ProyectoTD2025 - Grupo C

Tratamiento de Datos. Grado en Ciencia de Datos- UV

Alexandra Alfonso, Alberto Cabedo, Javier Marzo, Iker Pérez, David Puertes y Julio Sierra

2025-04-08

Índice

1. Introducción

Este proyecto se centra en realizar un análisis detallado de los tickets electrónicos proporcionados por Mercadona, una reconocida cadena de supermercados en España. Estos tickets electrónicos son enviados a los usuarios a través del correo electrónico en formato PDF.

Los datos utilizados en este análisis provienen de los tickets electrónicos implementados por Mercadona. Estos documentos contienen información detallada sobre las transacciones de compra realizadas en sus establecimientos.

El objetivo principal es examinar a fondo estos tickets electrónicos para obtener información valiosa sobre los hábitos de compra en Mercadona. Específicamente, se enfocarán en identificar las compras más comunes, los supermercados más visitados, los productos más populares y que localidad tiene mayor número de tickets entre otras cosas.

```
## Warning: package 'pacman' was built under R version 4.4.2
```

1.1 Carga de librerías necesarias para el análisis

Inicialmente, se cargan los paquetes necesarios para cada una de las fases del proyecto. Para hacerlo de forma más eficiente y ordenada, se utiliza el paquete pacman del lenguaje de programación R

1.2 Extracción de datos del fichero de texto

1.2.1 Recopilación de ficheros .txt

Con el propósito de simplificar el procesamiento de los datos contenidos en los tickets de Mercadona, se ha optado por utilizar un código en R que convierte los archivos PDF a archivos de txt.

Todos estos archivos se encuentran almacenados en una carpeta compartida designada para este tipo de archivos [./data], a la cual se accede para obtener todos los tickets en formato .txt.

Este enfoque permite realizar un análisis línea por línea de cada documento de manera eficiente y estructurada.

1.2.2 Recopilación de los datos

Una vez que todos los tickets han sido recolectados y transformados en texto plano, es momento de almacenar la información contenida en ellos. Para ello, hemos realizado 2 data.frames de trabajo:

- ***df.ticket***: Este data frame recopila la información general de cada ticket.
- ***df.productos***: Este data frame recopila la información de cada producto comprado y su precio

1.2.2.1 Recogida de datos generales del ticket Creamos un data.frame con los datos del ticket

1.2.2.2 Descripción de las variables del ticket Las variables creadas y almacenadas para esta tabla con datos más generales es la siguiente:

- ***numero.factura***: Es una variable de tipo texto que identifica unívocamente cada ticket analizado durante el transcurso del análisis.
- ***precio.total***: Es una variable de tipo numérico que informa sobre el valor del precio total del ticket
- ***precio.total.sin.IVA***: Es una variable de tipo numérico que informa sobre el valor del precio total del ticket, quitando todo el IVA añadido a cada producto.
- ***iva.anyadido***: Es una variable de tipo numérico que informa sobre el valor total de IVA añadido a todos los productos.
- ***direccion***: Es una variable de tipo categórico que informa sobre la dirección del Mercadona donde se realizó la compra.
- ***ciudad***: Es una variable de tipo categórico que informa sobre la ciudad donde se realizó la compra.
- ***codigo.postal***: Es una variable de tipo categórico que, análogamente a la ciudad, informa sobre el código postal que tiene la ciudad donde se realizó la compra.
- ***telefono***: Es un valor de tipo texto que indica el número de teléfono del Mercadona donde se realizó la compra.
- ***fecha***: Es una variable de tipo fecha que indica el día en el que se realizó la compra.
- ***hora***: Es una variable de tipo periodo que indica la hora exacta en la que se realizó el pago de la compra.
- ***tipo.tarjeta***: Es una variable de tipo categórico que representa las marcas de tarjetas de crédito o débito utilizadas junto con la indicación de VERIFICADO POR DISPOSITIVO, que representa un método adicional de autenticación.
- ***metodo.pago***: Es una variable de tipo categórico que representa los diferentes métodos de pago utilizados en las transacciones registradas.
- ***autorizacion.pago***: Es una variable de tipo texto que representa la autorización del banco con respecto al pago solicitado.
- ***identificacion.aplicacion.pago***: Es una variable de tipo texto que representa la identifficación de la aplicación bancaria utilizada a la hora de realizar el pago.

