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

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.

2. 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

3. Extracción de datos del fichero de texto

3.1 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.

3.2 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

3.2.1 Recogida de datos generales 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.

3.2.2 Recogida de datos de los productos del ticket Las variables creadas y almacenadas para la tabla de los productos es la siguiente:

- ***numero.factura***: Es una variable de tipo texto que identifica unívocamente cada ticket analizado durante el transcurso del análisis.
- ***cantidad***: Es una variable de tipo numérico que informa sobre la cantidad de productos comprados. En el caso en el que el producto esté condicionado por la variable peso, la cantidad será indicada como 1.
- ***nombre.producto***: Es una variable de tipo categórico que informa sobre el nombre del producto.
- ***precio.producto***: Es una variable de tipo numérico que indica el valor del producto por una unidad del mismo. Es decir, no tiene en cuenta la cantidad de productos de ese nombre comprados.

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.

Consideramos también la clasificació por categoría de los distintos productos almacenados en el ***df.productos***.

3.2.3 Tablas para facilitar la información que contienen los tickets Realizamos las tablas necesarias para realizar un resumen analítico de los datos proporcionados por los tickets.

TABLA 1: Resumen general de variables

Cuadro 1: Resumen General de Variables

Variable	Tipo	Niveles Ún.	Valor Más Frec.	Frec.	% Total	Faltantes (%)
Numero Factura	Texto	290	3075-010-680549	2	0.7%	0.0%
Precio Total	Numérica	NA	NA	NA	NA	0.0%
Precio Total Sin Iva	Numérica	NA	NA	NA	NA	0.0%
Iva Anyadido	Numérica	NA	NA	NA	NA	0.0%
Direccion	Categórica	34	C/ QUART 120	51	16.9%	0.0%
Ciudad	Categórica	19	VALENCIA	130	43.0%	0.0%
Codigo Postal	Categórica	29	46008	51	16.9%	0.0%
Telefono	Texto	34	963824500	51	16.9%	0.0%
Fecha	Fecha	NA	NA	NA	NA	0.0%
Hora	Hora	NA	NA	NA	NA	0.0%
Tipo Tarjeta	Categórica	2	ARC: 00	154	51.0%	0.0%
Metodo Pago	Categórica	1	TARJETA BANCARIA	302	100.0%	0.0%
Autorizacion Pago	Texto	289	077266	2	0.7%	0.0%
Identificacion Aplicacion Pago	Texto	3	A0000000041010	215	71.2%	0.0%
Autorizacion Transaccion	Categórica	2	ARC: 00	154	51.0%	0.0%
Numero Centro	Texto	34	098101017	51	16.9%	0.0%
Numero Caja	Texto	205	461428	6	2.0%	0.0%

Vemos como la tabla 1 presenta un resumen de las variables contenidas en df.ticket, indicando el tipo, número de niveles (si corresponde), valor más frecuente, su frecuencia relativa y la fracción de valores perdidos.

TABLA 2: Estadísticos de variables numéricas

Estadísticos Descriptivos de Variables Cuantitativas

Tabla 2

Variable	Media	Mediana	Desv. Est.	Mínimo	Máximo
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	12,315.15	0.00	0.00

Observamos como en la tabla 2 se muestran los principales estadísticos descriptivos de las variables cuantitativas como el precio total del ticket, el IVA añadido y el precio sin IVA.

