Gestión de Datos
Exploración gráfica de datos

Luis Antonio Ortega Andrés
Antonio Coín Castro

13 de diciembre de 2020

1. Problema a resolver

En este ejercicio pretendemos explorar visualmente un dataset para obtener información sobre los datos que representa. Concretamente, usaremos el lenguaje R para explorar un dataset[1] resultado de un Card Sorting sobre alimentos. Nos interesa la siguiente información:

1. Tipología y rango de los datos numéricos.
2. Tarjetas más relacionadas entre sí.

2. Lectura y limpieza de datos

En primer lugar, leemos los datos desde el recurso remoto empleando la función *built-in* de R `read.csv`. La estructura de datos principal que usaremos será el dataframe, por lo que también haremos uso de las funciones de R para manipularlos. Como solo nos interesa estudiar los datos numéricos, prescindimos de las columnas *Uniqid, Startdate, Starttime, Endtime, QID* y *Comment*. Para ello usamos la orden `subset` que nos permite quedarnos solo con columnas específicas del dataframe.

3. Estudio de los datos numéricos

Nos preguntamos qué tipos de datos numéricos alberga el dataset, así como el rango de los mismos. Para visualizar esta información, prescindimos momentáneamente de la columna `Category` (que sabemos que es la única restante que no es numérica) y realizamos un histograma de los datos. El resultado obtenido podemos verlo en la Figura 1. Como se puede apreciar, el conjunto está formado únicamente por valores binarios (0 ó 1), y además la proporción de 0s es mucho mayor (casi 10 veces mayor). Podemos concluir entonces que el dataset está configurado de forma que las observaciones están "desagrupadas"; es decir, por cada fila tenemos una observación de un usuario en una categoría concreta, habiendo un 1 en aquellas tarjetas que ha estimado que pertenecen a dicha categoría.

[1] http://cardsorting.net/tutorials/25.csv

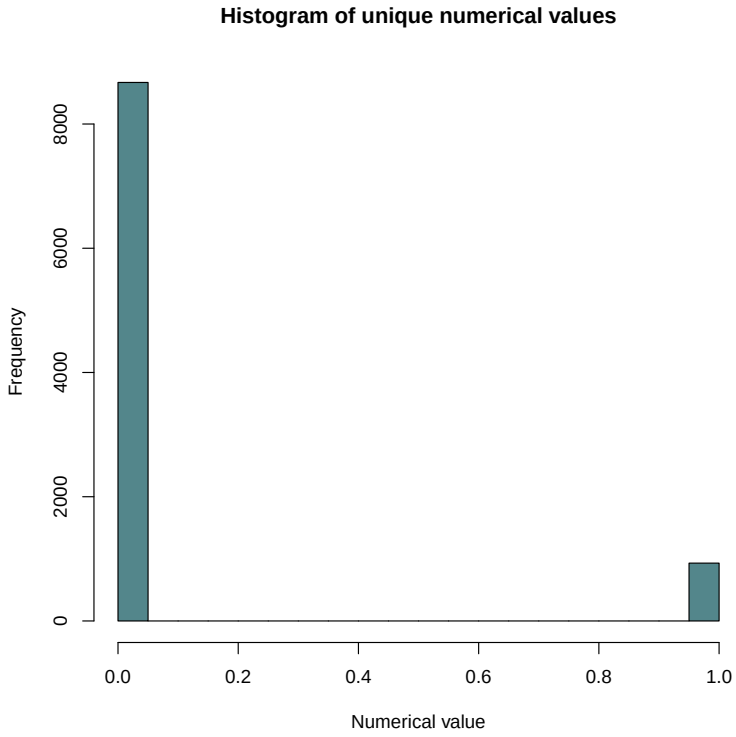

Figura 1: Histograma de los datos numéricos.

