Mini Proyecto TD 2024

Entrega parcial

Víctor Álvarez Palomares,
Belén Gallego Andreu,
Sergio Martínez Yagüe,
Ferran Medina Mompó,
Carles Pascual i Sivera,
y Úrsula Casaus Fabra

2024-03-29

Introducción

En esta primera parte de la entrega, vamos a importar y a analizar la información que nos aportan una serie de tickets de la compra que se nos proporcionan desde la asignatura.

Carga de librerías y datos necesarios para el análisis

Para empezar, se cargan todas las librerías necesarias para la realización del código. Esto se hace de manera más elegante utilizando el paquete pacman de RStudio. A continuación, se realiza la carga del conjunto de datos, los cuales se encuentran en formato .pdf. No obstante, no podemos importarlos desde este formato, así que vamos a convertirlos a .txt mediante una función que hemos creado en Python con la librería PyPDF2. De esta forma, obtenemos los ficheros .txt para poder importarlos en RStudio y trabajar con ellos a modo de dataframe.

Características generales de los datos

Los dataframes que representan la información de los tickets contienen tres variables: producto, unidades y precio. Las variables primera y la tercera son de tipo "char", mientras que la segunda variable es de tipo "num". El número de registros de cada dataframe depende del número de productos que se hayan comprado según el ticket.

Análisis de "missing data" en nuestro conjunto de interés

Por suerte, en nuestros conjuntos de datos no hay "missing data", ya que se tratan de tickets de la compra en los cuales cada producto tiene un nombre y precio asignados, y los tickets contienen la cantidad de veces que se añade a la compra.

Importación de los datos

Como ya hemos explicado previamente, hemos convertido nuestros ficheros a .txt desde .pdf. Ahora, vamos a crear otra función para separar las cadenas de caracteres de los ficheros y así importar los datos en los dataframes de forma que nos resulte más fácil trabajar con ellos.

```
# Primero que nada, cargamos las librerías:

library(pacman)
p_load(dplyr, stringr)

# Esta función separa los productos de los tickets en tres columnas, unidades,
# producto y precio.

separar <- function(a) {
  a <- iconv(a, to = "UTF-8") # Esta línea hace que no de error al leer
  # caracteres especiales
  productos<-a[8:(length(a)-13)] # Seleccionamos solo las filas con productos y
  # creamos un dataframe
  df<-data.frame(producto=productos )
  df <- df %>%
  mutate(unidades = as.numeric(str_extract(producto, "\\d+"))) %>%
  # Seleccionamos el número de unidades y hacemos una columna
  mutate(precio = (substr(producto, nchar(producto) - 3, nchar(producto))),
         # Seleccionamos el precio y hacemos una columna
         producto = sub("^\\d+", "", producto)) # Borramos las unidades de la
  # columna original
  return(df)
}


a <- readLines("./data/20231218 Mercadona 60,47 Ôé¼.txt")
```

```
## Warning in readLines("./data/20231218 Mercadona 60,47 Ôé¼.txt"): incomplete
## final line found on './data/20231218 Mercadona 60,47 Ôé¼.txt'
```

```
b <- readLines("./data/20231224 Mercadona 37,49 Ôé¼.txt")
```

```
## Warning in readLines("./data/20231224 Mercadona 37,49 Ôé¼.txt"): incomplete
## final line found on './data/20231224 Mercadona 37,49 Ôé¼.txt'
```

```
c <- readLines("./data/20231226 Mercadona 25,83 Ôé¼.txt")
```

```
## Warning in readLines("./data/20231226 Mercadona 25,83 Ôé¼.txt"): incomplete
## final line found on './data/20231226 Mercadona 25,83 Ôé¼.txt'
```

```
d <- readLines("./data/20231230 Mercadona 66,30 Ôé¼.txt")
```

```
## Warning in readLines("./data/20231230 Mercadona 66,30 Ôé¼.txt"): incomplete
## final line found on './data/20231230 Mercadona 66,30 Ôé¼.txt'
```

```
e <- readLines("./data/20240102 Mercadona 70,04 Ôé¼.txt")
```

```
## Warning in readLines("./data/20240102 Mercadona 70,04 Ôé¼.txt"): incomplete
## final line found on './data/20240102 Mercadona 70,04 Ôé¼.txt'
```

```
f <- readLines("./data/20240108 Mercadona 83,73 Ôé¼.txt")
```

```
## Warning in readLines("./data/20240108 Mercadona 83,73 Ôé¼.txt"): incomplete
## final line found on './data/20240108 Mercadona 83,73 Ôé¼.txt'
```

```
g <- readLines("./data/20240109 Mercadona 7,35 Ôé¼.txt")
```

```
## Warning in readLines("./data/20240109 Mercadona 7,35 Ôé¼.txt"): incomplete
## final line found on './data/20240109 Mercadona 7,35 Ôé¼.txt'
```

```
ticket_a <- separar(a)
ticket_b <- separar(b)
ticket_c <- separar(c)
ticket_d <- separar(d)
ticket_e <- separar(e)
ticket_f <- separar(f)
ticket_g <- separar(g)

# Cuando hay algún caracter especial como \ o -, sustituye esa fila por NAs,
# además no funciona para las frutas porque estan en otro formato distinto.
# También falta borrar el precio de la columna productos.
```

Preguntas

A continuación vamos a plantearnos diferentes cuestiones que nos podrían venir a la cabeza a la hora de hacer un análisis de los tickets de la compra.

1) ¿Qué productos son los más comprados?
2) ¿Cuál es el gasto medio de cada compra?
3) ¿Qué nos podría dar a entender la frequencia de compra de cada usuario (compras diarias, semanales...)?
4) ¿Qué productos se compran más según el precio de la compra?
5) ¿Qué días hay compras más grandes?

Además podríamos calcular las distribuiciones de compras por categorías de productos para saber qué tipos de productos son los más comprados. A partir de esto también podríamos obtener la distribución de gastos que hace el consumidor a la hora de realizar la compra.

Asimismo, podríamos plantearnos otras cuestiones como el gasto medio de compra, si se hace el pago con efectivo o tarjeta o incluso si los usuarios tienden a ir en coche o a pie (depende de el tamaño de compra, o incluso comodidad).