TABLA 3: Frecuencias de variables categóricas clave

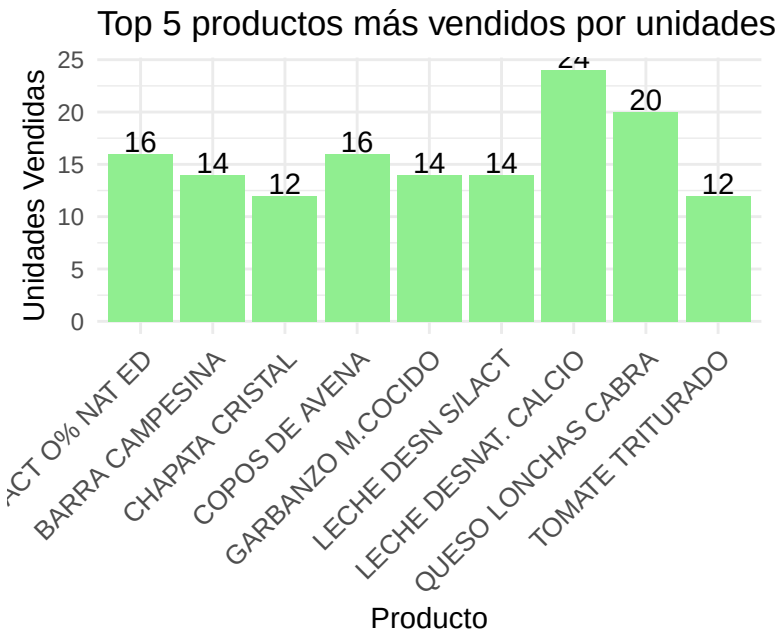
Frecuencia de Valores en Variables Categóricas			
Variable	**Valor**	**Frecuencia**	**Porcentaje**
Ciudad	VALENCIA	130	51.6%
	ALBORAIA/ALBORAYA	50	19.8%
	BURJASSOT	26	10.3%
	MURO	24	9.5%
	ALCOI/ALCOY	22	8.7%
Método de Pago	TARJETA BANCARIA	302	100.0%
Número de Centro	098101017	51	27.0%
	036426237	50	26.5%
	077763746	38	20.1%
	098101330	26	13.8%
	003586427	24	12.7%

Visualizamos como la tabla 3 muestra las frecuencias más altas de tres variables categóricas clave: ciudad, método de pago y número de centro, lo que permite comenzar a observar patrones interesantes en los tickets.

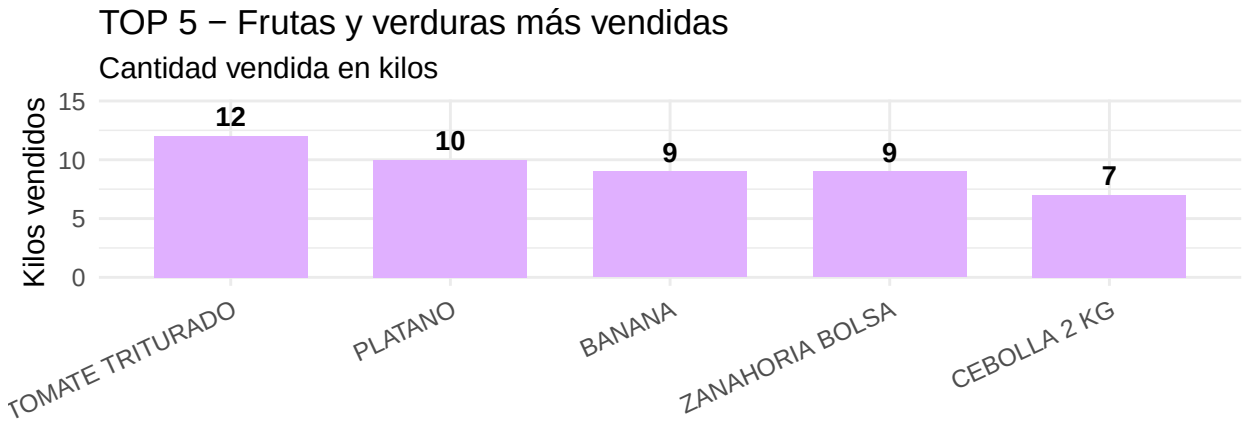
3. Preguntas propuestas

- Estas son las preguntas que tenemos que responder mediante el uso de gráficos:

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



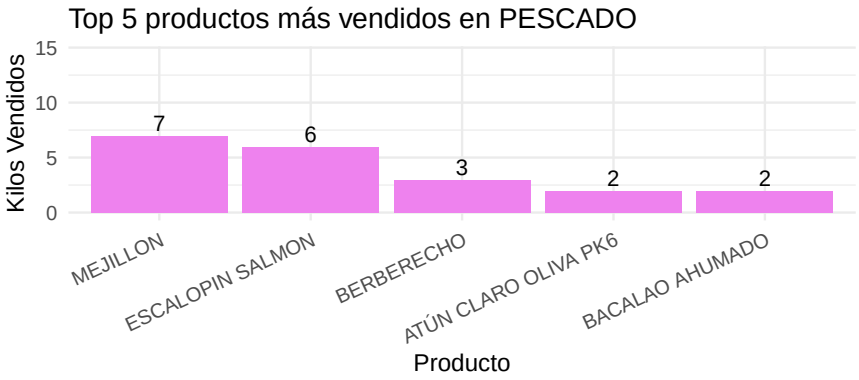
2) Si consideramos la categoría de FRUTAS Y VERDURAS. Cuáles son los 5 productos más vendidos?¿Cuántos kilos se han vendido de cada uno de estos productos?



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 ?

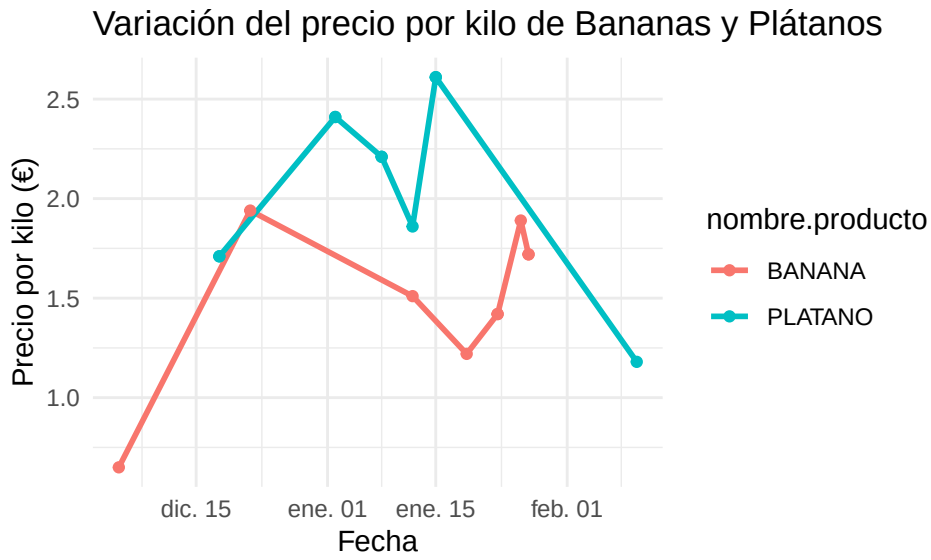
```
## # A tibble: 5 x 2
##   nombre.producto      Kilos_Vendidos
##   <chr>                         <dbl>
## 1 MEJILLON                          7
```

```
## 2 ESCALOPIN SALMON                6
## 3 BERBERECHO                      3
## 4 ATÚN CLARO OLIVA PK6            2
## 5 BACALAO AHUMADO                 2
```



