MEMORIA TRATAMIENTO DE DATOS

Laura Horjales Rivas Arnau Hernández Lucas Josep Vicent Morales Martorell
Rocío Bono Moreno Pau Pérez García

2025-05-10

Contents

1 Introducción

2 Contexto del proyecto

En la actualidad donde la información ha pasado a ser uno de los elementos más preciados, el análisis de datos juega un papel muy importante tanto en la toma de decisiones estratégicas como operativas. La gran cantidad de datos ha impulsado la necesidad de convertir estos mismos en conocimientos aplicables, lo que convierte el análisis exploratorio en una herramienta esencial en múltiples sectores.

Como futuros científicos de datos, nuestro trabajo consistirá en extraer la información más relevante de una gran cantidad de datos. En este proyecto, los tickets representaran nuestra fuente de datos, donde mediante estos, podremos descubrir: preferencias, hábitos y comportamientos de los consumidores. A través de ellos, también podemos interpretar patrones ocultos, prever necesidades futuras y generar estrategias que optimicen tanto la experiencia del cliente como la gestión empresarial.

Analizando los registros de venta recogidos en los tickets, es posible obtener conclusiones relevantes para áreas como el control de inventario, el diseño de promociones o la personalización de servicios. Esto no solo mejora la eficiencia interna, sino que permite responder de forma ágil y efectiva a las demandas del mercado.

2.1 Nuestro Proyecto

Este proyecto nace con la necesidad de convertir datos sin procesar - provenientes de tickets de Mercadona- en información valiosa mediante técnicas de limpieza, transformación y análisis exploratorio. A través de una serie de scripts en R, hemos diseñado un modelo de trabajo que permite capturar los aspectos más relevantes del consumo cotidiano, con el objetivo de ofrecer herramientas útiles para la toma de decisiones comerciales.

La metodología utilizada combina la programación en R con recursos de visualización y manipulación de datos. Nuestro trabajo no se limita en organizar los datos, sino que busca identificar correlaciones, tendencias y patrones de comportamiento. A través de este proceso, contribuimos al entendimiento de dinámicas reales del mercado y fomentamos un entorno donde el dato procesado se convierte en una ventaja competitiva.

2.2 Recursos utilizados y estructura de trabajo

Para llevar a cabo este análisis, hemos empleado librerías fundamentales de R como readr, dplyr, stringr, tibble y ggplot2, entre otras. Estas herramientas nos han permitido estructurar los datos de manera eficiente, realizar filtrados específicos, transformar cadenas de texto y crear visualizaciones útiles.

El proceso inicial incluyó la conversión de los archivos PDF a texto plano mediante el uso de las funciones de la librería pdftools. Luego, se procedió a extraer la información de interés, organizándola en dataframes preparados para su análisis posterior. Una parte esencial fue la limpieza de los datos: corregir errores, homogeneizar formatos y validar los campos esenciales para obtener un dataset sólido y confiable.

2.3 Variables y estruturas de los datos

En esta primera fase, trabajamos con un conjunto de diferentes variables , que abarcan tanto información general del ticket (fecha, hora, importe total, tienda, caja, número de ticket) como detalles específicos de los productos adquiridos (nombre del producto, cantidad, precio unitario, tipo de producto, peso o unidad, entre otros).

2.4 Preguntas obligatorias:

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

Realizando el análisis de top 5 ventas del Mercadona, podemos observar que los 5 productos más vendidos por unidades son: el atún claro oliva(62 unidades), queso lonchas cabra(53 unidades), bolsa plástico (51 unidades), leche desnatada de calcio (49 unidades) y, por último, yogur coco (40 unidades). Este resultado, muestra una gran tendencia hacia los lácteos ya que 3 de estos 5 productos son lácteos.

Table 1: Productos más Vendidos por nidad

Descripcion	Total_Unidades_Vendidas
ATUN CLARO OLIVA	62
BOLSA PLASTICO	50
QUESO LONCHAS CABRA	43
YOGUR COCO	40
PAN SEMILLAS	38

Pregunta 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?

