GD: Exploración Gráfica de Datos

Jon Zorrilla y Rubén Martínez

1. ENUNCIADO

El objetivo de este ejercicio consiste en utilizar los métodos exploratorios vistos en teoría para obtener información y conclusiones sobre un conjunto de datos de dominio.

Para ello, se utilizará un dataset relacionado con el resultado de un Card Sorting obtenido a través de una evaluación con distintos usuarios.

El dataset se dispone en forma de matriz (M), de forma que las categorías se sitúan en filas, constituyendo observaciones (n = 240), mientras que las tarjetas se distribuyen en columnas, constituyendo variables (p = 40). Por tanto, M[i,j]=N implica que N usuarios han clasificado la tarjeta j en la categoría i. Para el objetivo de este ejercicio, se puede prescindir de las columnas Uniqid, Startdate, Starttime, Endtime, QID y Comment.

Supongamos que os ponéis en el rol de un Ingeniero de Usabilidad, o en el del evaluador de Experiencia de Usuario, y queréis explorar visualmente el dataset para obtener la siguiente información:

- Tipología y rangos de los datos numéricos.

- Tarjetas más relacionadas entre sí. Entiéndase que, en este ámbito concreto, dos tarjetas se consideran similares si son ordenadas en categorías similares.

2. APARTADOS

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

Para leer el dataset desde el enlace utilizamos la funcionalidad **read.csv** acompañada del url de la base de datos.

```
data <- read.csv(url("http://cardsorting.net/tutorials/25.csv"))
head(data)
```

Description: df [6 × 47]

	Uniqid <int>	Category <chr>	Startdate <chr>	Starttime <chr>	Endtime <chr>
1	2249	Sides	10/8/2014	13:09:10	13:13:10
2	2249	meat	10/8/2014	13:09:10	13:13:10
3	2249	dinners	10/8/2014	13:09:10	13:13:10
4	2249	Snacks	10/8/2014	13:09:10	13:13:10
5	2249	breakfasat	10/8/2014	13:09:10	13:13:10
6	2249	Fruit and ve...	10/8/2014	13:09:10	13:13:10

6 rows | 1-6 of 47 columns

Figura 1: Tabla sin transformaciones

2.2. 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.

Como nuestro interés reside en los datos numéricos, prescindimos de las columnas *Uniqid, Startdate, Starttime, Endtime, QID y Comment*. Para esto, hacemos uso de la orden **subset** que permite obviar las columnas suministradas y quedarnos con las que nos interesen para el estudio.

```
data <- subset(data, select = -c(1,2,3,4,5,6, ncol(data)))
freqs <- data[,2:ncol(data)]
head(freqs)
```

Description: df [6 × 39]

	Apple <int>	Banana <int>	Bread <int>	Brocc... <int>	Butter <int>	Cake <int>	Cereal <int>	Cheese <int>	Chick... <int>
1	0	0	1	0	1	0	0	0	0
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	1	0	1	0
5	0	0	0	0	0	0	1	0	0
6	1	1	0	1	0	0	0	0	0

6 rows | 1-10 of 39 columns

Figura 2: Tabla con información de utilidad

2.3. 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 representar estos datos, separamos los datos se la columna Category y con los datos restantes realizamos un histograma. En este gráfico se puede observar que el conjunto está compuesto por dos valores numéricos, 0 y 1, donde la proporcion de valores 0s es mucho mayor que la de 1s. Esto se explica dado que por cada fila encontramos la observación de un usuario, donde el uno representa las veces que un alimento ha sido declarado frente al 0 que representa las veces que no.

```
#create table
freqs_unq <- data.frame(table(unlist(freqs)))
freqs_unq
#plot histograma
ggplot(freqs_unq, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", color="black", fill=c("aquamarine", "bisque1"))+
  geom_text(aes(label=Freq), vjust=1.6, color="black", size=3)+
  ggtitle("Histograma de las tarjetas") +
  theme(plot.title = element_text(hjust = 0.5, face="plain")) +
  xlab("Data") + ylab("Frecuencia") +
  theme(axis.line = element_line(colour = "black"),
        panel.grid.major = element_blank(),
        panel.grid.minor = element_blank(),
        panel.border = element_blank(),
        panel.background = element_blank())
```