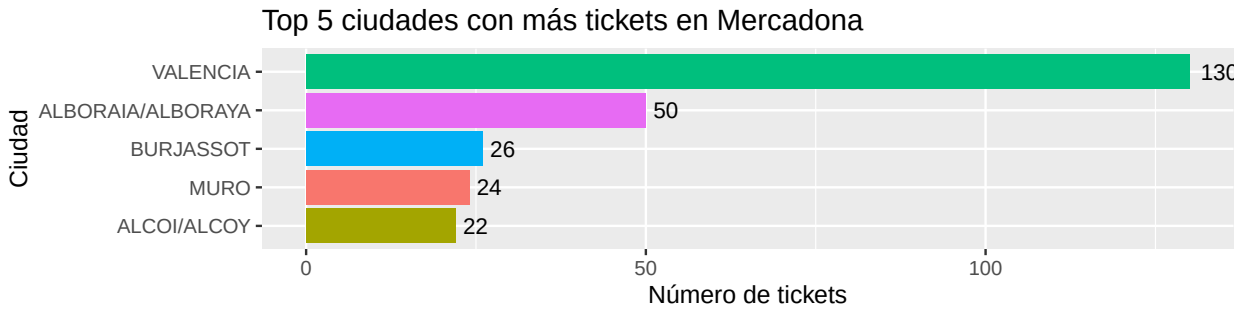
El diagrama indica que el mejillón es el producto más comercializado en la categoría de pescado, con 7 kilos, siendo el escalopín de salmón el más cercano con 6 kilos. Productos como el berberecho, el atún claro en oliva y el bacalao ahumado tienen ventas más bajas, con 3 y 2 kilos, respectivamente. Esto sugiere una evidente inclinación hacia productos marinos asequibles y sencillos de ingerir, como los mejillones y el salmón.

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.`



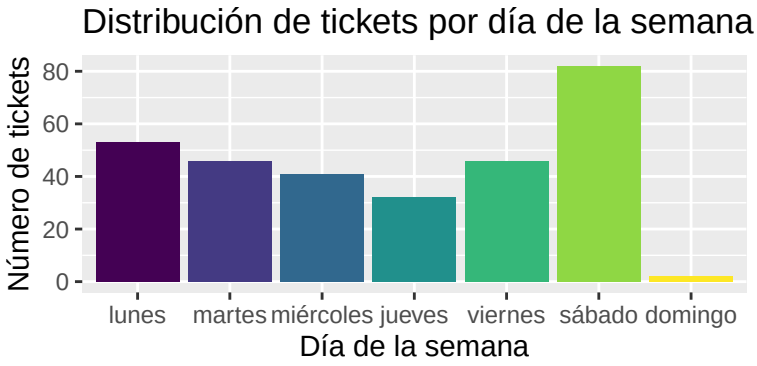
El diagrama ilustra el cambio en el costo por kilogramo de bananas y plátanos desde mediados de diciembre hasta comienzos de febrero. El costo de los plátanos (línea azul) sufre una fluctuación significativa, alcanzando un máximo a mediados de diciembre, seguido de un descenso. Por otro lado, el precio de las bananas (línea roja) inicia con un costo más bajo, incrementándose al final de diciembre y permaneciendo más constante en enero. Estas variaciones pueden atribuirse a elementos de oferta y demanda durante ese lapso.

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



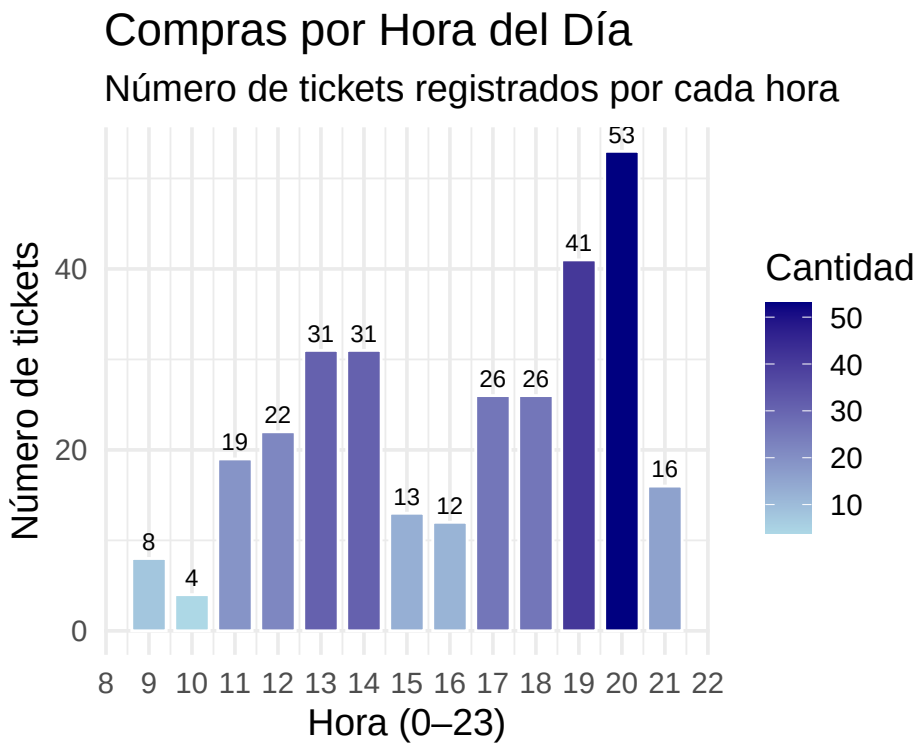
Valencia claramente lidera con 130 tickets. Alboraya aparece en segundo lugar con 50 tickets.