- ***autorizacion.transaccion***: Es una variable de tipo texto que representa el tipo de transacción realizada.
- ***numero.centro***: Es una variable de tipo texto que representa el número asociado al centro donde se ha realizado la compra.
- ***numero.caja***: Es una variable de tipo texto que representa la caja en la que el usuario ha sido atendido.

Los nombres de las variables son bastante descriptivos pero es necesaria una pequeña explicación sobre ellos para en caso de olvidarse o retomar el proyecto en un futuro sea fácil su entendimiento y rápida la puesta en marcha.

Cuadro 1: Tabla 1. Resumen general de las variables del conjunto de tickets.

Variable	tipo	niveles	topLevel	topCount	topFrac	missFrac
numero.factura	character	290	3075-010-680549	2	0.0066	0
precio.total	numeric	NA	NA	NA	NA	0
precio.total.sin.IVA	numeric	NA	NA	NA	NA	0
iva.anyadido	numeric	NA	NA	NA	NA	0
direccion	factor	34	C/ QUART 120	51	0.1689	0
ciudad	factor	19	VALENCIA	130	0.4305	0
codigo.postal	factor	29	46008	51	0.1689	0
telefono	character	34	963824500	51	0.1689	0
fecha	Date	NA	NA	NA	NA	0
hora	Period	NA	NA	NA	NA	0
tipo.tarjeta	factor	2	ARC: 00	154	0.5099	0
metodo.pago	factor	1	TARJETA BANCARIA	302	1.0000	0
autorizacion.pago	character	289	077266	2	0.0066	0
identificacion.aplicacion.pago	character	3	A0000000041010	215	0.7119	0
autorizacion.transaccion	factor	2	ARC: 00	154	0.5099	0
numero.centro	character	34	098101017	51	0.1689	0
numero.caja	character	205	461428	6	0.0199	0

```
## Warning: There was 1 warning in ‘summarise()‘.
## i In argument: ‘across(...)‘.
## Caused by warning:
## ! The ‘...‘ argument of ‘across()‘ is deprecated as of dplyr 1.1.0.
## Supply arguments directly to ‘.fns‘ through an anonymous function instead.
##
##   # Previously
##   across(a:b, mean, na.rm = TRUE)
##
##   # Now
##   across(a:b, \(x) mean(x, na.rm = TRUE))
```

Cuadro 2: Tabla 2. Estadísticos descriptivos de las variables numéricas.

Variable	Media	Mediana	DesvEst	Min	Max
precio.total	46.31	37.32	37.82	0.43	234.20
precio.total.sin.IVA	42.90	34.09	35.14	0.39	217.04