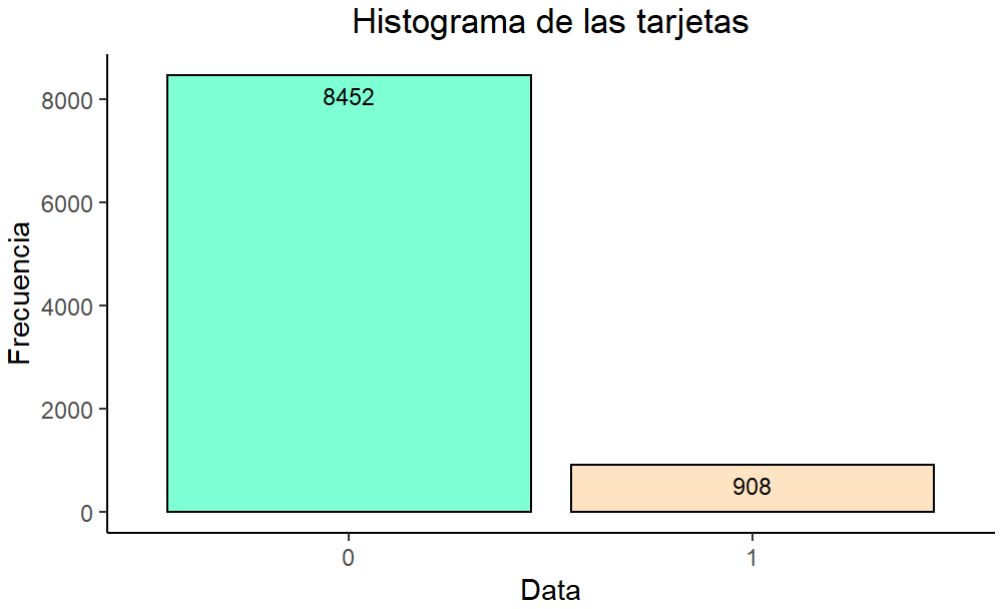

Figura 3: Histograma de tarjetas

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

Para crear una matriz de distancia o similitud de tarjetas, el tipo de visualización más adecuado es un **heatmap**. Para ello hacemos uso de la función **dist** de R y el método **euclidean**. Asimismo, para representar el mapa de calor podemos hacer uso de la función **gplots : : heatmap.2**, la cual ordena por similitud de filas la matriz y permite obtener una representación visual de los clusters de similitud.

El mapa de calor muestra la fuerza de las similitudes entre tarjetas dependiendo de la intensidad del color. Por lo que los grupos donde el color tienda al rojo y anaranjado la relación es mayor.

```
distancia = as.matrix(dist(t(freqs), method="euclidean"))
heatmap.2(distancia, symkey=FALSE, density.info="none", trace="none", dendrogram ="row")
title("Mapa de calor de las tarjetas")
```

