# Informe del Proyecto: Análisis exploratorio de Datos

Tratamiento de los Datos - Curso 2023/2024

Grupo D

2024-02-09

---

## Índice



---

# 1. Introducción

Este proyecto tiene por finalidad realizar un análisis exploratorio de los datos que se han recopilado de un supermercado, en concreto, Mercadona.

A partir de estos datos nos plantearemos unas cuestiones a resolver y realizaremos un estudio a partir de ellas, tratando de extraer conclusiones.

# 2. Materiales

El material utilizado son diferentes tickets de distintos supermercados de una misma cadena.

Una parte de los datos con los que hemos trabajado han sido proporcionados por el profesorado y otra recopilados por los autores del proyecto.

## 2.1. Métodos de importación

Realizamos la importación de los datos, para ello crearemos una lista con las cabeceras, otra con los productos (con sus cantidades y precios) y una última con las últimas líneas.

```
# Cargar la librería readr
suppressWarnings({ # Usamos esta función para que no se muestren los 'warnings'
  library(readr)   # al compilar el documento

# Ruta de la carpeta que contiene los archivos de tickets en formato texto
ruta <- "data"

# Obtener la lista de archivos en la carpeta
archivos <- list.files(path = ruta, full.names = TRUE)

# Crear listas para almacenar las cabeceras, productos y últimas líneas de
# cada archivo
cabeceras <- list()
productos <- list()
ultimas_lineas <- list()

# Recorrer cada archivo y procesar su contenido
for (archivo in archivos) {
  # Leer el contenido del archivo
```

```
  contenido <- read_lines(archivo)

  # Encontrar el índice de la palabra "TOTAL"
  indice_total <- grep("TOTAL", contenido)

  # Extraer las cabeceras (primeras 7 líneas)
  cabecera <- contenido[1:min(7, length(contenido))]

  # Extraer los productos
  productos_temp <- contenido[8:(indice_total - 1)]

  # Extraer las últimas líneas (después de "TOTAL")
  ultimas_lineas_temp <- contenido[(indice_total + 1):length(contenido)]

  # Almacenar las líneas en las listas correspondientes
  cabeceras[[archivo]] <- cabecera
  productos[[archivo]] <- productos_temp
  ultimas_lineas[[archivo]] <- ultimas_lineas_temp
}

})

#Mostrar las primeras líneas para ver el formato de cada lista
head(cabeceras[1])
```

```
## $`data/20231213 Mercadona 4,85 .txt`
## [1] "MERCADONA, S.A.   A-46103834"
## [2] "C/ PROFESOR BELTR\xc1N B\xc1GUENA S/N"
## [3] "46009 VALENCIA"
## [4] "TEL\xc9FONO: 963470267"
## [5] "13/12/2023 18:32  OP: 348873"
## [6] "FACTURA SIMPLIFICADA: 2427-012-371161"
## [7] "Descripci\xf3n P. Unit Importe"
```

```
head(productos[1])
```

```
## $`data/20231213 Mercadona 4,85 .txt`
## [1] "1TORTITAS ARROZ YOGUR 1,60" "1CARAMELOS VITA-C 2,35"
```

```
## [3] "1CAFE LIGHT S/LAC 0,90"
```

```
head(ultimas_lineas[1])
```

```
## $`data/20231213 Mercadona 4,85 .txt`
##  [1] "TARJETA BANCARIA 4,85"
##  [2] "IVA BASE IMPONIBLE (\x80) CUOTA (\x80)"
##  [3] "10% 4,41 0,44"
##  [4] "TOTAL 4,41 0,44"
##  [5] "TARJ. BANCARIA:  **** **** **** 8438"
##  [6] "N.C: 098100860  AUT: 154052"
##  [7] "AID: A0000000041010   ARC: 3030"
##  [8] "MASTERCARD"
##  [9] "Importe: 4,85 \x80  Mastercard"
## [10] "SE ADMITEN DEVOLUCIONES CON TICKET"
```

### 2.2. Método ejemplo

(Aquí el material que hemos obtenido y que métodos vamos a usar para tratarlos)

## 3. Exploración y visualización

Observando nuestros datos, nuestro equipo ha planteado unas ciertas preguntas basadas en el estudio realizado:

1. Siguiendo la distribución de los datos, ¿se puede observar alguna tendencia a la compra de determinados productos específicos?
2. Observando las listas y relacionandolas, ¿Se puede encontrar alguna relación entre la cantidad de productos y si el cliente ha utilizado el parking o no?
3. ¿Existirá una diferencia relevante entre los clientes que han utlizado el mercadona de una localización concreta y aquellos que han asistido a otro de diferente localización?
4. ¿Se puede llegar a relacionar las variables categóricas de nuestro estudio con las variables númericas?¿Entre cuáles podemos establecer relación si se diera el caso?
5. ¿Habrá una cierta tendencia por parte del cliente a comprar más cantidad de productos si los precios son menores o más reducidos?¿El nombre influye en esta relación?
6. En nuestro conjunto de datos, ¿podríamos aplicar el algoritmo del *clustering* para encontrar datos atípicos o excepciones en nuestro estudio? ¿O no sería posible?

