Salmon	33	18
Apple	2	23
Pineapple	26	23
Watermelon	39	23
Cake	7	25
Apple	2	26
Orange	23	26
Watermelon	39	26
Lobster	18	33
Apple	2	39
Orange	23	39
Pineapple	26	39

Figura 4: Tarjetas más similares

Tenemos, por *clustering*, los siguientes grupos marcados:

- Apple, Watermelon, Banana, Orange y Pineapple.
- Pie, Cakes (y un poco con Cookies).
- Pancakes y Waffle.
- Broccoli, Lettuce, Carrots, Onions y Corn.
- Salmon y Lobster.
- Chicken, Steak y Hamburger.