ProyectoTD2025

Andrea Fu Castelló Sala Sergio Del Carme Moreno María Martínez Marí
Jorge Navarro Rodríguez Iván Pérez Alonso Edurne Serigó Troyano

2025-05-13

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 la 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.

Carga de librerías y datos necesarios para el análisis

En cuanto a datos necesarios para la realización del proyecto, destacamos la carga de librerías como: pdftools, knitr, ggplot2, dplyr, tidyverse y stringr para el desarrollo de nuestro código. Gracias a estas librerías, seremos capaces de poder utilizar las funciones que nos permitan almacenar la información contenida en los tickets en nuestro dataframe.

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`. 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) y Venta al peso (PESCADO)

Importación de datos

Este bloque de código realiza la importación y estructuración de los tickets en formato PDF de Mercadona para su posterior análisis. El objetivo es transformar los tickets en un data frame legible y manipulable en R, donde cada fila representa una línea del ticket, junto con el nombre del archivo (ticket) y el número de línea correspondiente.

Primero se define la ruta donde están almacenados los archivos PDF y luego se genera una lista de todos los archivos dentro de esa ruta que contienen la palabra "Mercadona" en su nombre.

A continuación, aplicamos una función de lectura a cada uno de los archivos PDF encontrados. Dentro de esta función, pdf_text() se encarga de leer el contenido del PDF como texto, devolviendo una lista de páginas. Todas las páginas de un ticket se unen en un solo texto, y luego este texto se divide en líneas individuales usando el carácter de salto de línea.

Por cada archivo, se construye un pequeño data frame con tres columnas: archivo, linea y texto.

Luego, todos estos pequeños data frames se combinan en uno solo, formando el objeto df_lineas, que contiene todas las líneas de todos los tickets.

Por último, se muestra una tabla con las primeras 52 líneas de este data frame (el primer ticket del data frame) para visualizar cómo se ve el contenido de los tickets ya estructurado. Esta visualización permite confirmar que la lectura fue correcta y que los datos están listos para su análisis.

Table 1: Tabla de tickets de mercadona

archivo	linea	texto
20231125 Mercadona 37,76 €.pdf	1	MERCADONA, S.A. A-46103834
20231125 Mercadona 37,76 €.pdf	2	CTRA. GARRUCHA A VERA S/N
20231125 Mercadona 37,76 €.pdf	3	04630 GARRUCHA
20231125 Mercadona 37,76 €.pdf	4	TELÉFONO: 950133380
20231125 Mercadona 37,76 €.pdf	5	25/11/2023 09:09 OP: 78800
20231125 Mercadona 37,76 €.pdf	6	FACTURA SIMPLIFICADA: 2916-010-520925
20231125 Mercadona 37,76 €.pdf	7	
20231125 Mercadona 37,76 €.pdf	8	
20231125 Mercadona 37,76 €.pdf	9	
20231125 Mercadona 37,76 €.pdf	10	
20231125 Mercadona 37,76 €.pdf	11	Descripción P. Unit Importe
20231125 Mercadona 37,76 €.pdf	12	5 DONACIÓN 1,00 5,00
20231125 Mercadona 37,76 €.pdf	13	1 BARREÑO 2,20
20231125 Mercadona 37,76 €.pdf	14	3 PALO ANTIDESLIZANTE 1,90 5,70
20231125 Mercadona 37,76 €.pdf	15	1 MULTIUSOS 2,55
20231125 Mercadona 37,76 €.pdf	16	2 AGUA MINERAL 0,73 1,46
20231125 Mercadona 37,76 €.pdf	17	2 B.BASURA EXT.C.FÁCIL 1,60 3,20
20231125 Mercadona 37,76 €.pdf	18	1 ESCURRE FÁCIL 2,85
20231125 Mercadona 37,76 €.pdf	19	1 CUBO FREGAR C/RUEDAS 3,85
20231125 Mercadona 37,76 €.pdf	20	1 FRIEGASUELOS PINO 0,95
20231125 Mercadona 37,76 €.pdf	21	2 FIBRAS CRUZADAS 2,10 4,20
20231125 Mercadona 37,76 €.pdf	22	1 LOTE BAYETAS SUAVES 1,20
20231125 Mercadona 37,76 €.pdf	23	1 FREGONA HILO MIR. 2,80
20231125 Mercadona 37,76 €.pdf	24	1 LOTE 3 BAYETAS MICRO 1,80
20231125 Mercadona 37,76 €.pdf	25	1 PARKING 0,00
20231125 Mercadona 37,76 €.pdf	26	ENTRADA 08:28 SALIDA 09:08
20231125 Mercadona 37,76 €.pdf	27	TOTAL (€) 37,76
20231125 Mercadona 37,76 €.pdf	28	TARJETA BANCARIA 37,76
20231125 Mercadona 37,76 €.pdf	29	

archivo	linea	texto
20231125 Mercadona 37,76 €.pdf	30	IVA BASE IMPONIBLE (€) CUOTA (€)
20231125 Mercadona 37,76 €.pdf	31	10% 1,33 0,13
20231125 Mercadona 37,76 €.pdf	32	21% 25,87 5,43
20231125 Mercadona 37,76 €.pdf	33	TOTAL 27,20 5,56
20231125 Mercadona 37,76 €.pdf	34	
20231125 Mercadona 37,76 €.pdf	35	DONACIÓN A BANCO DE
20231125 Mercadona 37,76 €.pdf	36	ALIMENTOS NO REEMBOLSABLE 5,00
20231125 Mercadona 37,76 €.pdf	37	
20231125 Mercadona 37,76 €.pdf	38	TARJ. BANCARIA: **** **** **** 4422
20231125 Mercadona 37,76 €.pdf	39	N.C: 072700743 AUT: 612848
20231125 Mercadona 37,76 €.pdf	40	AID: A0000000031010 ARC: 3030
20231125 Mercadona 37,76 €.pdf	41	
20231125 Mercadona 37,76 €.pdf	42	
20231125 Mercadona 37,76 €.pdf	43	VISA CREDITO/DEB
20231125 Mercadona 37,76 €.pdf	44	Importe: 37,76 € VISA DEBITO
20231125 Mercadona 37,76 €.pdf	45	
20231125 Mercadona 37,76 €.pdf	46	
20231125 Mercadona 37,76 €.pdf	47	
20231125 Mercadona 37,76 €.pdf	48	
20231125 Mercadona 37,76 €.pdf	49	SE ADMITEN DEVOLUCIONES CON TICKET
20231125 Mercadona 37,76 €.pdf	50	
20231125 Mercadona 37,76 €.pdf	51	DISPONE DE 20 MINUTOS
20231125 Mercadona 37,76 €.pdf	52	PARA RETIRAR SU VEHÍCULO

Preguntas planteadas por el profesor

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

Queremos identificar los cinco productos más vendidos en términos de unidades. Para ello, debemos analizar aquellas líneas de los tickets que corresponden a productos vendidos por cantidad (unidades) y no por peso o medida. Este tipo de productos suele aparecer en las líneas que comienzan por un número, indicando la cantidad vendida.