7. ¿Dependerá la localización del supermercado, de si los clientes hacen uso o no del parking de este? En caso de que no, ¿Se podría relacionar este hecho con otra variable?
8. ¿Las variables categóricas tienen alguna influencia o dependen unas sobre otras? ¿Y las numericas?
9. Observando el estudio, ¿se podría hacer una predicción de cuál es la media de precio que gastan los clientes en este supermercado o faltarían datos?

## 4. Análisis

(Aquí la realización del estudio)

## 5. Resultados

(Aquí los resultados que hemos obtenido)

## 6. Conclusiones

(Aquí las conclusiones que hemos sacado sobre el proyecto)

## 7. Ejemplos provisionales de figuras con pie de página

**Ejemplo 1:** Vamos a observar un gráfico hecho con código, la Figura 1 (si se le da lleva a la figura, es un enlace).

**Ejemplo 2:** Vamos a observar una imagen, la Figura 2 (si se le da lleva a la figura, es un enlace).

**Ejemplo 3:** Vamos a observar una figura hecha con *látex*, por ejemplo, veremos mejor una ecuación, la Ecuacion 1 (si se le da lleva a la figura, es un enlace).

$$e^{i\pi} + 1 = 0 \tag{1}$$

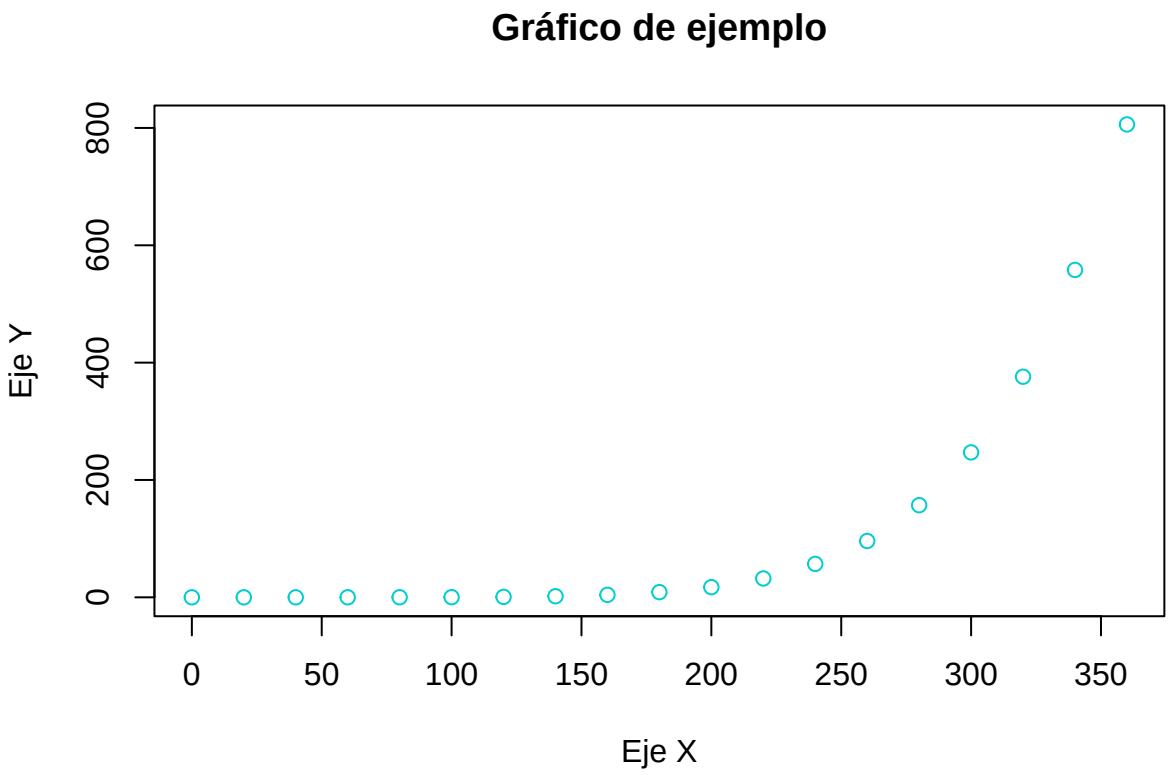


Figura 1: Gráfico de ejemplo para mostrar un pie de imagen

Figura 2: Pie de página de la imagen de ejemplo