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?



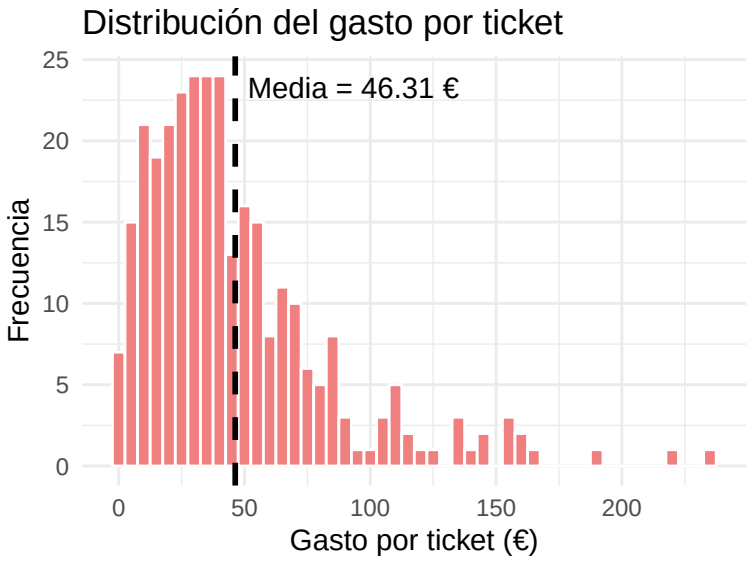
- Además de las preguntas propuestas por el profesorado, como equipo hemos formulado algunas cuestiones adicionales que podemos abordar con los datos disponibles. Estas son las siguientes:

  7) ¿Existe algún patrón en cuanto a la hora del día en que se realizan las compras?



Las horas a las que mas compras hay es justo antes de cenar, donde la gente vuelve del trabajo y baja a comprar la cena y donde menos hay es por la mañana, a la hora de comer y a la hora de cenar, lo que tiene sentido ya que están comprando lo que van a necesitar para preparar su comida o su cena

8) ¿Cuánto dinero de media se gastan los compradores?