iva.anyadido	3.39	2.64	2.88	0.00	17.16
hora	0.00	0.00	12315.15	0.00	0.00

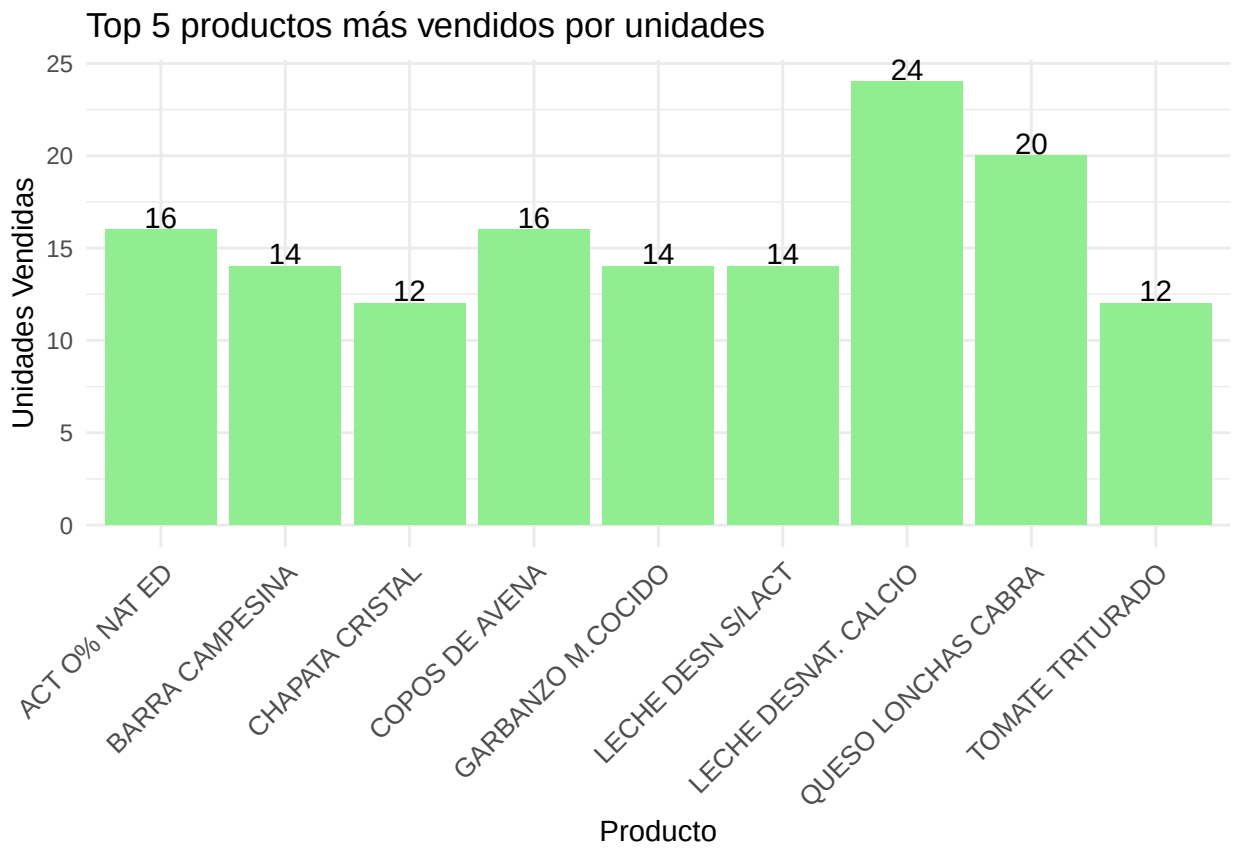
Cuadro 3: Tabla 3. Frecuencia de los valores más comunes en variables categóricas clave.

Variable	Valor	Frecuencia
ciudad	VALENCIA	130
ciudad	ALBORAIA/ALBORAYA	50
ciudad	BURJASSOT	26
ciudad	MURO	24
ciudad	ALCOI/ALCOY	22
metodo.pago	TARJETA BANCARIA	302
numero.centro	098101017	51
numero.centro	036426237	50
numero.centro	077763746	38
numero.centro	098101330	26
numero.centro	003586427	24

1.3 Preguntas propuestas

- Estas son las preguntas que tenemos que responder mediante el uso de gráficos:
- ¿Cuáles son los 5 productos, de los vendidos por unidades, con más ventas ? ¿Cuántas unidades de cada uno se han vendido ?Ç
- 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 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.
- ¿ 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 ?
- ¿Existe algún patrón en cuanto a la hora del día en que se realizan las compras?
- ¿Cuánto dinero de media se gastan los compradores?
- ¿Cuanto dinero se suele añadir con el IVA?

