Práctica exploración de datos

Rocío Rico Sanche-Mateos,

Iker Villegas Labairu

4 de diciembre de 2022

Introducción y objetivos

En esta práctica se va a llevar a cabo un estudio sobre el resultado de un *Card Sorting* sobre distintos alimentos. El objetivo será extraer la máxima cantidad de información de los datos y establecer una serie de conclusiones a través de los mimos. Disponemos de un dataset con los datos dispuestos a modo de matriz con 40 variables que hacen alusión a las diferentes tarjetas que representan cada alimento, y 240 entradas correspondientes a las categorías creadas por cada usuario para clasificar los alimentos. Basándonos en dicho dataset, vamos a realizar las tareas que se muestran a continuación.

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

Se nos pide leer un documento de excel a través de un enlace web. Para ello utilizaremos la función de R *read.csv()* que nos permitirá almacenar los datos en una variable lista. Además, al estar el fichero almacenado en una dirección web, utilizaremos la función *url()* dentro de *read.csv()* para

establecer la conexion con dicho archivo .csv.

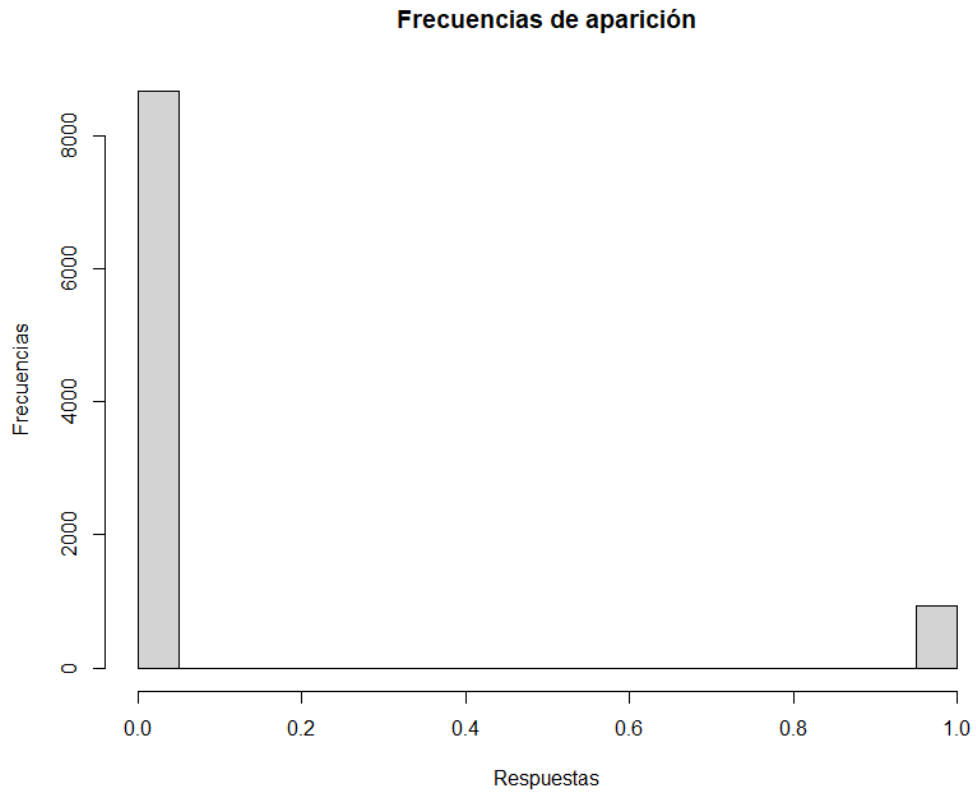
b. Realizar las transformaciones que se consideren convenientes para trabajar de manera efectiva con las categorías y las tarjetas. Se deberá obviar toda la información que no sea de utilidad.