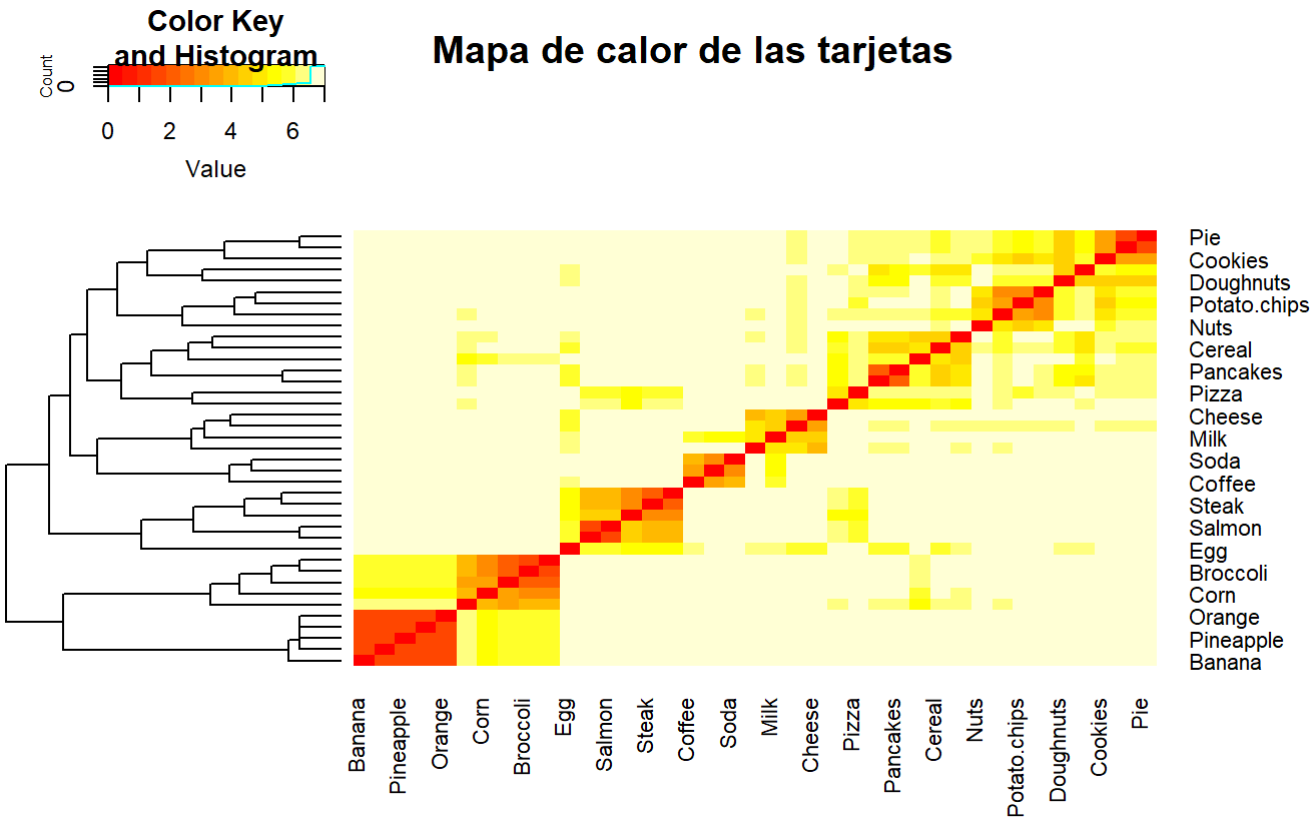

Figura 4: Mapa de calor de las tarjetas

2.5. 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.

El grafo es un elemento gráfico muy útil para poder analizar de un modo más profundo las relaciones entre tarjetas. Para representarlo se hace uso de la librería **qgraph** mediante la métrica de la inversa de las distancias.

En el grafo, las aristas con mayor grosor muestran la relación más fuerte entre variables, mientras que las de menor grosor muestran lo contrario.

```
qgraph(1/(1 + distancia), labels=colnames(distancia), layout="spring", vsize=5, theme="colorl
title("Grafo ponderado de la similitud entre tarjetas")
```