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



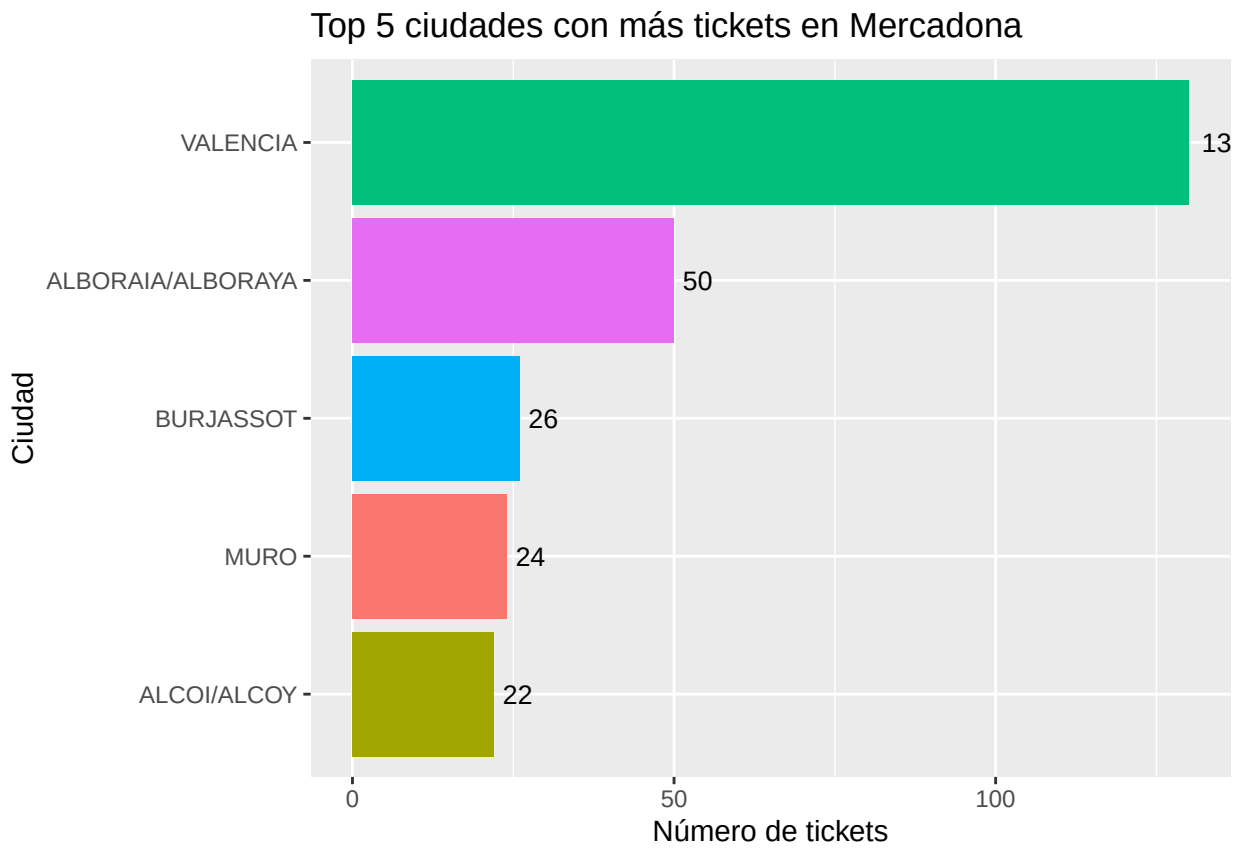
En esta gráfica podemos observar los productos más vendidos por unidades. El producto con mayor número de ventas es la leche desnatada sin lactosa, con 24 unidades, seguido de las lonchas de queso de cabra con 20. También destacan productos como los yogures Activia, los copos de avena o la barra de pan campesina. Aunque el título indica un Top 5, se han incluido más productos para ofrecer una visión más completa del comportamiento de ventas. Este análisis permite identificar los productos más demandados.

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?

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 ?