Comenzamos filtrando del dataframe original (df_lineas) únicamente las líneas que comienzan por un número. Esto lo haremos mediante una expresión regular que identifique líneas que comienzan con uno o más dígitos seguidos de un espacio. Después, extraeremos las partes clave del texto: cantidad, nombre del producto y, si aparece, el precio total.

Una vez extraídos los datos, creamos un nuevo dataframe con dos columnas: nombre del producto y cantidad vendida. A continuación, agrupamos los productos, sumamos las cantidades y seleccionamos los cinco con mayor número de unidades vendidas. Por último, mostramos los resultados.

Table 2: Top 5 productos vendidos por unidades

producto	unidades_vendidas
LECHE DESNAT. CALCIO	24
QUESO LONCHAS CABRA	20
COPOS DE AVENA	16
BARRA CAMPESINA	14
GARBANZO M.COCIDO	14

Como podemos observar en la tabla, los cinco productos más vendidos por unidades son: leche desnatada con calcio (24 unidades), queso lonchas cabra (20 unidades), copos de avena (16 unidades), barra campesina(14 unidades) y garbanzo m.cocido (14 unidades).

2. 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?

En este análisis queremos centrarnos exclusivamente en los productos pertenecientes a la categoría de frutas y verduras. Para ello, se parte de una lista de palabras clave que identifican nombres comunes de estos productos. El objetivo es identificar las líneas de los tickets que mencionan alguno de estos productos y extraer la cantidad vendida en kilos.

Primero buscamos en el dataframe df_lineas todas las líneas que contienen alguna de estas palabras clave, ignorando mayúsculas o minúsculas. Luego, recorremos una a una esas líneas e intentamos extraer la cantidad en kilos asociada. Si no encontramos la cantidad en la línea principal, revisamos la línea siguiente por si la información estuviera allí. Este valor se convierte a formato numérico unificando el uso de comas y puntos.

Después, limpiamos y estandarizamos los nombres de los productos, convirtiéndolos a mayúsculas y eliminando posibles espacios o caracteres extraños. A continuación, agrupamos por nombre de producto y sumamos los kilos vendidos para cada uno. Finalmente, mostramos los cinco productos más vendidos en esta categoría.

Table 3: 5 frutas y verduras más vendidas (kg)

producto	kilos_totales
NARANJA	82
PATATA	75
MANDARINA	31
CEBOLLA	22
PEPINO	20
MELON PIEL SAPO	16

Como podemos observar en la tabla, los cinco productos más vendidos dentro de la categoría de frutas y verduras son: naranja con 82.062Kg, patata con 74.800Kg, mandarina con 31.186Kg, cebolla con 22.000Kg y pepino con 19.624Kg.

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

En este caso nos centramos exclusivamente en los productos pertenecientes a la categoría de pescado y marisco. Para identificar estos productos, definimos una lista de palabras clave. Estas palabras nos permiten buscar en el dataframe df_lineas aquellas líneas de texto que contienen referencias a este tipo de productos.

A continuación, para cada línea detectada, extraemos el nombre del producto y buscamos la cantidad de kilos que se ha vendido, tanto en la línea actual como, si es necesario, en la línea siguiente. El peso se convierte en formato numérico estandarizado. Posteriormente, se limpia y normaliza el texto del nombre del producto (todo en mayúsculas, sin espacios sobrantes ni caracteres innecesarios).

Por último, agrupamos los datos por nombre de producto y sumamos los kilos vendidos para cada uno. Se ordenan de mayor a menor y se extraen los cinco productos más vendidos en kilos.

Table 4: 5 pescados más vendidos (kg)

producto	kilos_totales
BACALADILLA	7,4
RODABALLO	3,7
SEPIA LONJA	3,6
DORADA	2,9
SEPIA FRESCA	2,6
SEPIA SUCIA REFRIG	2,4

Como podemos observar en la tabla, los cinco productos de la categoría pescado más vendidos por kilos son:bacaladilla con 7,4Kg, rodaballo con 3.672Kg, sepia lonja con 3.584Kg, dorada con 2.934Kg y sepia fresca con 2.642Kg.

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

Para analizar la evolución temporal del precio por kilo de plátanos y bananas, se realiza un filtrado de todas las líneas del dataframe df_lineas que contengan alguna de esas dos frutas. Una vez identificadas estas líneas, se extrae el nombre del archivo (ticket) correspondiente para localizar la fecha de cada transacción.

Las fechas de los tickets se extraen buscando la línea número 5 (que contiene la fecha en cada ticket) y emparejándola con los tickets que contienen banana o plátano. Esta fecha se convierte a formato de fecha para poder utilizarla en un gráfico temporal.

Luego, se obtiene el precio por kilo. Para ello, se asume que el precio aparece en la línea siguiente a donde aparece el nombre del producto, por lo que se extrae el texto de esa línea y se filtra con una expresión regular que busca el formato típico de precio: dos decimales y la etiqueta "€/kg". Posteriormente, se limpia y transforma este texto a un valor numérico decimal para poder graficarlo.

Finalmente, se normaliza el nombre del producto para identificar si se trata de BANANA o PLÁTANO, y se genera un gráfico de líneas que representa cómo ha variado el precio por kilo de cada fruta a lo largo del tiempo.