Ahora disponemos de nuestros datos en una tabla con 240 observaciones y 40 variables. Para nuestro análisis tendremos que prescindir de las columnas *Uniquid, Startdate, Starttime, Endtime, QID* y *Comment*. Para eliminarlas utilizaremos la función *select()* (disponible en el paquete *dplyr*) que nos permite seleccionar aquellas variables/columnas que necesitemos, Pasaremos como argumentos, nuestro conjunto de datos y un vector con un signo negativo que contenga el nombre de estas columnas de las cuales vamos a prescindir. De esta forma podremos almacenar en una nueva variable (o sobreescribiendo en la misma), una nueva tabla con las columnas que necesitaremos para nuestro estudio.

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.

Nuestro objetivo en este apartado será estudiar los datos numéricos de nuestro dataset. Para ello, en primer lugar, eliminamos todos aquellos que no son numéricos, que en nuestro caso se corresponderá con la columna *Category*. Así, almacenamos en una nueva variable la tabla que contiene

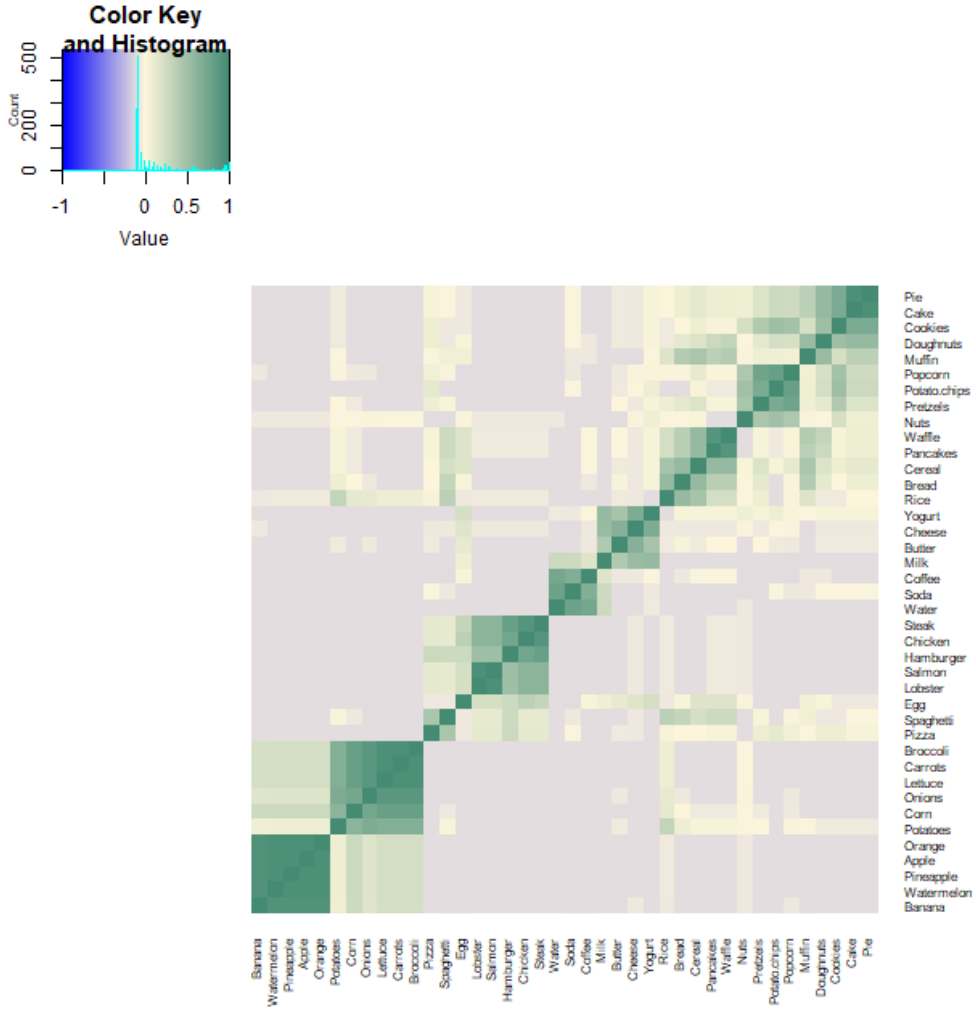
únicamente unos y ceros, correpondientes a la identificación de tarjetas. Este proceso lo llevamos a cabo del mismo modo que hicimos en el apartado anterior utilizando la función *select()*. Ahora dibujaremos el histograma. Nuestros datos numéricos están almacenados en una lista, y queremos que pasen a un formato de vector. Para ello hemos razonado dos formas diferentes: una de ellas, consiste en inicializar (con valor *NULL*) el vector que utilizaremos para la representación, y, a través de un bucle, ir introduciendo las columnas de la tabla. La otra, en cambio, consistirá en pasar la tabla a través de la función *unlist()*, la cuál nos permitirá transformar dicha tabla en un vector y, después, aplicar *as.integer()*, para transformar los elementos del vector a formato int (pues los valores son ceros y unos). En términos de coste computacional, nos será más ventajosa esta segunda opción. Por último, nos queda representar estos datos. Así, recurriremos a la función *hist()*, que nos permite sacar por pantalla, mediante un histograma, las frecuencias de estos datos.