Dentro de la categoría de FRUTAS y VERDURAS los 5 productos más vendidos son: el plátano (vendiendo un total de 62.868 kg), la banana (28.140 kg), la sandia baja semillas (22.843 kg), el pepino (19.624 kg) y el calabacín verde (17.754 kg).

Table 2: Productos más Vendidos

Descripcion	total_kilos_vendidos
PLATANO	58.376
BANANA	25.978
SANDIA BAJA SEMILLAS	22.843
PEPINO	19.624
MELON PIEL SAPO	16.178

Pregunta 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?

Table 3: Pescados Más Vendidos

Descripcion	Peso_Total	Cantidad
LENGUADO FRESC EUROP	1.248	3
SEPIA FRESCA	1.602	3
SEPIA LONJA	2.000	3
SEPIA SUCIA REFRIG	2.402	3
DORADA	2.934	2

En la tabla 3 se puede observar los 5 tipos de pescado más vendidos junto a la cantidad en kg de cada uno. Se puede apreciar facilmente que la cantidad vendida de cada uno es muy similar.

Esta tabla tambien nos muestra que apesar de que la dorada dentro de los 5 peces más vendidod es el que menos se consume, lo compensa en kg comprados, la cual cosa puede explicar porque se vende menos.

Pregunta 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 ello primero sacaremos todas las veces que se ha comprado bananas y platanos(como necesitamos que este ordenado por la fecha haremos un inner_join)