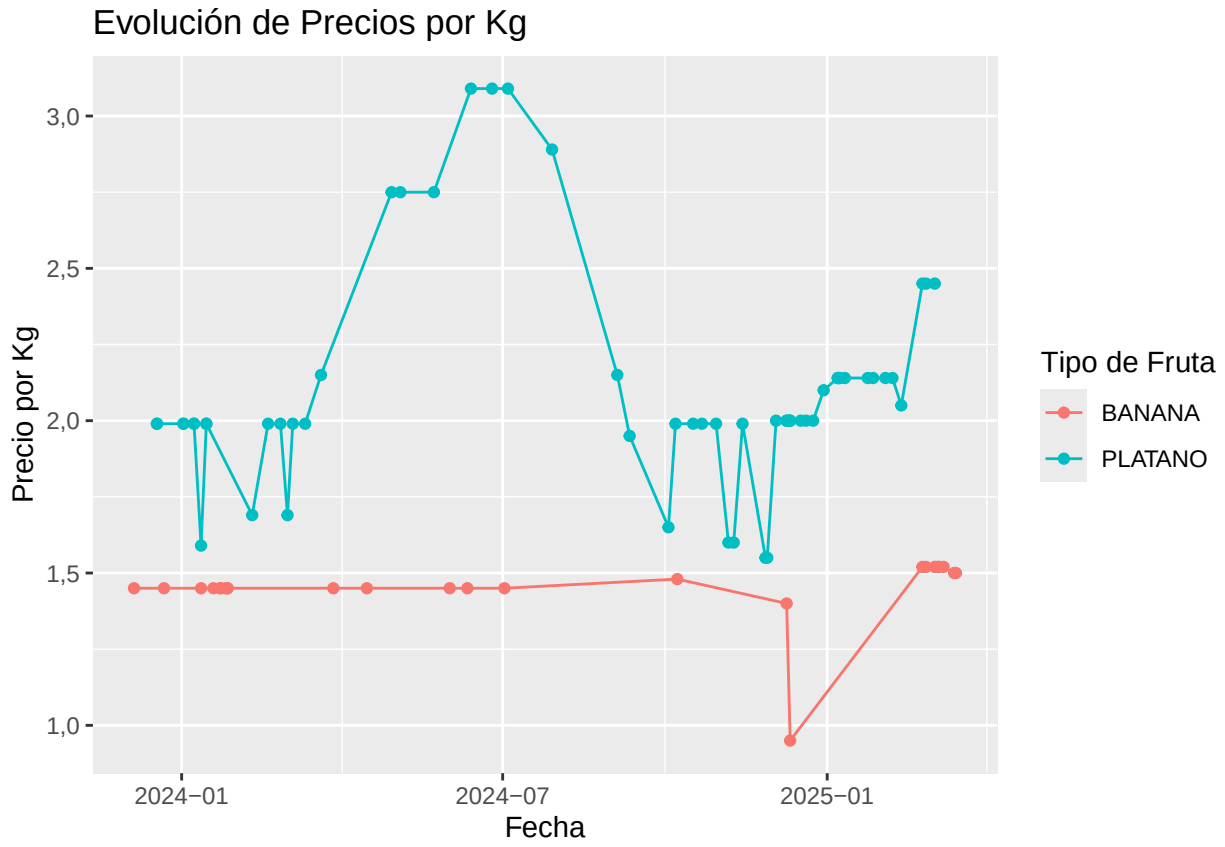


Figure 1: Evolución del precio por kg del platano y la banana

El gráfico permite visualizar de manera clara cómo ha cambiado el precio por kilo de plátanos y bananas en las distintas fechas en las que fueron comprados. Se observan posibles fluctuaciones de precio a lo largo del tiempo, lo que puede ser útil para analizar tendencias de mercado o patrones de compra.

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

Para responder a esta pregunta, se analiza la línea número 3 de cada ticket, ya que en ese campo se encuentra la información de la procedencia, es decir, la ciudad o pueblo en el que se emitió el ticket.

Primero, se extrae todo el contenido de la línea 3 del dataframe df_lineas, que contiene el texto correspondiente a la dirección o localidad. Luego, se limpia ese texto eliminando espacios en blanco al principio y al final, así como cualquier número, con el objetivo de quedarse únicamente con el nombre del municipio o localidad.

Una vez obtenidos los nombres limpios de las ciudades o pueblos, se crea una tabla de frecuencias que cuenta cuántas veces aparece cada uno. Esta tabla muestra el número de tickets por localidad, lo que permite saber qué ciudad o pueblo es el que ha generado más tickets en el conjunto de datos.

Finalmente, se ordena la tabla de manera descendente, de modo que la ciudad con mayor número de tickets aparece en la parte superior de la tabla, y se muestra en un formato presentable.

Table 5: Procedencia de los tickets

	procedencia	Freq
17	VALENCIA	130
1	ALBORAIA/ALBORAYA	50
7	BURJASSOT	26
10	MURO DE ALCOY	24
2	ALCOI/ALCOY	22
6	BUÑOL	12
5	BENIJOFAR	9
3	ALGINET	5
4	BENIFAIO	4
9	GARRUCHA	4
18	VERA	3
13	SAGUNT/SAGUNTO	2
8	GANDIA	1
11	ONTINYENT	1
12	REQUENA	1
14	SAN ANTONIO DE BENAGEBER	1
15	SANT JOSEP DE SA TALAIA	1
16	SANTA EULÀRIA DES RIU	1

En la tabla podemos ver el resultado de la procedencia de los tickets, siendo valencia la ciudad más frecuente con 130 tickets, seguido de alboraia con 50 tickets y después burjassot con 26 tickets.

6. 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?

Para responder a esta pregunta, se analizan las fechas de los tickets con el objetivo de identificar en qué día de la semana se ha generado cada uno, y así contar cuántos tickets corresponden a cada día.

Primero, se filtran las líneas correspondientes a la fecha del ticket, que se encuentran en la línea 5 del dataframe df_lineas. Luego, se extrae la fecha en formato dd/mm/yyyy y se convierte en un objeto Date, lo cual permite identificar automáticamente a qué día de la semana pertenece cada fecha.

A continuación, se genera un recuento de tickets por día de la semana, lo que nos da la frecuencia de tickets para cada día. Este recuento se convierte en un data.frame y se ordenan los días de la semana manualmente para asegurar que el gráfico muestre los días en el orden lógico (de lunes a domingo).

Finalmente, se construye un gráfico de barras que muestra claramente cuántos tickets se han generado en cada día. Esto permite visualizar qué días son más activos y cuáles menos, ayudando a responder la pregunta de negocio: ¿Qué día sería el más adecuado para cerrar?

El día con menos tickets generados sería, en principio, el más adecuado para cerrar, ya que implicaría un menor impacto económico.