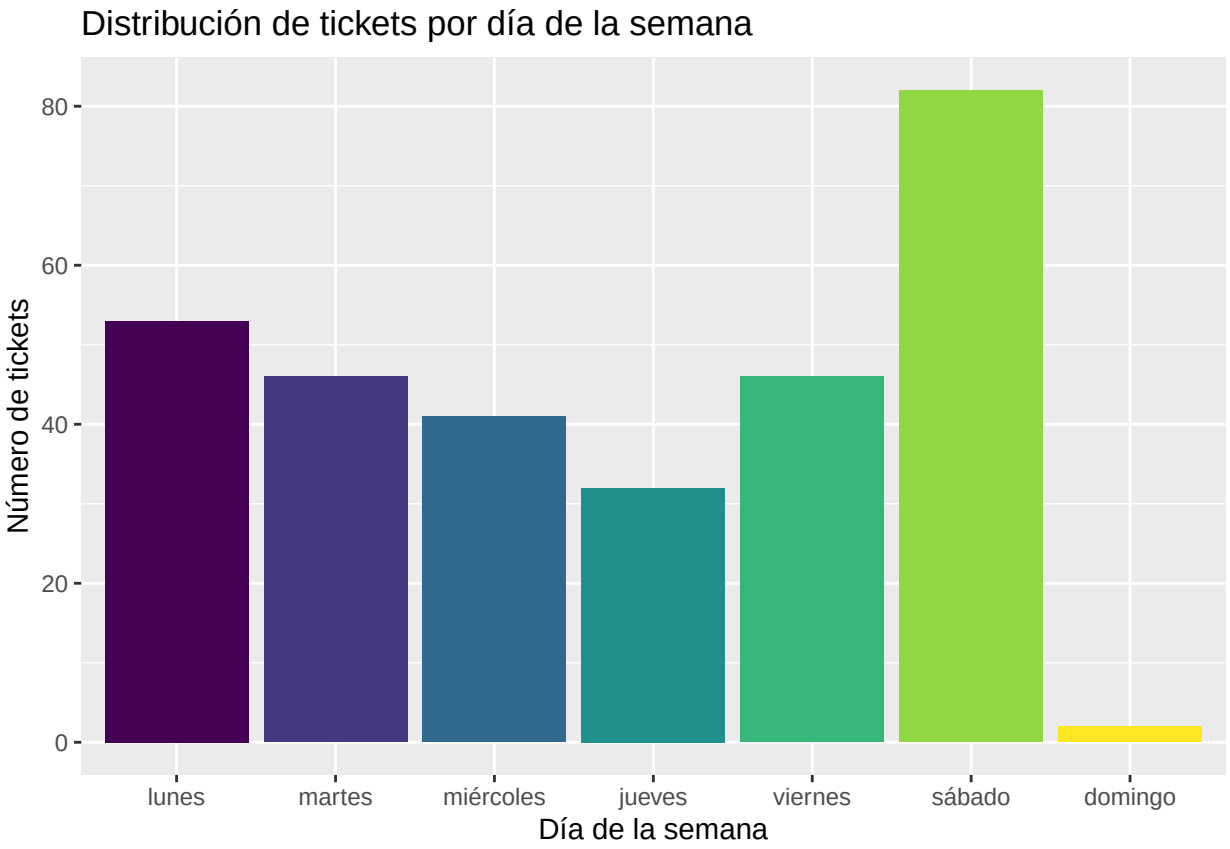
4) Muestra mediante un gráfico de líneas como ha variado el precio por kilo de las bananas y los plátanos en los tickets disponibles, a lo largo del tiempo.

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



Valencia claramente lidera con 130 tickets. Alboraya aparece en segundo lugar con 50 tickets.

