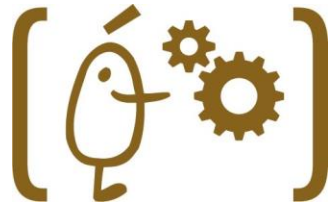


Proyecto Tratamiento de datos

Roberto Millán Brea
romibre@alumni.uv.es

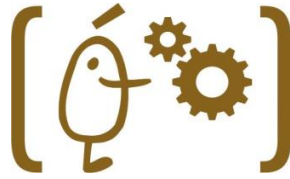


Objetivos del Proyecto



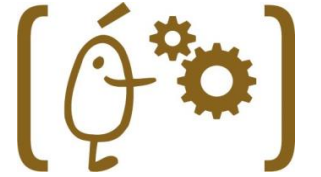
- **Objetivo General:** Analizar los recibos de compra del supermercado Mercadona para extraer insights significativos.
- **Objetivos Específicos:**
 - Importar y limpiar datos de recibos en formato PDF.
 - Estructurar los datos en un formato adecuado para su análisis.
 - Realizar visualizaciones y estadísticas descriptivas sobre los productos vendidos.
 - Interpretar los resultados para obtener conclusiones relevantes.

Metodología y herramientas utilizadas



- **Metodología:**
 - Uso de R para el análisis exploratorio de datos.
 - Extracción y limpieza de datos mediante las librerías pdftools y stringr.
 - Visualización de datos con ggplot2.
 - Análisis estadístico utilizando dplyr y tidyverse.
- **Herramientas:**
 - Librerías específicas de R: pdftools, stringr, ggplot2, dplyr, entre otras.
 - Preparación del entorno de trabajo y gestión de paquetes en R mediante un script inicial.

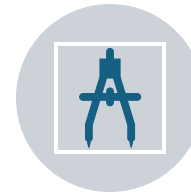
Carga de ficheros y datos



Cargaremos los nombres de los ficheros pdf, de los tickets de compra, en una variable con el objetivo de poder leerlos en un bucle y asignar los datos a un dataframe.



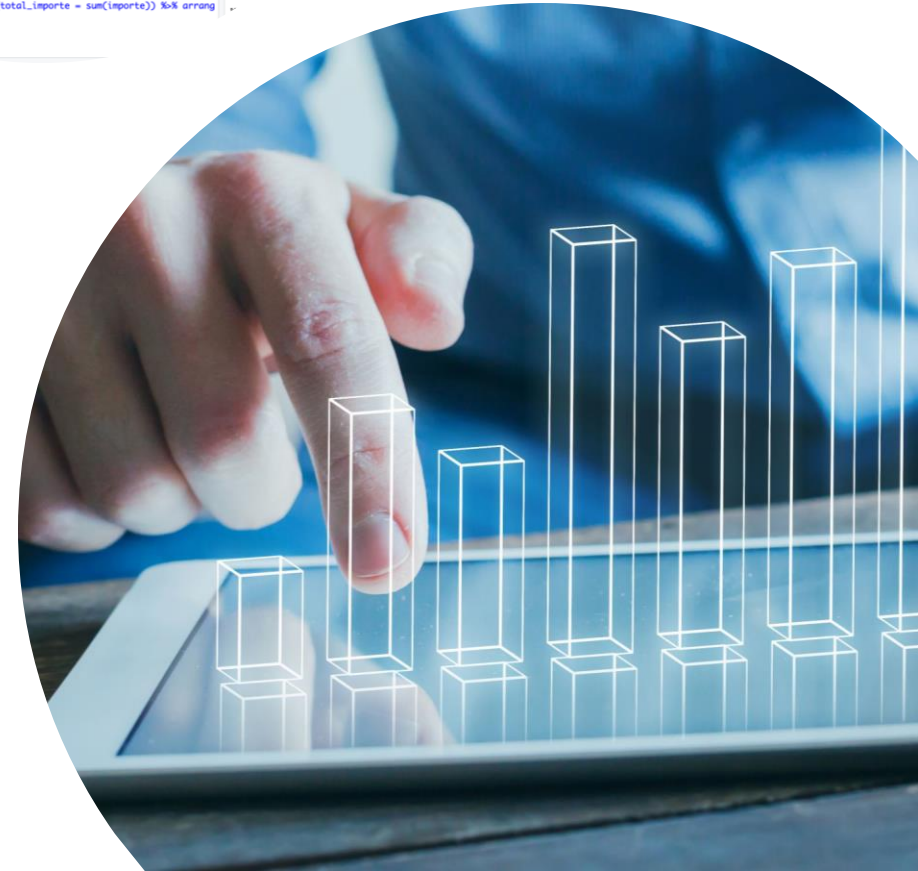
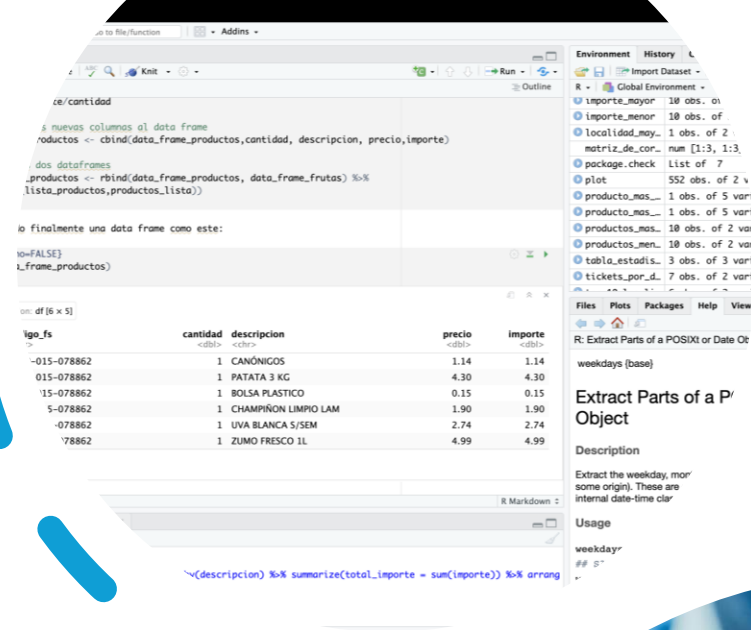
Crearemos un data frame con los datos y los modificaremos para que tengan un formato y clase adecuados. Para leer los archivos PDF usaremos las funciones de la librería pdftools.