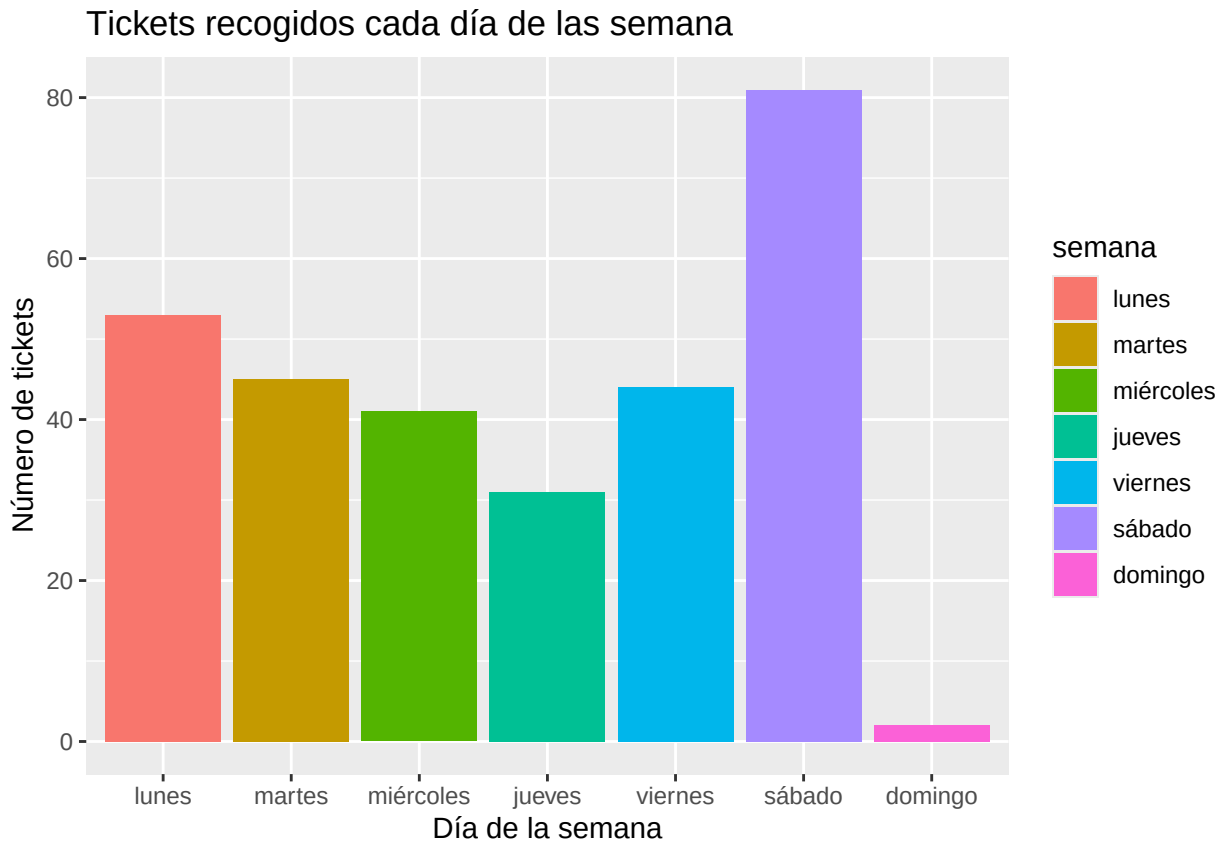


Figure 2: Tickets recogidos de cada dia de la semana

Como se observa en la figura, si se tuviese que cerrar un día entre semana seria el jueves.

Preguntas planteadas por nosotros

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

Se calcula la media de productos vendidos por ticket según el día de la semana. Para ello, se extrae la fecha desde el nombre del archivo y se identifican las líneas que representan productos reales. Luego, se cuenta cuántos productos hay por ticket y se asocia cada uno a su día de la semana. Finalmente, se calcula la media de productos por día, ordenando los días de lunes a domingo.

```
## # A tibble: 7 x 2
##   dia_semana media_productos
##   <chr>                <dbl>
## 1 lunes                 17.4
## 2 martes                14.5
## 3 miércoles             15.8
## 4 jueves                11.3
## 5 viernes               16.0
## 6 sábado                19.3
## 7 domingo               14
```

Finalmente, en la tabla podemos ver el resultado: la media de productos los lunes es de 17, la de los martes es de 14, la de los miércoles es de 16, la de los jueves es de 11, la de los viernes es de 16, la de los sábados es de 19 y la de los domingos es de 30. Podemos observar que, de normal, el dia que menos productos se venden es el jueves.

8. ¿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.

Este análisis se centra en visualizar cómo ha cambiado el precio por kilo de las manzanas en los tickets de compra. Para ello, se filtran los productos con unidad €/kg, se extrae el precio por kilo y se asocia al nombre del producto de la línea anterior. Luego se limpia y se filtra solo el producto manzana, generando un gráfico de su precio en el tiempo.

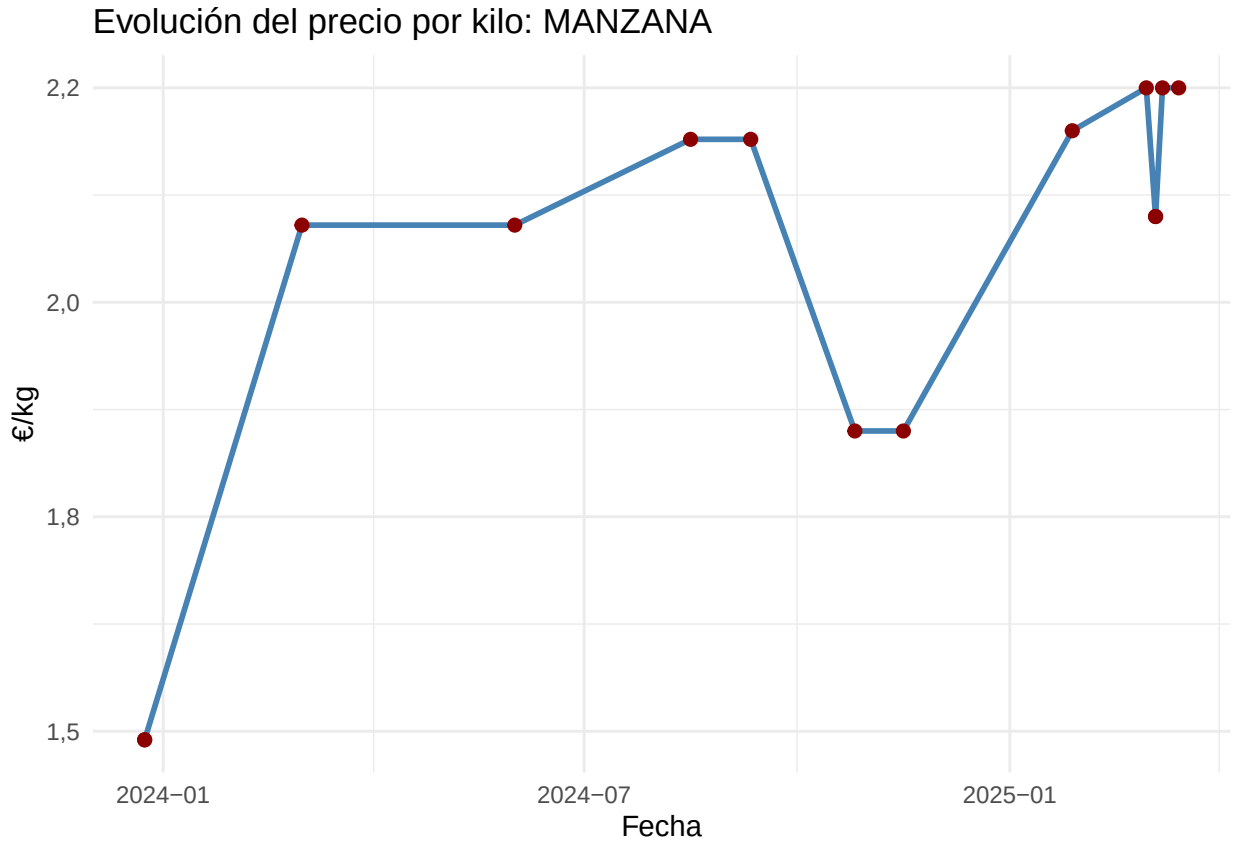


Figure 3: Evolución del precio por kg de manzana

Como podemos ver en el gráfico, el precio por kilo tiene picos de subida en la primera mitad del año 2024 y luego baja mucho y se mantiene constante.