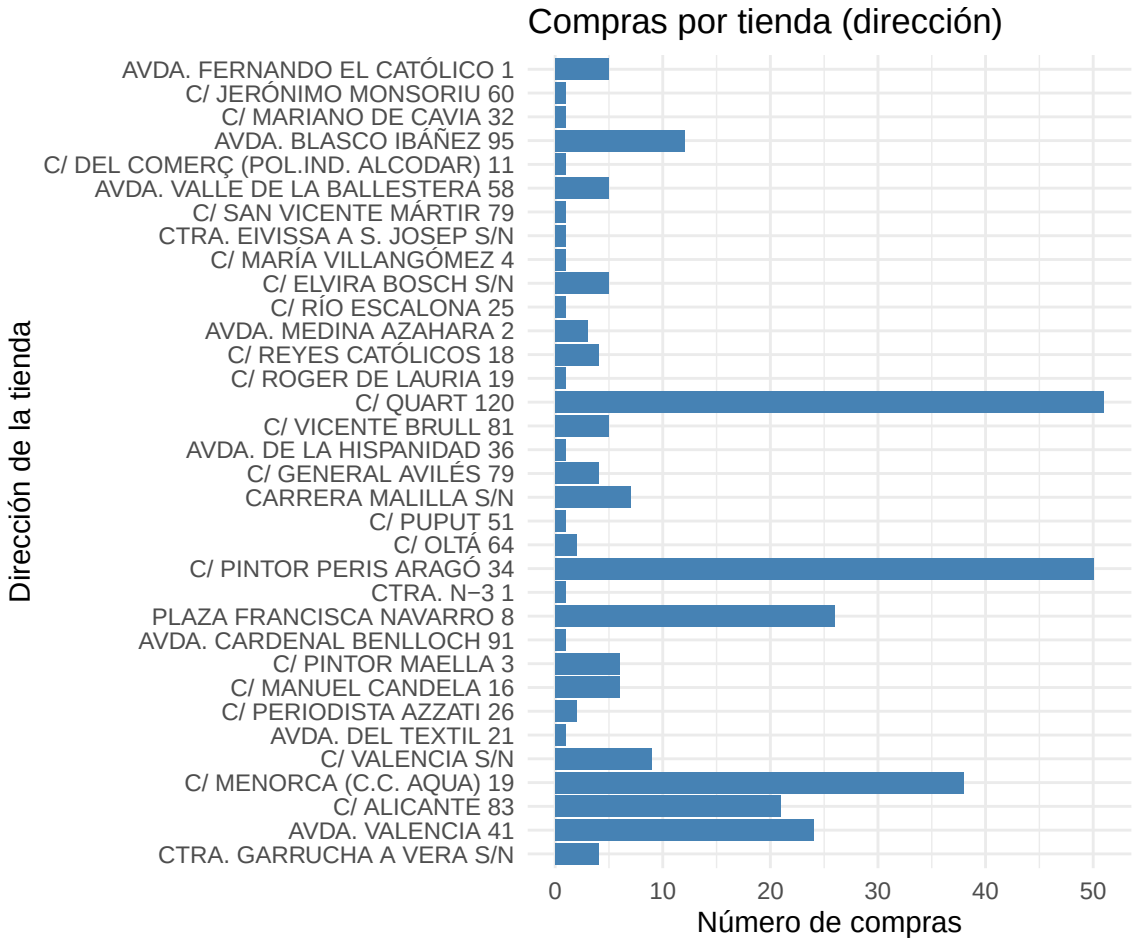


A partir de esta gráfica se puede ver que la media da 46,31€ pese a que la mayoria de las compras sean mas baratas debido a algunas compras muy grandes que han habido en nustro conjunto de datos

9) ¿Cuanto dinero se suele añadir con el IVA?

IVA Promedio: 6.02 €

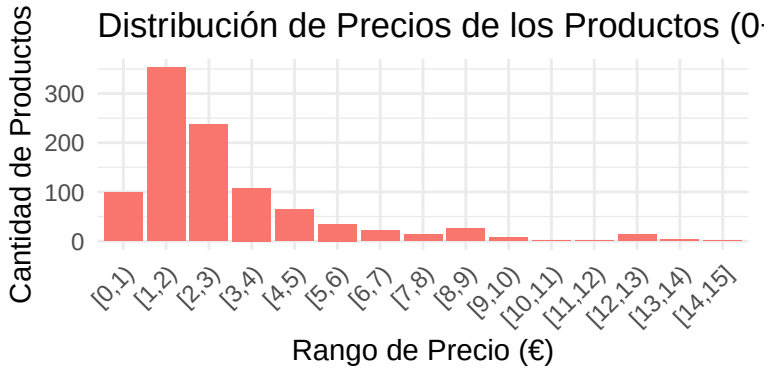
10) ¿Cuál es la tienda en la que más se ha comprado? (Dirección tienda)