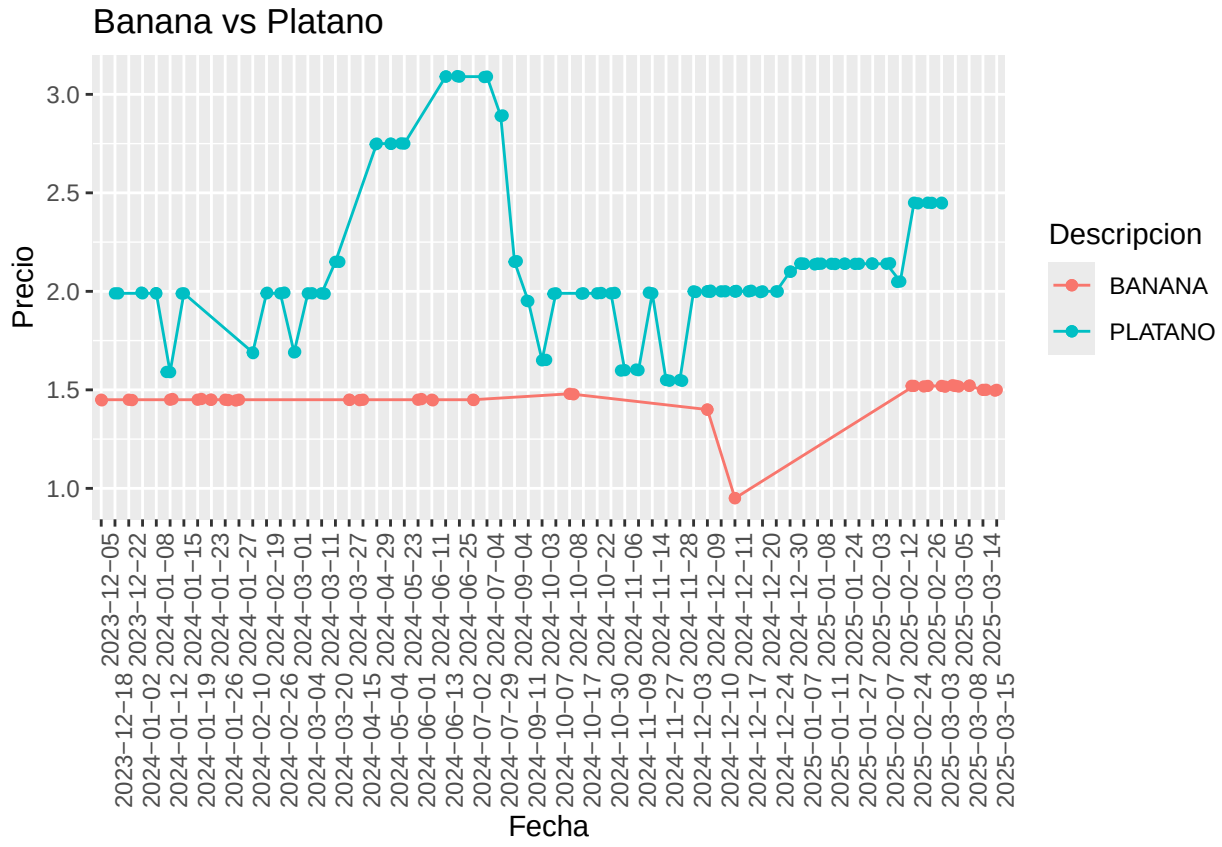

Figure 1: Evolución Precio Bananas y Platanos

Se puede observar en las gráficas que el precio del platano es mucho más variado, es decir desde marzo a julio tiene un valor superior y luego baja. En cambio la banana tiene un valor más constante y independientemente de la temporada tiene prácticamente el mismo precio.

También hay que considerar que tenemos el doble de observaciones de platano por lo que puede ser que no sea tan precisa la representación de la banana, donde también se puede ver que hay un outlier el dia 12 de noviembre de 2024.

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

Para esto ya hemos creado una columna llamada Ciudad donde estan las procedencias de los tickets. Para saber las ciudades donde se compra más agruparemos por ciudades i contaremos el numero de casos que hay.

Table 4: Procedencia Tiquets

Ciudad	Compras
VALENCIA	120
ALBORAIA/ALBORAYA	50
BURJASSOT	27
MURO DE ALCOY	24
ALCOI/ALCOY	22
BUÑOL	12
BENIJOFAR	8
ALGINET	5
MONCADA	5
BENIFAIO	4
GARRUCHA	4
VERA	3
GANDIA	1
ONTINYENT	1
REQUENA	1
SAGUNT/SAGUNTO	1
SAN ANTONIO DE BENAGEBER	1
SANT JOSEP DE SA TALAIA	1
SANTA EULÀRIA DES RIU	1

Como era previsible al ser los tickets de los alumnos, era mucha más probable que estos fueran de Valencia o de alrededores, por lo que hay 120 de Valencia y a continuación Alboraya y Burjassot.

Pregunta 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?