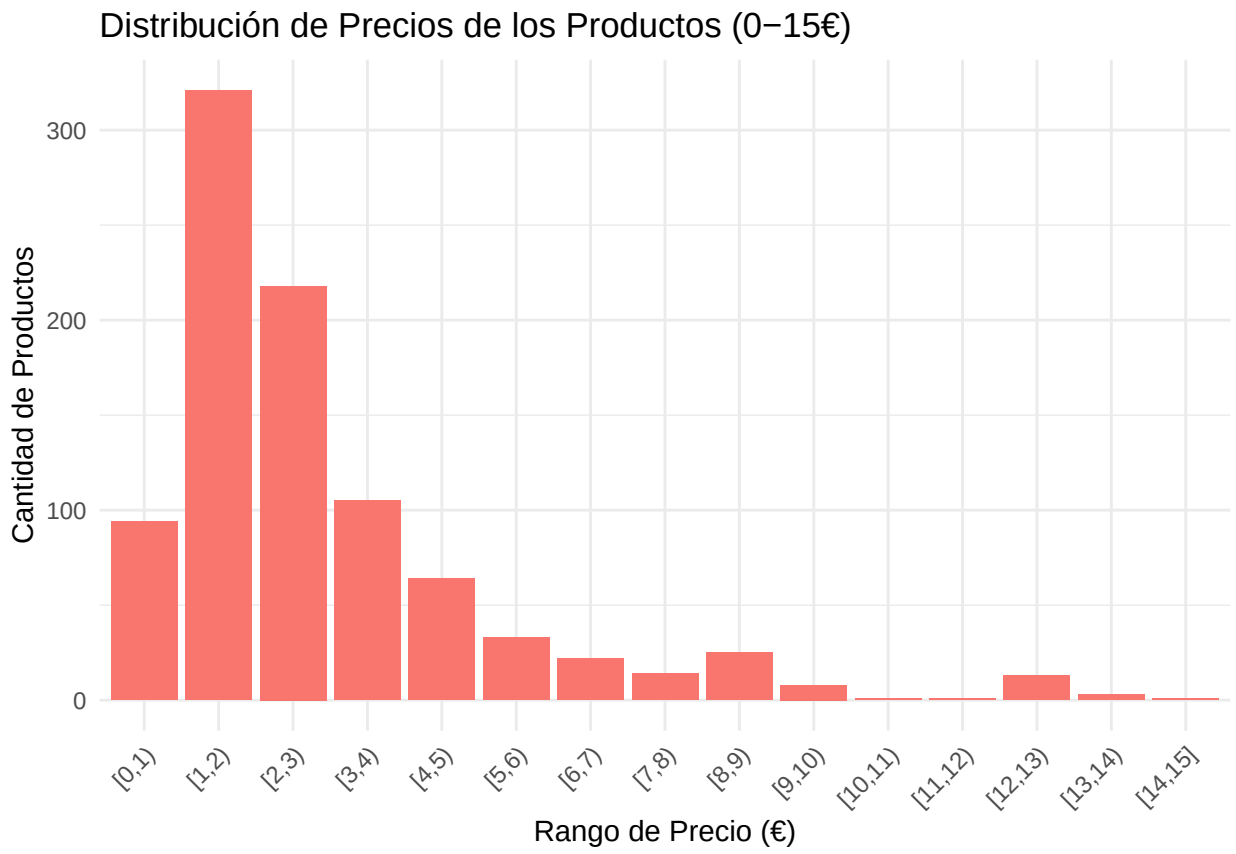
6) Muestra mediante un diagrama el número de tickets recogidos cada día de las semana. ¿Si tuvieses que cerrar un día entre semana qué día lo harías?



- Además de las preguntas propuestas por el profesorado, como equipo hemos formulado algunas cuestiones adicionales que podemos abordar con los datos disponibles. Estas son las siguientes:

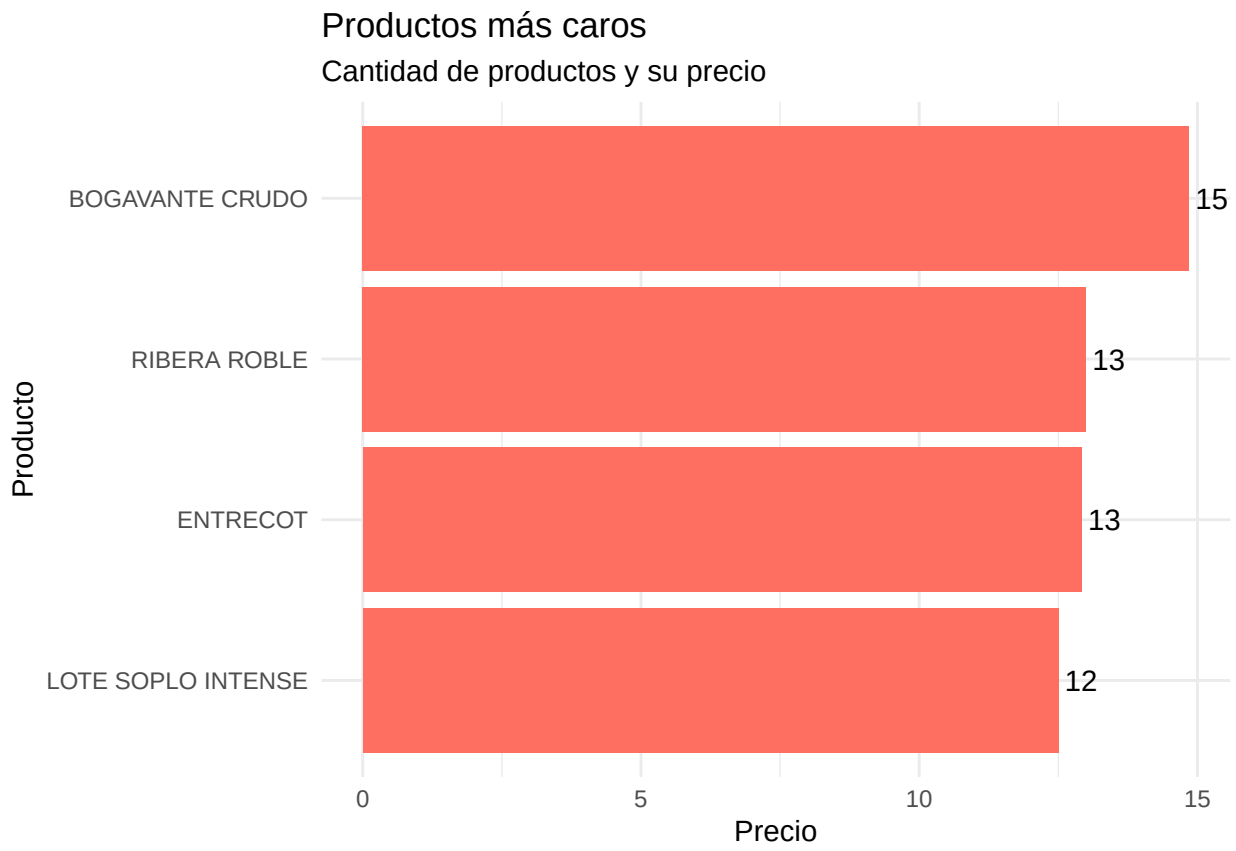
    7) ¿Existe algún patrón en cuanto a la hora del día en que se realizan las compras?
    8) ¿Cuánto dinero de media se gastan los compradores?
    9) ¿Cuanto dinero se suele añadir con el IVA?
    10) ¿Cuál es la tienda en la que más se ha comprado? (Dirección tienda)
    11) ¿Cuál es el alimento comprado más caro?

Como para hacer un gráfico de todos los productos que se han comprado sería un poco difícil su comprensión ya que hay demasiados productos diferentes, podemos optar por agrupar los productos por rangos de precios y mostrar cuántos hay en cada rango. Esto genera una visión general de la distribución de precios de los productos.



‘‘‘

Y ahora, para encontrar los productos que están en el rango de precios entre 12 y 15 euros y representarlos en un gráfico, primero filtraremos el dataframe original para incluir solo los productos dentro de ese rango y luego mostraremos esos productos en un gráfico.



Por tanto, como podemos observar el alimento comprado más caro es el bogavante crudo el cual solo se ha comprado 1 vez, el cual su precio es de casi 15€.