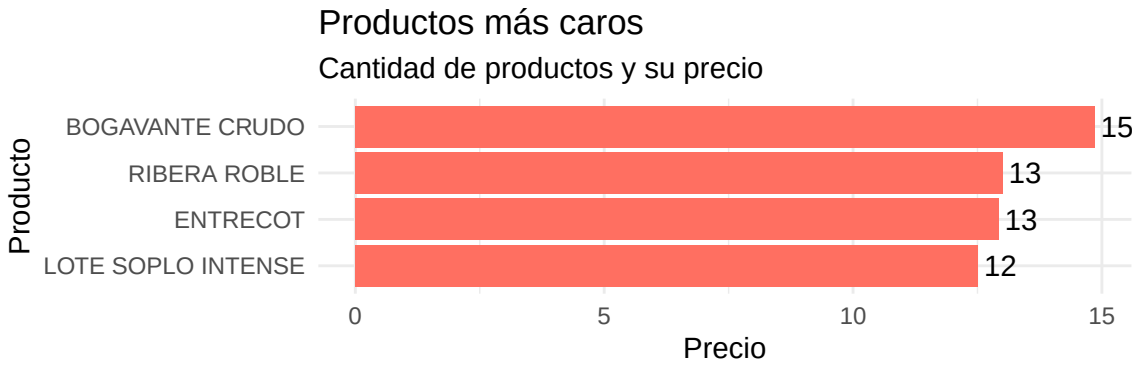
En esta gráfica muestra el número de compras que se ha hecho en los Mercadonas según los datos que tenemos, podemos observar que la dirección que más se repite es C/QUART 120, donde se ha comprado un total de 51 ocasiones; esta calle está situada en el corazón de Valencia (España) en una zona céntrica y bien comunicada. También hay que destacar la dirección C/ PINTOR PERIS ARAGÓ 34, la cual esta ubicada en el centro de Alboraya, donde se ha comprado 50 veces, lo cual concuerda con la anterior gráfica de la pregunta 5

`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€.

4. Conclusiones

Para concluir, nos gustaría expresar los beneficios que nos ha supuesto realizar este trabajo. La resolución de las preguntas planteadas sobre los tickets de Mercadona y la extracción de las variables que tiene cada ticket nos han ayudado a saber más acerca de Mercadona, ya sean los producto más vendidos en diferentes categorías (fruta, pescado), la hora de más afluencia de gente comprando, el dinero de media que se gastan los compradores...

Este proyecto ha sido fundamental para ampliar nuestro conocimiento sobre el conjunto de datos. Además, hemos adquirido experiencia en la resolución de los desafíos comunes que surgen al trabajar con datos. Esta experiencia nos ha permitido comprender mejor las complejidades involucradas en el análisis de datos y cómo abordarlas de manera efectiva en futuros proyectos.

Este proyecto ha sido de suma importancia, ya que nos ha proporcionado la invaluable experiencia de abordar una problemática desde cero, trabajando con datos y extrayendo conclusiones significativas.

Es importante señalar que los resultados de este proyecto deben ser interpretados con precaución debido a la limitada cantidad de datos disponibles. En estudios similares, se suelen analizar millones de tickets electrónicos para obtener conclusiones más sólidas y representativas de los hábitos de compra.

Este proyecto ha sido una oportunidad invaluable para mejorar nuestras habilidades de trabajo en equipo. Durante el proceso, hemos enfrentado decisiones conjuntas que han fortalecido nuestra colaboración y capacidad para trabajar de manera efectiva como grupo. Estas experiencias seguramente nos serán útiles en futuros proyectos, donde la capacidad de tomar decisiones en equipo juega un papel crucial en el éxito general del trabajo.

En definitiva, este trabajo nos ha servido para mejorar en valores y para iniciarnos en el mundo de los datos y su análisis.