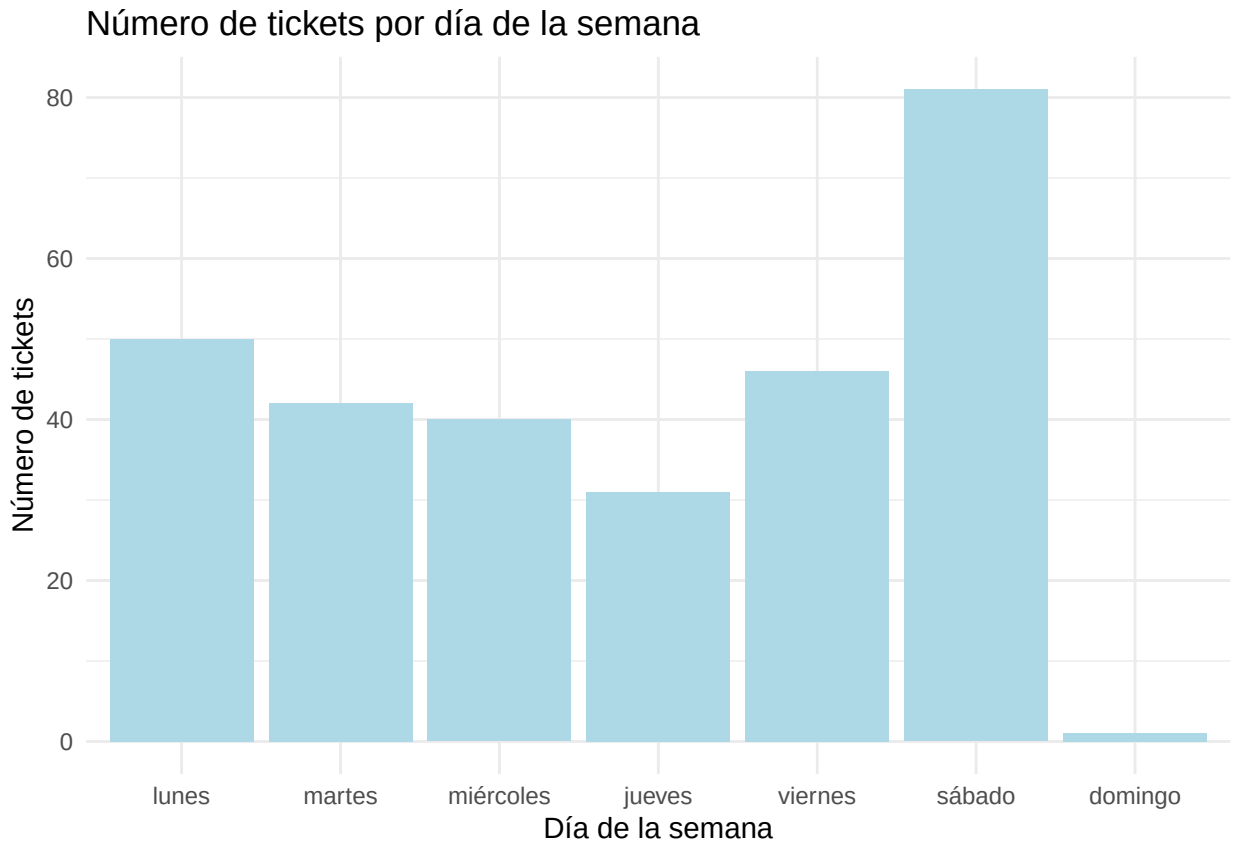

Figure 2: Número de tickets por día de la semana

Según el diagrama el día entre semana con menos tickets y por lo tanto ventas, es el jueves. Por lo que lo más inteligente sería cerrar este día, ya que es cuando menos gente va a Mercadona.

2.5 Resto de preguntas propuestas:

Pregunta 7 : ¿Cuál es la hora más habitual para realizar la compra? ¿Este horario varía entre los días laborales y los fines de semana?

Table 5: Hora más Habitual Comprar

Tipo_dia	Hora_hora	N_compras
Fin de semana	13	20
Laboral	20	48

Como podemos comprobar el fin de semana la hora con más compras (21) es a la 13 a diferencia que los dias laborales que es a las 20 con 48 tickets, esto puede darse, debido a que entre semana tenemos obligaciones y extraescolares que atender y por eso las 20 es más frecuente mientras que los fines de semana aprovechamos las mañanas para poder comprar y tener unas tardes más relajadas.

Pregunta 8 : ¿Existe alguna diferencia clara en el perfil de compra entre los días de semana y los fines de semana?

Table 6: Diferencias entre Compra Fin de Semana y Entre

Tipo_dia	Descripcion	Frecuencia
Laboral	PLATANO	50
Laboral	QUESO LONCHAS CABRA	32
Laboral	BOLSA PLASTICO	31
Laboral	FILETE PECHUGA	25
Laboral	UVA BLANCA S/SEM	24
Fin de semana	PEPINO	18
Fin de semana	ALMENDRA NATURAL	16
Fin de semana	CACAHUETE CHOCOLATE	15
Fin de semana	KIWI VERDE	15
Fin de semana	Z.NARANJA 330ML REF.	14

Durante los fines de semana, los productos más comprados tienden a ser snacks, frutos secos y frutas como el pepino, kiwi verde y el zumo de naranja.En cambio, los días laborales, los productos más vendidos son elementos básicos como plátanos, queso en lonchas o carne, la gran compra de bolsas nos indica que entre semana se realizan compras más grande, por eso, su gran cantidad de venta, mientras que los fines de semana es una compra más oriendata al consumo inmediato o pequeños caprichos.