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

Se analizan los tickets para identificar los productos no vendidos al peso, filtrando aquellos que no contienen "€/kg" y comienzan por una cantidad. Luego, se limpian y estandarizan los nombres de los productos para evitar duplicados. A continuación, se cuentan las apariciones de cada producto. Finalmente, se muestran los 10 más vendidos y los 10 menos vendidos.

```
## # A tibble: 10 x 2
##    producto                n
##    <chr>               <int>
##  1 A ALIÑADA PICADEDOS     1
##  2 ACEITE GIRASOL          1
##  3 ACEITUNA CHUPADEDOS     1
##  4 AGUA MINERAL            1
##  5 AJO Y PEREJIL           1
##  6 ALBONDIGAS DE BACALA    1
##  7 ALBONDIGAS UNID         1
##  8 ALCACHOFA PEQUEÑA       1
##  9 ALMENDRA LAMINADA       1
## 10 ALTRAMUCES              1
```

```
## # A tibble: 10 x 2
##    producto               n
##    <chr>              <int>
##  1 QUESO LONCHAS CABRA   20
##  2 BROTES TIERNOS MAXI   10
##  3 COPOS DE AVENA        10
##  4 PAN DE PUEBLO         10
##  5 PLATANO               10
##  6 QUESO FRESCO CABRA    10
##  7 BANANA                 9
##  8 ZANAHORIA BOLSA        9
##  9 LECHE DESNAT CALCIO    8
## 10 PARKING                8
```

Los productos más vendidos son queso lonchas cabra y brotes tiernos maxi y los menos vendidos son aceite de girasol y aliñada picadedos.

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

Buscamos identificar cuál es el producto que más se compra por unidades en los tickets disponibles. Para ello, primero se filtran aquellas líneas del conjunto de datos (df_lineas) que comienzan con un número seguido de un espacio.

Una vez filtradas esas líneas, el código divide cada una de ellas en varias columnas separadas: unidades, descripción del producto, precio por unidad e importe total. Después se realiza una limpieza adicional para eliminar productos que no corresponden a una venta real, que a veces aparecen registradas pero no deben considerarse en el análisis.

Luego, se seleccionan únicamente las columnas de unidades y descripción del producto, ya que el análisis se centra en saber qué producto se ha comprado en la mayor cantidad de unidades. Estos datos se ordenan de mayor a menor en función del número de unidades, con el fin de destacar los productos que se compran en grandes cantidades. Finalmente, se muestra el primer resultado de esta lista ordenada, que corresponde al producto más comprado por unidades en una sola transacción.

Table 6: Producto con el mayor número de unidades en un ticket

unidades	descripcion
8	COLA SIN CAFEINA

Como podemos observar en la tabla, el producto más comprado por unidades en una sola transacción es la cola sin cafeina (8 unidades).

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

El objetivo de esta pregunta es identificar qué pares de productos suelen comprarse juntos en un mismo ticket. Para comenzar, se filtran las líneas del dataset que contienen información sobre productos vendidos por unidad. Estas líneas se dividen en varias columnas como unidades, descripción del producto, precio por unidad e importe total. Se excluyen las líneas que hacen referencia a "DONACIÓN" para evitar sesgos en los datos. Luego, se seleccionan solo las columnas relevantes, en este caso el archivo (que identifica el ticket) y la descripción del producto. Posteriormente, se agrupan los datos por ticket y se eliminan aquellos que solo contienen un producto, ya que no pueden formar combinaciones.

Una vez obtenidos los tickets con al menos dos productos, se generan todas las combinaciones posibles de dos productos dentro de cada ticket. Cada par se ordena alfabéticamente para evitar que el orden afecte el conteo. Luego, los pares se combinan en una sola cadena de texto para facilitar su conteo. Finalmente, calculamos cuántas veces aparece cada combinación en todos los tickets, ordenando los resultados de mayor a menor frecuencia.

Table 7: Pares de productos diferentes que se compran más veces juntos

productos	n
BROTES TIERNOS MAXI QUESO LONCHAS CABRA	16
LECHE DESNAT. CALCIO QUESO LONCHAS CABRA	14
PAN DE PUEBLO QUESO LONCHAS CABRA	14
PLATANO QUESO LONCHAS CABRA	14
C. POLLO 100% NAT QUESO LONCHAS CABRA	12
QUESO FRESCO CABRA QUESO LONCHAS CABRA	12

Como podemos observar en la tabla, las combinaciones más comunes son brotes tiernos maxi y queso lonchas cabra (comprados 16 veces).

12. ¿Hay estacionalidad en plátanos? (se compran más en cierta fecha del año)

Esta pregunta busca detectar si existe una estacionalidad en la compra de plátanos, es decir, si hay meses específicos en los que se compran más plátanos que en otros. Para analizar esto, primero se filtran las líneas que contienen la palabra "plátano" (ignorando mayúsculas y minúsculas). Luego, se identifican los tickets (archivos) que contienen estos productos. A partir de esos tickets, se recuperan las fechas asociadas a cada compra de plátanos, que normalmente se encuentran en la línea 5 de cada ticket, y se extrae la fecha en formato texto para convertirla en un formato de fecha válido en R.

Una vez que se tiene la fecha, se necesita conocer la cantidad de plátanos comprados. Esto se logra localizando la línea inmediatamente después de donde aparece la palabra "plátano", ya que allí suele aparecer la cantidad en kilogramos, el precio por kilo y el precio total. Esta información se extrae con expresiones regulares y se separa en columnas llamadas cantidad, precio(kg) y precio(total). A continuación, se limpian los datos eliminando la unidad "kg", reemplazando las comas por puntos, y convirtiendo la cantidad a formato numérico. Después, se transforma la fecha para separar el año y el mes, y se agrupa por "año-mes" para sumar cuántos kilos de plátanos se vendieron en cada periodo. Finalmente, se genera un gráfico de puntos para visualizar si existe algún patrón o tendencia estacional en la compra de plátanos a lo largo del tiempo.