En nuestro cardsorting, para las distintas tarjetas, cada usuario ha introducido las diferentes categorías para llevar a cabo la clasificación, asociando el valor 1 cuando la tarjeta se correspondía con dicha categoría creada; y 0 cuando no. Por este motivo nos encontramos en nuestro histograma tal cantidad de frecuencias de aparición para el dato 0, frente al 1. Esto nos da una pista de que no hay una fuerte relación entre las tarjetas, pues una gran aparición de ceros delata que se han creado una cantidad considerable de categorías diferentes.

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

Para medir la similitud hemos utilizado la función *cor()* de R que estudia hasta qué punto dos variables están relacionadas linealmente, en este caso, al introducir una matriz de elementos devuelve la correlación que hay entre entre las distintas columnas, es decir, entre las distintas tarjetas.
Para poder visualizarlo nos hemos ayudado de la función *heatmap.2* del paquete *gplots* resultando en la gráfica siguiente:

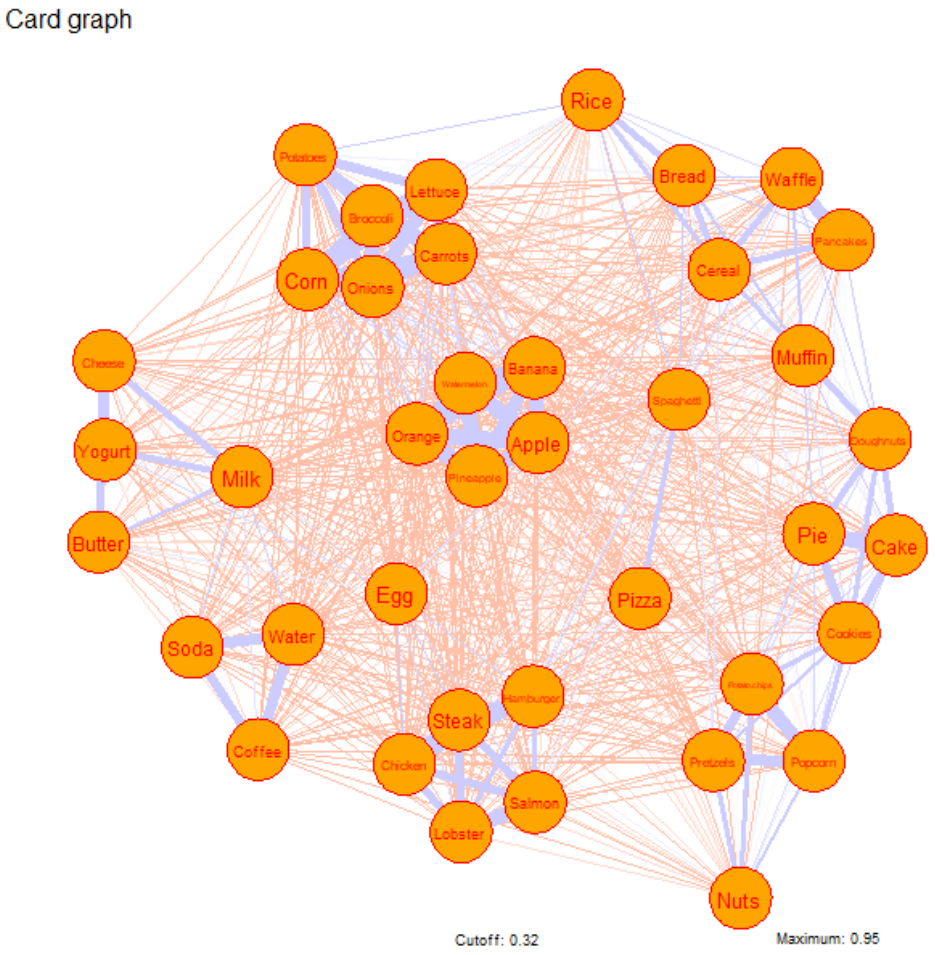


Como se podía predecir se ha creado una matriz $40x40$ donde la diagonal es una línea de unos ya que la correlación de una tarjeta consigo misma es la máxima similitud. A parte de eso, rodeando a la diagonal destacan zonas con alta correlación ya que si nos fijamos están relacionadas semánticamente las tarjetas: dentro del campo de la fruta caben las palabras 'banana', 'pineaple' y 'orange', igual que dentro del campo de las verduras podemos alojar las tarjetas 'corn', 'lettuce' y 'brocoli'. Como estos dos ejemplos están

muy asociados entre sí en el modelo mental colectivo vemos que también hay una correlación no despreciable entre estos campos. Este tipo de relaciones se repiten muchas más veces dentro de este ejercicio de *Card Sorting* como por ejemplo para aquellas tarjetas que se pueden englobar la repostería o la carne, etc. Para un análisis más extensivo tendríamos que irnos fijando detalladamente en las relaciones menos obvias que aparecen fuera de la diagonal como por ejemplo 'rice' y 'corn'.