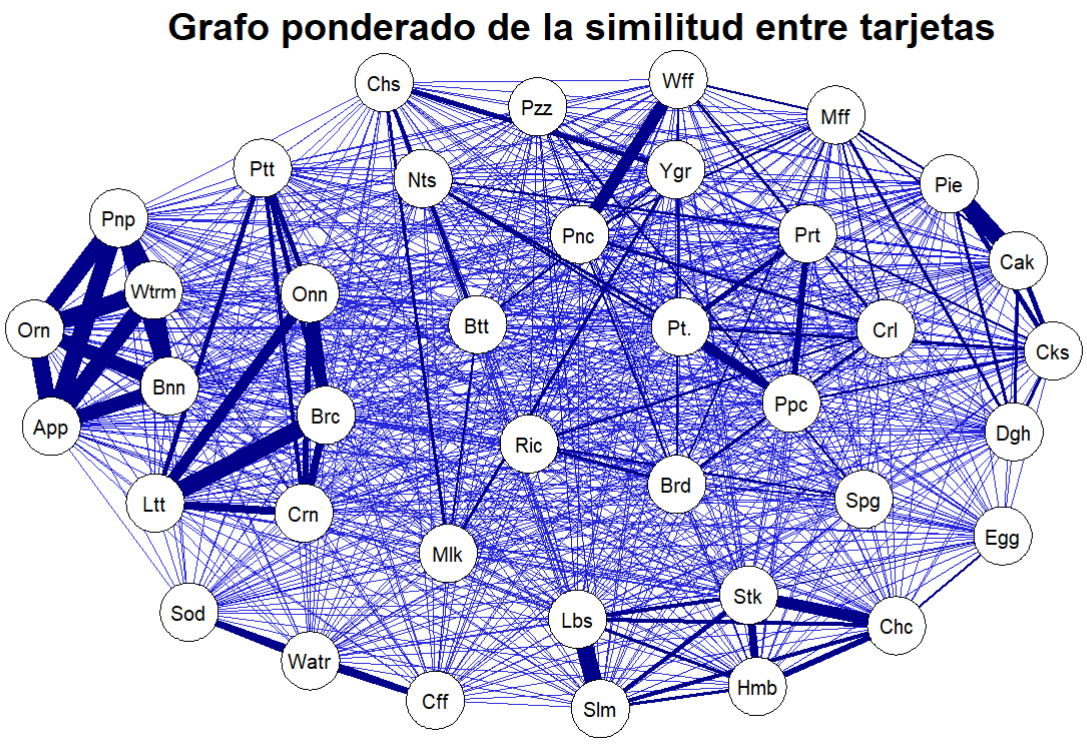

Figura 5: Grafo ponderado de la similitud entre tarjetas

2.6. 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)?

```
cat("Cards with higher similarity:\n")
min_distancia <- min(distancia)
which(as.matrix(distancia)==min_distancia, arr.ind=TRUE)
```

Entre las **distintas tarjetas podemos distinguir grupos que muestran mayores similitudes muestran entre si** son:

- Frutas: Pineapple, Watermelon, Orange, Banana y Apple.
- Tartas: Pie, Cake y Cookies
- Tortas: Waffle y Pancakes.
- Verduras: Lettuce, Broccoli, Onions, Corn y Potatoes.
- Bebidas: Soda, Water y Coffee.
- Carnes y pescados: Lobster, Salmon, Steak, Chicken y Hamburger.
- Snacks: Pretzels, Popcorn y Potato chips.

- "Lacteos": Milk, Butter, Nuts, Cheese, Pizza y Yogurt

Cards with higher similarity:

	row	col
Orange	22	1
Pineapple	25	1
Watermelon	38	1
Lettuce	16	4
Pie	24	6
Broccoli	4	16
Salmon	32	17
Apple	1	22
Pineapple	25	22
Watermelon	38	22
Cake	6	24
Apple	1	25
Orange	22	25
Watermelon	38	25
Lobster	17	32
Apple	1	38
Orange	22	38
Pineapple	25	38

Figura *6*: Tabla de tarjetas con más similitudes

Con este análisis, podemos concluir que las relaciones realizadas por los usuarios tienen sentido dadas las categorías que forman. Para alcanzar estos resultados, ambos métodos, mapas de calor y grafos, son eficaces y complementarios para el analista.