4. Similitud entre tarjetas

Nos preocupamos ahora de estudiar cómo se relacionan las diferentes tarjetas a la luz de los datos de los que disponemos. Como primer enfoque, calculamos una **matriz de distancias** entre las mismas, de forma que podamos cuantificar cómo de cerca están las unas de las otras. La métrica que elegimos para hacer esto es la métrica euclídea, de forma que cada componente represente una categoría. Así, la contribución de las categorías donde las tarjetas coincidan será nula, mientras que aquellas para las que no coincidan contribuirán con un factor de 1 en la suma correspondiente. Para calcular esta matriz usamos la función `dist` de R.

Por otro lado, para visualizar la matriz de distancias empleamos un enfoque doble: hacemos un mapa de calor y un dendrograma. Estos dos gráficos suelen ir de la mano, y de hecho el comando `heatmap.2` de la librería `gplots` los pinta a la vez por defecto. El resultado para nuestra matriz de distancias entre tarjetas se observa en la Figura 2.

Lo que nos interesa es medir la similitud de las tarjetas, entendiendo que se parecerán cuando coincidan un número elevado de categorías en las que se ordenan. En el mapa de calor se puede observar cómo hay varios grupos o *clústers* de tarjetas a poca distancia, representadas por el color rojo/naranja oscuro. Esta misma información se puede obtener del dendrograma, viendo como algunas categorías no se dividen hasta casi el final del eje. Ambas representaciones se complementan bastante bien, haciendo que de un vistazo podamos obtener la información que buscamos.

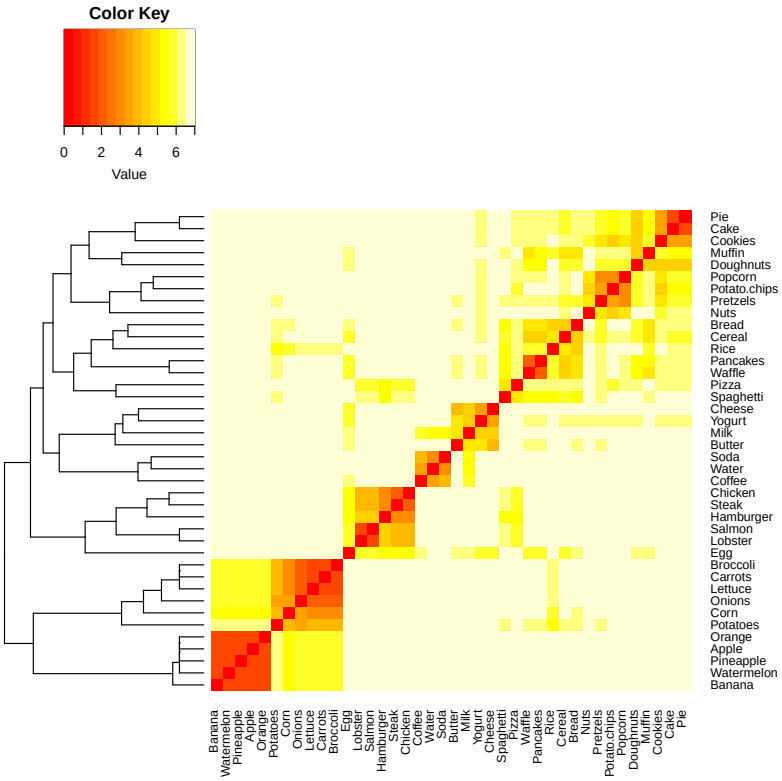

Figura 2: Mapa de calor y dendrograma asociado a las distancias.

Finalmente, podemos estudiar las relaciones entre las tarjetas mediante un grafo, donde los nodos serán las tarjetas y todos estarán relacionados con todos. El peso de las conexiones será la similitud entre las tarjetas que representan, haciendo que visualmente las líneas sean más gruesas cuanto más similares sean. Para visualizar esto utilizamos la función `qgraph` de la librería homónima, teniendo en cuenta que debemos pasar una matriz de similitud y no una matriz de distancia. Como métrica de similitud escogemos, por ejemplo, la matriz de los inversos de las distancias, de forma que sean todas positivas y aumenten conforme menor sea la distancia. El resultado se puede ver en la Figura 3.