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.

Para este apartado tendremos que hacer uso del paquete *qgraph*, el cuál nos permitirá representar a través de un grafo las relaciones que hay entre tarjetas. Una vez instalado y cargado el paquete, utilizaremos la función *qgraph* para realizar dicha representación. Pasaremos como argumento la matriz de correlaciones calculada en el apartado anterior.



Observando el grafo, vemos que nuestras conclusiones del apartado anterior se cumplen. Se aprecia una fuerte relación entre las tarjetas que hacen alusión a frutas y verduras. También se diferencian otros grupos con claras relaciones como lacteos, bebidas y carnes y pescados. Mientras que, por el otro lado, etiquetas como *rice* o *nuts*, podrían tener relación para varias categorías. Sin embargo, se aprecia claramente que la tarjeta *egg*, no presenta ninguna relación considerable con el resto de variables a clasificar, lo

cuál nos indica que la mayoría de los usuarios la habrán clasificado en una categoría única.

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

Las tarjetas que están más relacionadas cuentan con un 0,9519134 de correlación, y como se suponía, son aquellas ligadas por su próxima conexión a nivel semántico:

- Verduras:
  - 'Broccoli'y 'Carrots'
  - 'Broccoli' y 'Lettuce'
  - 'Lettuce' y 'Carrots'
- Frutas:
  - 'Orange' y 'Apple'
  - 'Orange' y 'Pineapple'
  - 'Orange' y 'Watermelon'
  - 'Watermelon' y 'Pineapple'
  - 'Watermelon' y 'Apple'
  - 'Apple' y 'Pineapple'
- Repostería:

- 'Pie' y 'Cake'

- Pescadería:
  - 'Salmon' y 'Lobster'

Por tanto tiene sentido esta relación a nivel semántico, pues se pueden clasificar todas dentro de una misma categoría.

Además. vemos que entre todas estás tarjetas con mayor grado de correlación aparecen la mayoría de frutas y verduras, así como 'salmon' y 'lobster', y 'pie' y 'cake'. De este modo, podemos concluir que estos pares de tarjetas son aquellas que los usuarios más relación encuentran entre sí, y por tanto, que en la mayoría de casos se encuentren categorizadas bajo el mismo nombre.