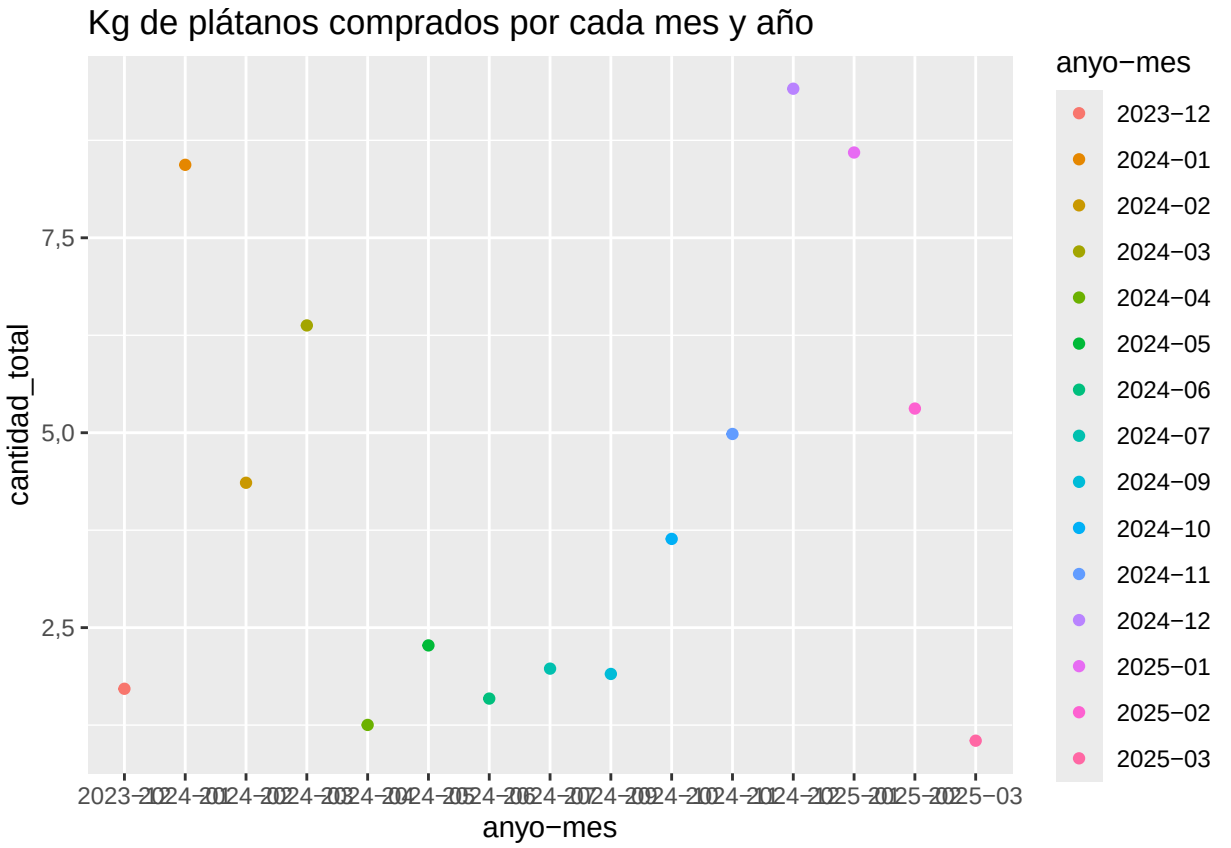


Figure 4: Cantidad de plátanos comprados por mes

Como se puede comprobar en la gráfica se compran más platanos en los meses de enero, febrero y marzo (es decir, en invierno), por tanto, los plátanos tienen estacionalidad.

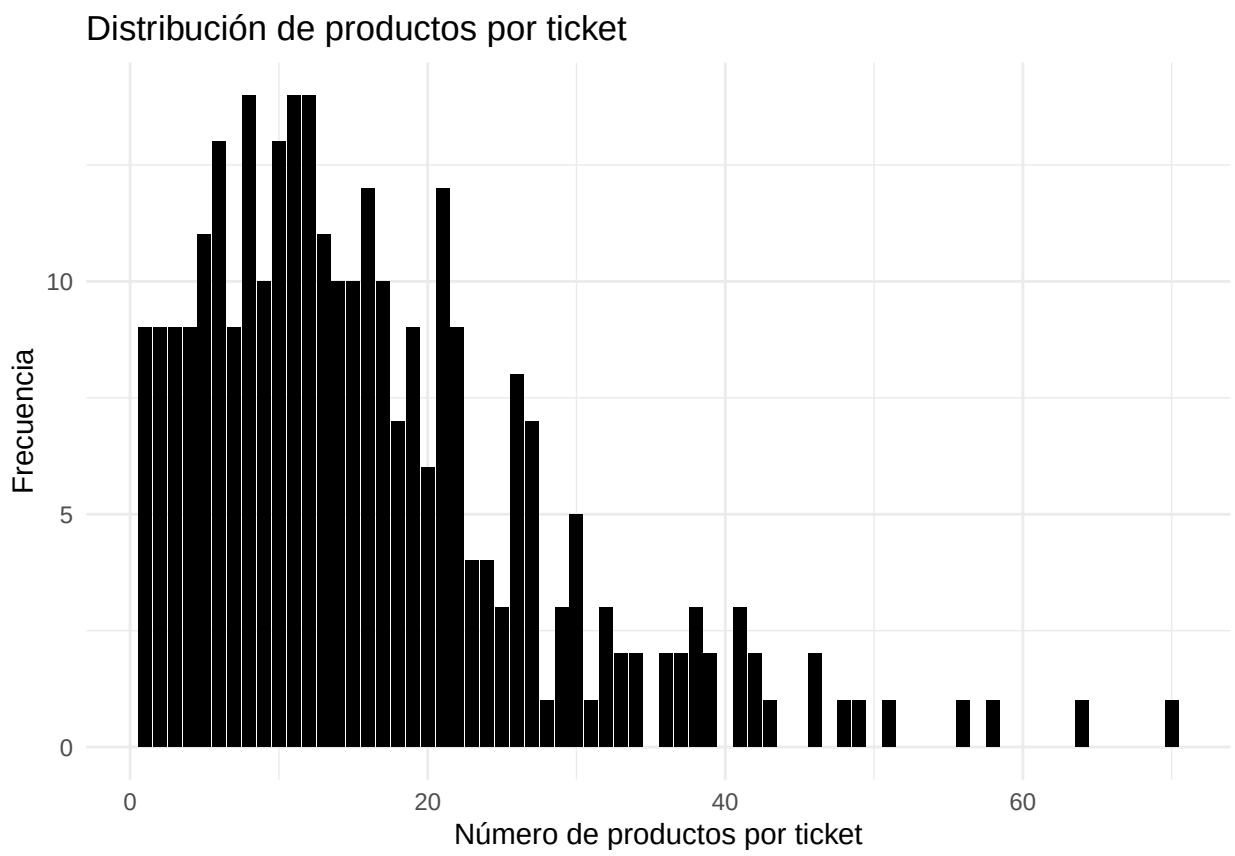
13. ¿Cuántos productos se compran por ticket, en promedio?

Esta pregunta calcula cuántos productos se compran, en promedio, por ticket. Para ello, se agrupan las líneas por ticket y se identifica hasta qué punto del texto aparecen los productos, usando como referencia las palabras “TOTAL” o “PARKING”.

Luego, se cuentan las líneas que representan productos, detectando aquellas que empiezan con un número seguido de un espacio (por ejemplo: “1 LIMPIADOR 1,50”), que es el formato típico en los tickets de Mercadona.