5. Tarjetas más relacionadas y significado semántico

Utilizando la representación en grafo de las similitudes, es bastante sencillo localizar las tarjetas más similares. En concreto, tenemos un grupo de tarjetas muy parecidas: **Banana, Apple, Watermelon, Pineapple** y **Orange**. También tenemos otros grupos de tarjetas parecidas, como **Pie** y **Cake**; **Salmon** y **Lobster**; **Pancake** y **Waffle**, y también **Brocoli**, **Carrots** y **Lettuce** (y en este último también en menor medida, **Onion**).

Podemos comprobar que efectivamente todos los grupos que hemos mencionado arriba tienen la distancia minimal entre ellos (de entre todas las tarjetas), utilizando las órdenes `min` y `which` de R. Esta distancia es $\sqrt{2}$, es decir, que las tarjetas más parecidas coinciden en todas las categorías excepto en 2. Además, podemos ver que los grupos obtenidos tienen sentido desde un punto de vista semántico: el primero representa frutas, el segundo postres y cosas dulces; el tercero pescado, y el cuarto verduras y hortalizas.

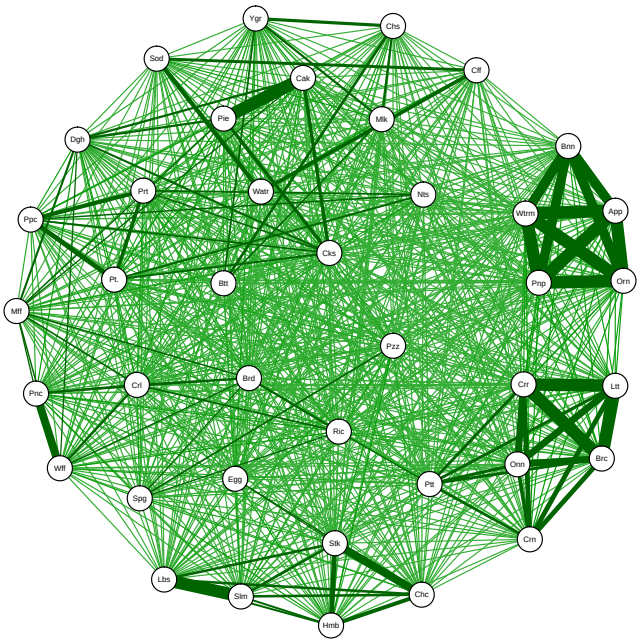

Figura 3: Representación mediante un grafo ponderado de la similitud entre tarjetas.

Como último paso de este análisis, podemos echar un vistazo a las categorías en las que han sido clasificadas algunas de las tarjetas más parecidas. Por ejemplo, los histogramas de categorización para **Apple** y **Banana** pueden verse en la Figura 4, mientras que los de **Pie** y **Cake** está representados en la Figura 5. Como vemos, las categorías más frecuentes para los dos primeros son aquellas que de alguna forma describen frutas, y para los dos últimos son categorías que representan dulces, postres o comida rápida[2]. Sin disponer de información adicional no podemos interpretar las categorías que se han codificado como un número, pero aun así dentro de las tarjetas similares hay algunas categorías en esta situación que han coincidido para 2 o más usuarios.

[2]Notamos que lo que estamos representando en el histograma es la frecuencia con la que se elige cada categoría; no es una comparación de cuándo coincide y cuando no la categorización de las dos tarjetas para un usuario en concreto.

