ProyectoTD2025

Grupos E

2025-04-13

Introducción

Mercadona y otros supermecados como Lidl, Dia y Comsum han implantado el ticket electrónico y los usuarios que lo deseen pueden recibirlo como documento pdf en su correo electrónico, en lugar de hacerlo en papel.

Como alumnos del GCD nos plantemos el siguiente objetivo:

Queremos desarrollar un programa que permita analizar los tickets para realizar un seguimiento de evolución de precios, compras más habituales, productos más consumidos, supermercado habitual, hora de compra, etc.

Dado que el formato no es el mismo nos vamos a centrar en tickets de Mercadona.

Con este proyecto se espera desarrollar una herramienta capaz de leer y analizar automáticamente tickets de compra de Mercadona, extraídos en formato PDF, para obtener información útil sobre los hábitos de consumo.

Material y métodos

Para este proyecto se ha utilizado un conjunto de tickets de compra proporcionados en formato PDF. Todos los tickets pertenecen, como hemos comentado antes, a supermercados Mercadona.

El análisis de datos se realiza utilizando el lenguaje de programación R.

Como sabemos, la librería `pdftools` y la función `pdf_text` sirven para cargar el contenido del ticket en un vector de texto. Por lo tanto, los tickets serán leídos uno a uno utilizando la función `pdf_text()` de la librería `pdftools`, que permite convertir el contenido del PDF en texto plano. A partir de ese texto, se ha extraído:

- Encabezado (previo a productos) FIJO
- Parte final (después del Total).
- Si hay aparcamiento o no (línea extra)
- Productos: · Venta por unidades · Venta al peso, FRUTA y VERDURA · Venta al peso, PESCADO · Venta LISTO COMER (No hay en los tickets proporcionados)

Importación de datos

A continuacón realizaremos la importación de datos para cargar los tickets en pdf de mercadona:

Primero cargamos todos los datos.

```
library(pdftools)
```

```
## Warning: package 'pdftools' was built under R version 4.4.3
```

```
## Using poppler version 23.08.0
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.4.3
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.2
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
ruta_tickets<- 'data/'
archivos_pdf <- list.files(path = ruta_tickets, pattern = "Mercadona",
                           full.names = TRUE, recursive = TRUE)

# Leer todos los archivos y construir un data frame
df_lineas <- do.call(rbind, lapply(archivos_pdf, function(archivo) {
  paginas <- pdf_text(archivo)
  texto <- paste(paginas, collapse = "\n")       # Unir todas las páginas
  lineas <- unlist(strsplit(texto, "\n"))        # Separar en líneas

  # Crear un data.frame con nombre del archivo, número de línea y texto
  data.frame(
    archivo = basename(archivo),
    linea = seq_along(lineas),
    texto = lineas,
    stringsAsFactors = FALSE
  )
}))

head(df_lineas) #Se muestra el df de los tickets del mercadona
```

```
##                            archivo linea
## 1 20231125 Mercadona 37,76 Ôé¼.pdf     1
## 2 20231125 Mercadona 37,76 Ôé¼.pdf     2
## 3 20231125 Mercadona 37,76 Ôé¼.pdf     3
## 4 20231125 Mercadona 37,76 Ôé¼.pdf     4
## 5 20231125 Mercadona 37,76 Ôé¼.pdf     5
## 6 20231125 Mercadona 37,76 Ôé¼.pdf     6
##                                        texto
```

```
## 1                   MERCADONA, S.A. A-46103834
## 2                    CTRA. GARRUCHA A VERA S/N
## 3                                04630 GARRUCHA
## 4                       TELÉFONO: 950133380
## 5                   25/11/2023 09:09 OP: 78800
## 6       FACTURA SIMPLIFICADA: 2916-010-520925
```

Preguntas planteadas por el profesor

¿Cuáles son los 5 productos, de los vendidos por unidades, con más ventas? ¿Cuántas unidades de cada uno se han vendido?

Si consideramos la categoría de FRUTAS Y VERDURAS. Cuáles son los 5 productos más vendidos? ¿Cuántos kilos se han vendido de cada uno de estos productos?

Si consideramos la categoría de PESCADO. Cuáles son los 5 productos más vendidos? ¿Cuántos kilos se han vendido de cada uno de estos productos?

Muestra mediante un gráco de líneas como ha variado el precio por kilo de las bananas y los plátanos en los tickets disponibles, a lo largo del tiempo.

¿Cuál es la procedencia de los tickets? ¿Qué ciudad/ pueblo tiene un mayor número de tickets?

Muestra mediante un diagrama el número de tickets recogidos cada día de las semana. ¿Si tuvieses que cerrar un día entre semana qué día lo harías?

Preguntas planteadas por nosotros

¿Cuál es la media de ventas para cada día de la semana?

¿Cómo ha evolucionado el precio por unidad o por kilo de un producto específico a lo largo del tiempo? Hacer el gráfico de cómo varía el precio de un producto a lo largo del tiempo.

¿Cuáles son los productos menos vendidos en el conjunto de tickets disponibles?

En una compra, ¿cuál es el producto por unidad que se adquiere en mayor cantidad? (NO FRUTAS Y VERDURAS NI PESCADO)

¿Cuáles son las combinaciones de productos más frecuentes (dos productos que se compren juntos)?

¿Hay estacionalidad en ciertos productos (produtos que se compran más en cierta fecha del año)?

¿Cuántos productos se compran por ticket, en promedio?

¿Cuál es el importe medio por ticket? ¿Cuál es el ticket más caro registrado? ¿Y el más barato?

¿Cuál es la cantidad total de dinero que se obtiene en impuestos cuando se venden alimentos con un 10% de IVA? ¿Y con un 21% y un 5%?

¿A qué horas se suele ir más a comprar en los supermercados? ¿Cuáles son las que menos?

¿Qué método de pago es más frecuente en los tickets: tarjeta o efectivo? Muéstralo mediante un box plot ¿Cuánto se ha gastado en total con cada método de pago?