Una vez contado el número de productos en cada ticket, se calcula el promedio total. Además, se crea un gráfico de barras que muestra la frecuencia de tickets según la cantidad de productos que contienen.

```
## [1] 16
```



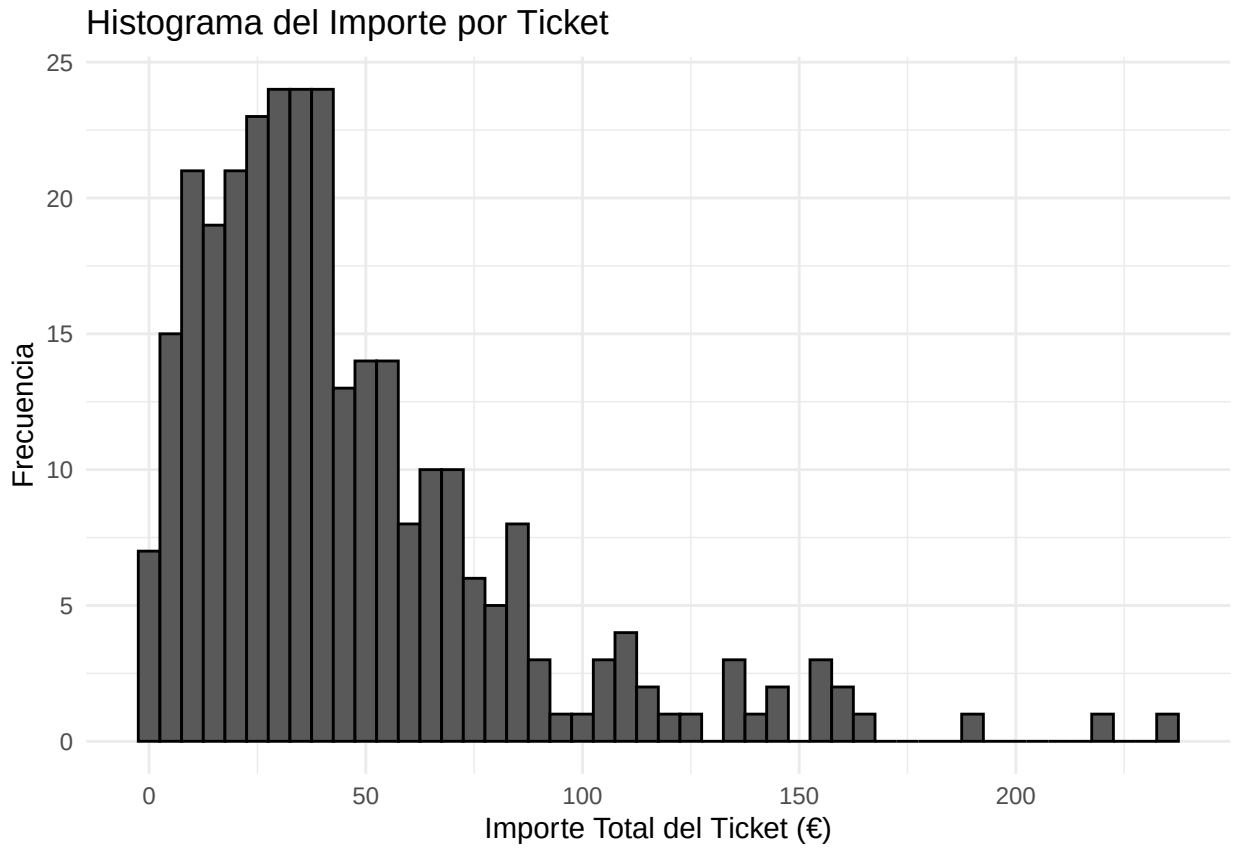
Como vemos en el resultado, en promedio se compran 16 productos por ticket.

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

Esta pregunta analiza los importes de los tickets para saber cuánto se gasta en promedio y detectar el ticket más caro y el más barato.

Primero se agrupan los datos por ticket y se identifica el método de pago (tarjeta, efectivo o desconocido). Luego se localiza la línea que contiene "TOTAL (€)" y se extrae el importe total del ticket.

Con esos datos, se calcula el importe medio por ticket y se muestra un histograma para ver la distribución de importes. Finalmente, se busca qué ticket tiene el importe más alto y cuál tiene el más bajo.




```
## [1] 46
```

```
## # A tibble: 1 x 3
##   archivo                       Metodo_Pago Importe_Total
##   <chr>                         <chr>              <dbl>
## 1 20240323 Mercadona 234,20.pdf Tarjeta             234.
```

```
## # A tibble: 1 x 3
##   archivo                       Metodo_Pago Importe_Total
##   <chr>                         <chr>              <dbl>
## 1 20241227 Mercadona 0,43 €.pdf Tarjeta             0.43
```

Como vemos en el resultado, el importe medio por ticket es de 46€. El ticket más caro registrado es: 20240323 Mercadona 234,20.pdf y el más barato es: 20241227 Mercadona 0,43 €.pdf.

15. ¿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%?

Queremos calcular la cantidad total de dinero recaudado en concepto de impuestos (IVA) a partir de los tickets de compra proporcionados. En este caso, nos piden centrarnos solo en 3 tipos de IVA (21%, 10% y 5%).

Para poder realizar este análisis, primero debemos partir del dataframe original (df_lineas). Como los datos del IVA se pueden encontrar en diferentes lineas utilizaremos una expresión regular que detecte información sobre el IVA. Esto lo haremos mediante la función str_detect() de la librería stringr. Además usaremos una serie de funciones para transformar y organizar los datos.

Ya hemos creado el dataframe que contiene toda la información del IVA proporcionada en los tickets. Ahora a partir de este agruparemos todas las líneas por porcentaje de IVA y sumaremos todas las cuotas asociadas a los tipos de IVA. Una vez lo tengamos utilizaremos la función filter() para así centrarnos únicamente en los tres tipos de IVA que queremos analizar, que son IVA del 5%, 10% y 21%.

Table 8: Total de IVA por tipo de porcentaje

porcentaje	Cuota
21	256,7
10	666,0
5	8,2

La tabla @ref:tabla-TiposIVA revela que la mayor recaudación por IVA, con un total de 666.01€, corresponde a los productos gravados al 10%. En cambio, los productos con un IVA del 21% generaron una recaudación de 256.74€, mientras que aquellos con un IVA del 5% aportaron 8.21€.

Por último, queremos verificar si se han tenido en cuenta todos los tickets y que no se haya perdido ningún dato en el proceso de extracción del IVA. Podemos comparar el número total de tickets únicos en el dataframe original (df_lineas) con el número de tickets para los cuales se pudo extraer distinta información del IVA en el dataframe lineas_IVA.

```
## Tickets totales: 297
```

```
## Tickets con IVA: 297
```

Como podemos ver, ambos números son iguales (total de tickets únicos y tickets con el IVA extraido), entonces no hay ningún dato perdido.