Pregunta 9 : ¿El precio total de la compra varían según la ciudad o zona?

Table 7: Precio Por Zona

Ciudad	n_tickets	media_total	mediana	sd_total	minimo	maximo
ONTINYENT	1	112.02000	112.020	NA	112.02	112.02
ALCOI/ALCOY	22	102.55409	128.645	71.4846960	9.66	234.20
REQUENA	1	89.18000	89.180	NA	89.18	89.18
MONCADA	5	66.98800	57.330	25.5122455	49.94	111.08
SAGUNT/SAGUNTO	1	66.30000	66.300	NA	66.30	66.30
BENIFAIO	4	64.68750	62.350	48.3358089	12.56	121.49
BUÑOL	12	58.12167	60.845	28.5231794	18.90	105.91
VERA	3	52.07333	52.120	0.8609491	51.19	52.91
SAN ANTONIO DE BENAGEBER	1	49.74000	49.740	NA	49.74	49.74
BENIJOFAR	8	48.12125	46.165	22.4782488	6.27	80.53
MURO DE ALCOY	24	45.02083	32.490	36.0782802	3.20	161.22
VALENCIA	120	41.18833	34.655	31.0746227	0.43	187.52
BURJASSOT	27	38.15846	35.515	17.5759747	3.07	82.81
ALGINET	5	33.48200	30.940	16.0079221	18.64	55.35
ALBORAIA/ALBORAYA	50	32.19840	24.200	23.8732468	1.75	90.03
GARRUCHA	4	24.89250	24.855	11.7489556	12.10	37.76
GANDIA	1	3.90000	3.900	NA	3.90	3.90
SANTA EULÀRIA DES RIU	1	3.08000	3.080	NA	3.08	3.08
SANT JOSEP DE SA TALAIA	1	0.77000	0.770	NA	0.77	0.77

Para verlo de manera más visual, podemos hacaer un diagrama de boxplot, donde la linea central refleja la mediana y los cuartiles la caja.