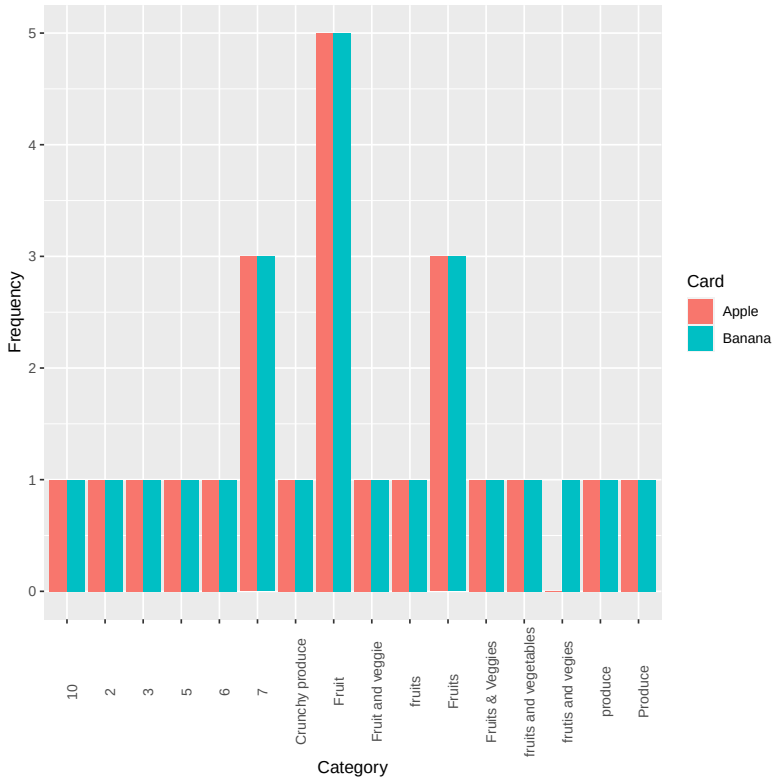

Figura 4: Histogramas de categorías en las que han sido colocados Apple y Banana.

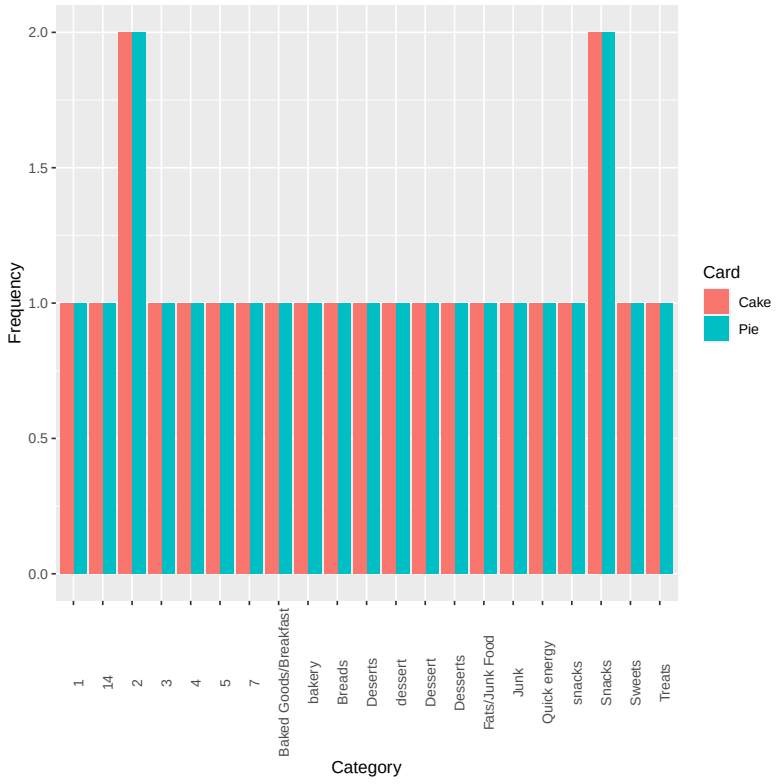

Figura 5: Histogramas de categorías en las que han sido colocados Pie y Cake.