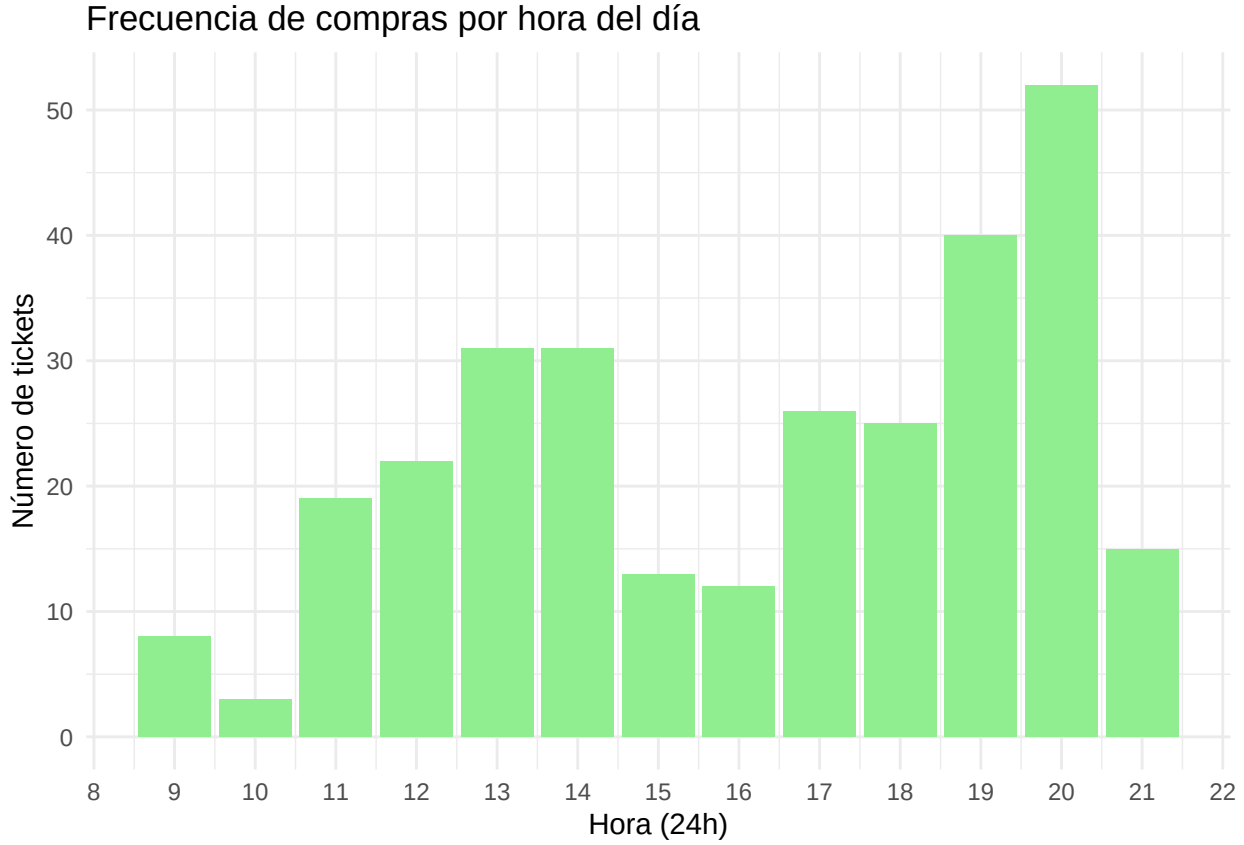
16. ¿A qué horas se suele ir más a comprar en los supermercados? ¿Cuáles son las que menos?

Queremos saber qué horas del día tienen mayor y menor afluencia en supermercados. Para ello, necesitamos extraer la hora de compra registrada en cada ticket.

Dado que la hora puede encontrarse en diferentes líneas dentro del ticket, aplicamos una expresión regular que detecta el formato "hh:mm". Esto se realiza mediante la función str_extract() de la librería stringr. Además usamos una serie de funciones para transformar y organizar los datos.

Este proceso nos permite construir un nuevo data frame (df_horas) con la hora de compra de cada ticket.

A partir del nuevo data frame (df_horas) contamos con que frecuencia de tickets por cada hora del día y lo visualizamos en un gráfico de barras. Así, podemos identificar con claridad las horas pico de afluencia y las horas más tranquilas.



El gráfico de barras @ref:figura-Frecuenciahoras muestra que las horas de mayor afluencia en el supermercado son las 20h y las 19h, con poco más de 50 y 40 tickets registrados respectivamente. En cambio, las horas de menor afluencia, sin considerar el horario de cierre (8h y 22h), son las 10h y las 9h, con menos de 5 tickets a las 10h y menos de 10 a las 9h.

Por último, queremos verificar si se han tenido en cuenta todos los tickets y que no se haya perdido ningún dato en el proceso de extracción de la hora. Podemos comparar el número total de tickets únicos en el dataframe original (df_lineas) con el número de tickets para los cuales se pudo extraer una hora en el dataframe df_horas.

```
## [1] 297
```

```
## [1] 297
```

Como podemos ver, ambos números son iguales (total de tickets únicos y tickets con hora extraída), entonces no hay ningún dato perdido.

17. ¿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?

Nos piden averiguar cual es el método de pago que más se repite: efectivo o tajeta.

En este caso se repite el hecho de que la forma de pago puede encontrarse en diferentes líneas dentro del ticket, por eso vamos a aplicar la funcion str_detect() para ver si es Tarjeta o Efectivo. Además queremos calcular también el importe total de cada método de pago.

Tenemos ya el dataframe resumen_tickets que contiene solo las líneas de texto que mencionan el método de pago. A partir de este dataframe calcularemos el importe total de cada método de pago.

Table 9: Importe total de cada método de pago

Metodo_Pago	Total_Gastado
Tarjeta	13650

Como se puede observar en la tabla cada ticket ha coincidido en el método de pago, todos han sido pagados con tarjeta. También vemos que el importe total de todos los tickets es 13.649,8€.

Conclusión

A lo largo de este proyecto, se ha desarrollado una herramienta funcional y automatizada en R para la lectura, limpieza, estructuración y análisis de tickets de compra en formato PDF, concretamente de supermercados Mercadona. Este sistema permite transformar documentos no estructurados en un conjunto de datos ordenado y analizable, lo que constituye un paso esencial en el procesamiento de información real proveniente del ámbito del consumo diario.

Gracias al uso de librerías como pdftools, dplyr y stringr, ha sido posible extraer con precisión la información clave de los tickets: desde los productos comprados hasta sus cantidades, precios, fechas y patrones de compra. Además, se han podido explorar aspectos interesantes como las combinaciones de productos más frecuentes o el posible comportamiento estacional en la compra de ciertos alimentos, como los plátanos.

Este trabajo no solo demuestra el valor del análisis de datos aplicado a la vida cotidiana, sino que también sienta las bases para posibles desarrollos futuros. Por ejemplo, se podrían integrar tickets de otros supermercados, analizar tendencias de precios a lo largo del tiempo o incluso construir recomendaciones personalizadas para consumidores.

En resumen, este proyecto muestra cómo la tecnología y el análisis de datos pueden ayudar a entender mejor los hábitos de consumo, ofrecer información útil para la toma de decisiones y fomentar una gestión más inteligente del gasto doméstico.