Los nombres de las variables serán asignados, y guardaremos las variables como vectores, almacenados en un data frame

Análisis de productos

- Luego de haber realizado ya nuestra carga de ficheros realizaremos el análisis de productos, creando un data frame a partir del original, clasificándolos y depurándolos



Exploración y visualización de datos

- **Exploración Inicial:** Estadísticas descriptivas sobre el conjunto de datos.
- Notemos que los estadísticos nos indican que las compras habitualmente oscilan entre los 30 y los 50 euros de media, aunque al haber outliers, por arriba la media es mayor que la mediana
- Obtención de matrices de correlación

The screenshot displays the RStudio environment with the following components:

- Source Editor:** Contains R code for calculating and printing a correlation matrix.
- Environment:** Lists loaded objects including 'importe_mayor', 'importe_menor', 'localidad_mayor', 'matriz_de_correlacion', 'package.check', 'plot', 'producto_mas_vendido', 'productos_mas_vendidos', 'productos_menos_vendidos', 'tabla_estadistica', and 'tickets_por_destino'.
- Viewer:** Shows the output of the correlation matrix calculation, displaying a 3x3 matrix of Pearson correlation coefficients.
- Console:** Displays the execution of R commands, including the calculation of the correlation matrix and the printing of summary statistics for 'importe_mayor' and 'importe_menor'.

R Code in Source Editor:

```
301  
302  
303  
304  
305 # Calcular la matriz de correlación  
306 matriz_de_correlacion <- cor(data_frame_numeric)  
307  
308 # Mostrar la matriz de correlación  
309 print(matriz_de_correlacion)  
310  
311  
312  
186:1 Analizamos los productos :
```

Correlation Matrix Output:

	total_compra	base_imponible	cuota_iva
Media	46.83860	43.33765	4.118217
Mediana	38.22000	34.67000	2.980000
DesvEst	38.34398	35.60024	9.230350

Correlation Matrix Output:

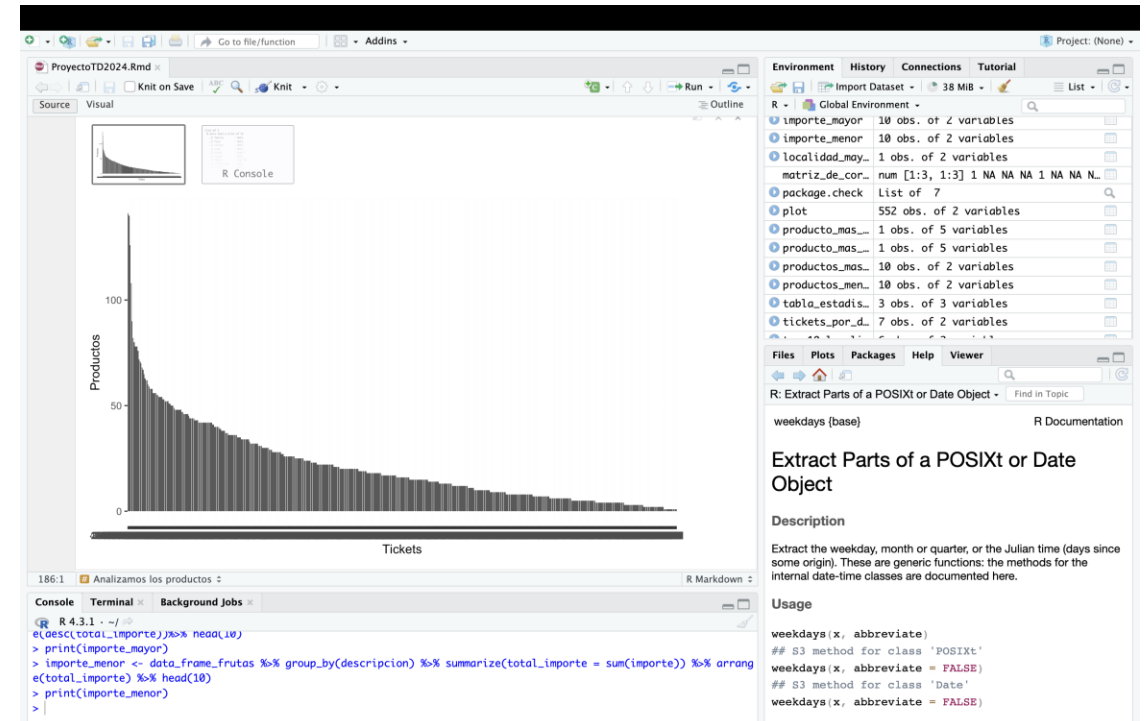
	total_compra	base_imponible	cuota_iva
total_compra	1	NA	NA
base_imponible	NA	1	NA
cuota_iva	NA	NA	1

Console Output:

```
R 4.3.1 ~/  
e(desc(total_importe))>> head(10)  
> print(importe_mayor)  
> importe_menor <- data_frame_frutas %>% group_by(descripcion) %>% summarize(total_importe = sum(importe)) %>% arrange  
e(total_importe) %>% head(10)  
> print(importe_menor)  
>
```

Exploración y visualización

- Aquí podemos ver la relación entre productos y tickets. Notemos que hay tickets con muchos productos, que serán los de las compras caras, y otros con pocos, que serán las baratas.



The screenshot shows the RStudio IDE with the following components:

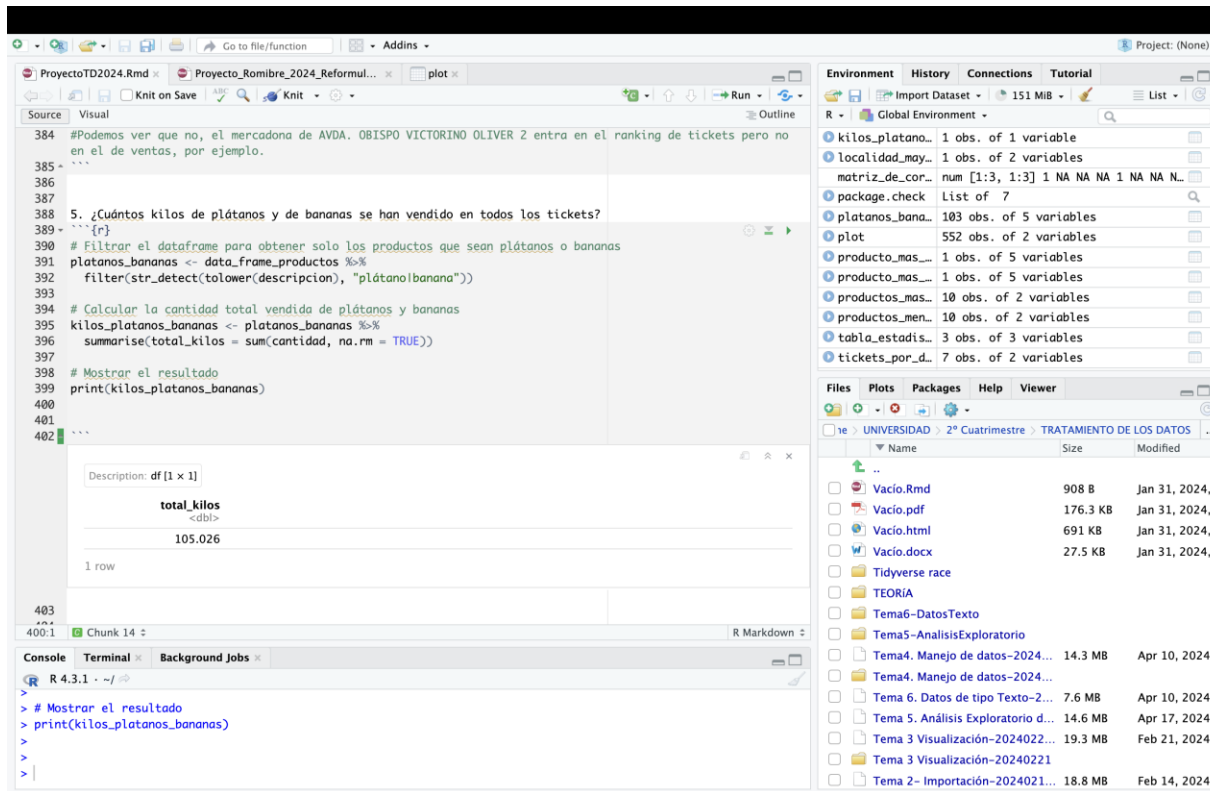
- Source Editor:** Contains R code for analyzing ticket data. Line 339 prints the number of distinct tickets (552). Line 342 asks for the top 10 localities with the most tickets. Lines 343-351 show the code to create a data frame, group by store direction, summarize the number of tickets, and print the top 10.
- Environment Pane:** Lists objects in the Global Environment, including `importe_mayor` (10 obs. of 2 variables), `importe_menor` (10 obs. of 2 variables), `localidad_mayor` (1 obs. of 2 variables), `matriz_de_cor...` (num [1:3, 1:3] 1 NA NA NA 1 NA NA N...), `package.check` (List of 7), `plot` (552 obs. of 2 variables), `producto_mas...` (1 obs. of 5 variables), `producto_mas...` (1 obs. of 5 variables), `productos_mas...` (10 obs. of 2 variables), `productos_men...` (10 obs. of 2 variables), `tabla_estadis...` (3 obs. of 3 variables), and `tickets_por_d...` (7 obs. of 2 variables).
- Viewer Pane:** Displays a tibble with 6 rows and 2 columns: `direccion_tienda` (character) and `num_tickets` (integer). The data is as follows:

direccion_tienda	num_tickets
C/ VIRGEN DEL LOSAR 54	57
AVDA. VALENCIA 41	44
CAMINO VIEJO DE LEGANÉS 58	44
AVDA. OBISPO VICTORINO OLIVER 2	41
AVDA. CASTELLÓN 33	35
C/ JERÓNIMO MONSURIU 60	33
- Console:** Shows the execution of R code, including `e(desc(total_importe))`, `print(importe_mayor)`, `importe_menor <- data_frame_frutas %>% group_by(descripcion) %>% summarize(total_importe = sum(importe)) %>% arrange(desc(total_importe)) %>% head(10)`, and `print(importe_menor)`.

Exploración y visualización de datos

- Aquí podemos ver la cantidad de tickets que hay, así como el top-10 de más tickets. Notemos que en las localidades más céntricas o concurridas, como podrían ser avenidas, hay más tickets
- Aunque en otra pregunta respondemos que tener más tickets, no siempre conlleva tener más productos

Exploración y visualización de datos



The screenshot displays the RStudio environment with the following components:

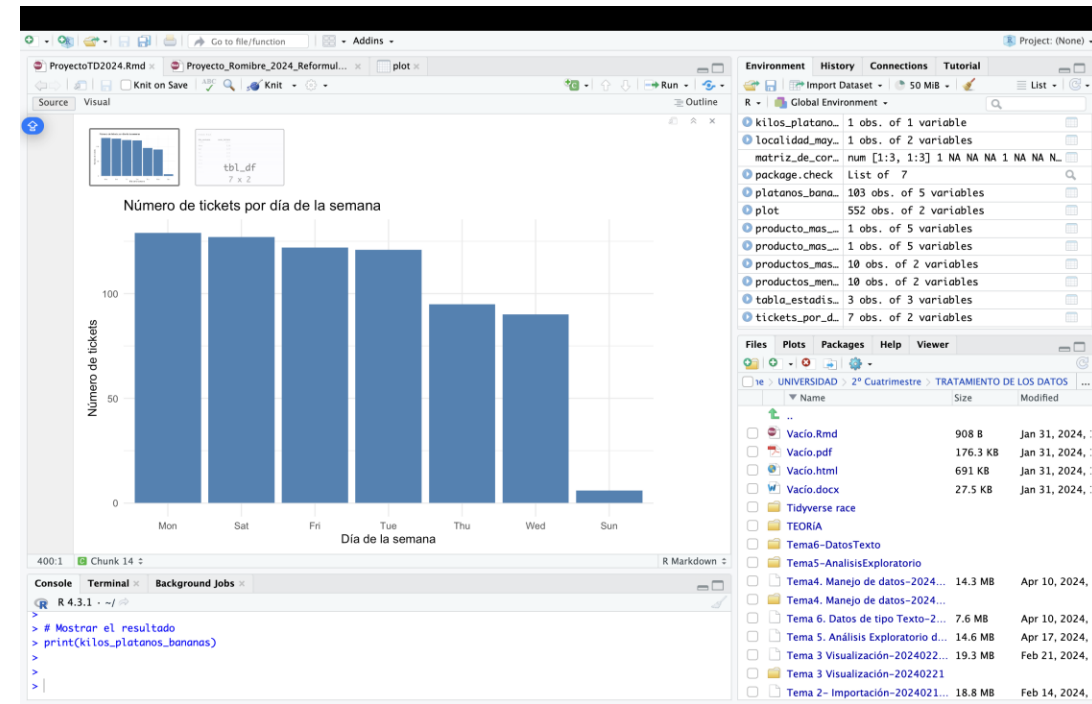
- Source Editor:** Contains R code for data analysis. The code filters a dataset for 'plátanos' and 'bananas', calculates the total kilograms sold, and prints the result.
- Environment:** Lists objects in the Global Environment, including 'kilos_platanos...', 'localidad_may...', 'matriz_de_cor...', 'package.check', 'platanos_bana...', 'plot', 'producto_mas...', 'productos_mas...', 'productos_men...', 'tabla_estadis...', and 'tickets_por_d...'.
- Files:** Shows a list of files in the '2º Cuatrimestre' directory, including 'Vacio.Rmd', 'Vacio.pdf', 'Vacio.html', 'Vacio.docx', 'Tidyverse race', 'TEORIA', 'Tema6-DatosTexto', 'Tema5-AnalisisExploratorio', 'Tema4. Manejo de datos-2024...', 'Tema 6. Datos de tipo Texto-2...', 'Tema 5. Análisis Exploratorio d...', 'Tema 3 Visualización-2024022...', 'Tema 3 Visualización-20240221', and 'Tema 2- Importación-2024021...'.
- Console:** Shows the output of the R code, including the command 'print(kilos_platanos_bananas)' and the result '105.026'.
- Summary Table:** A table with 1 row and 1 column, showing the total kilograms sold for 'plátanos y bananas'.

total_kilos
105.026

Aquí respondemos por ejemplo a la pregunta de cuántos kilos de plátanos se han vendido en total:

Exploración y visualización de datos

- También nos hemos hecho más preguntas para un análisis más exhaustivo de los datos como: Qué días se han obtenido más tickets. Como decimos en el Rmd, se nota una clara tendencia a la baja en ese sentido según el pasar de la semana. Veámoslo:



Resultados y Conclusiones

- **Hallazgos Clave:**
- Análisis de productos más vendidos y menos vendidos

Hemos visto que el producto más caro por ejemplo era el jamón.

Esos productos y los más exclusivos se venden menos. Otra tendencia es que en horario laboral la gente compra más que en horario festivo o de descanso, tanto en horas, como en días

Gracias por la atención

The screenshot displays the RStudio environment. The main editor window shows R code for data manipulation. The Environment pane on the right lists objects like `kilos_platano`, `localidad_may`, `matriz_de_cor`, `package.check`, `plot`, `producto_mas...`, `productos_mas...`, `productos_men...`, `tabla_estadis...`, and `tickets_por_d...`. The Files pane at the bottom shows a directory structure for a project named 'TRATAMIENTO DE LOS...'. The Terminal window at the bottom shows the command `str(kilos_platanos_bananas)`.

```
8. ¿Cuál es el producto más caro, y el más barato?  
```{r}  
producto_mas_caro <- data_frame_productos %>%
 arrange(desc(precio)) %>%
 head(1)
print(producto_mas_caro)

#Si incluimos al parking, este será el producto más barato, ya que hay parking gratuito en el Mercadona
producto_mas_barato <- data_frame_productos %>%
 arrange(precio) %>%
 head(1)
print(producto_mas_barato)
```
```

| codigo_fs | cantidad | descripcion | precio | importe |
|-------------------|----------|----------------------|--------|---------|
| 1 2465-011-597185 | 1 | JAMON DE CEBO IBE.G. | 129 | 129 |

Proyecto Tratamiento de datos

Roberto Millán Brea
romibre@alumni.uv.es