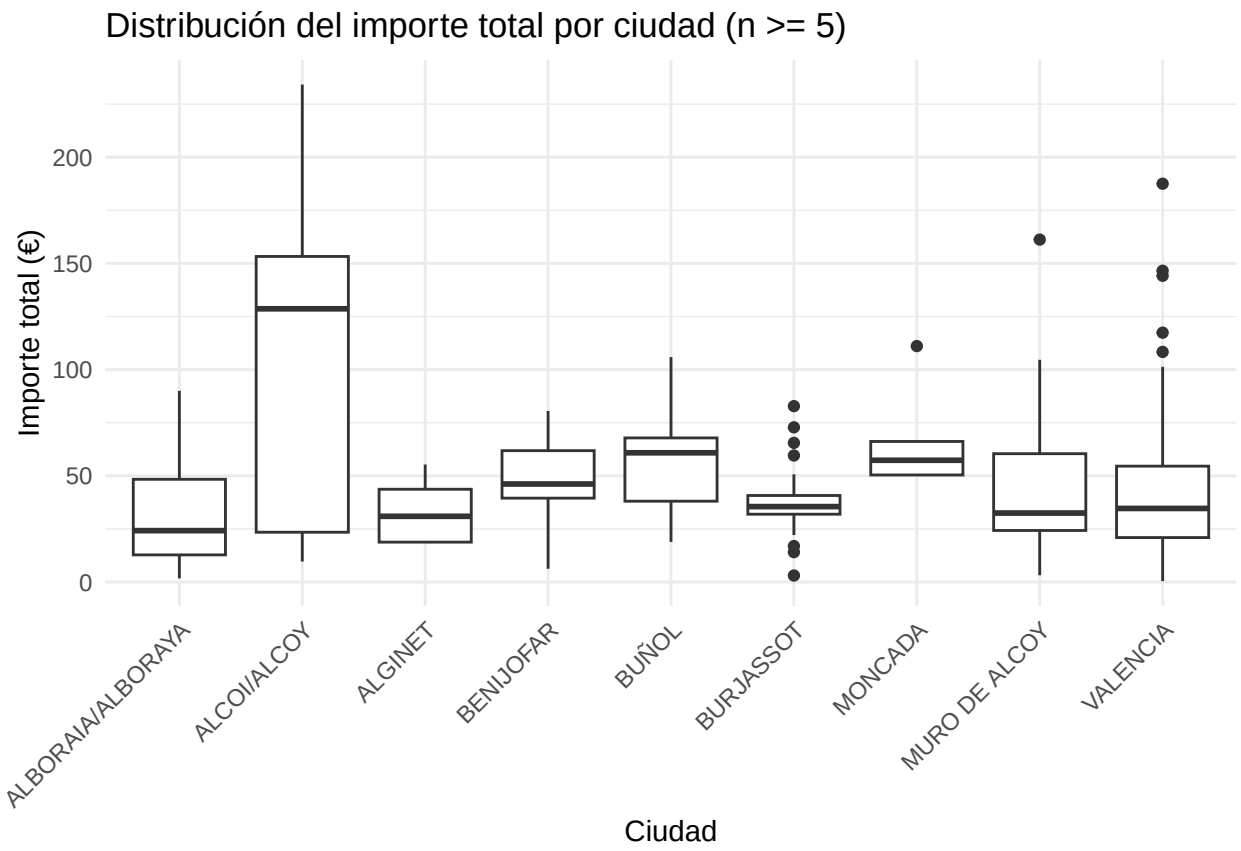

Figure 3: Precio Por Zona

Como podemos ver en el boxplot Alcoy es la ciudad con más tickets y donde más se gasta en promedio, ciudades como valencia y burjassot tienen una media bastante estable en todas las compras aunque encontramos casos donde hay compras de más de 100€. La elevada media de Alcoy puede deberse a que la mayoría de los tickets son de compras muy altas lo que también podría estar relacionada con los precios en ese Mercadona, mientras en los otros municipios la media se mantiene bastante constante y dentro de rangos más bajos.

Pregunta 10 : ¿Cuánto dinero de media se gasta cada cliente en una compra?

Table 8: Media Gasto por Cliente

n_compras	gasto_medio	gasto_mediana	sd_gasto
291	45.88152	37.05	37.9408

En nuestro estudio tenemos 291 tickets, el gasto medio de cada cliente es de 45,8€ por compra, mientras que la mediana nos indica que el 50% de las compras cuestan menos de 37,05. La desviación de 38 nos indica que el gasto por compra varía bastante de un cliente a otro, es decir hay compras pequeñas de 20€y otras muy grandes de más de 100€.

Pregunta 11 : ¿Cuál es el mes o período en el que más gastos se realizan? ¿Durante las vacaciones de navidad?

Para resolver esta pregunta podemos realizar un gráfico de barras para observar mejor la información. En este gráfico se representan los datos de distintos tickets de 3 años diferentes (2023, 2024 y 2025). Al tener diferentes cantidades de tickets por mes de un año a otro, tenemos que decir un período por cada año de tickets (en este caso 3 meses). Por lo que este período no tiene porque coincidir todos los años.

Table 9: Gastos Mensuales

Año	Mes	Gasto_Total
2024	mar	1545.86
2025	feb	1422.81
2024	ene	1214.46
2024	feb	1004.36
2025	ene	960.52
2024	dic	948.86

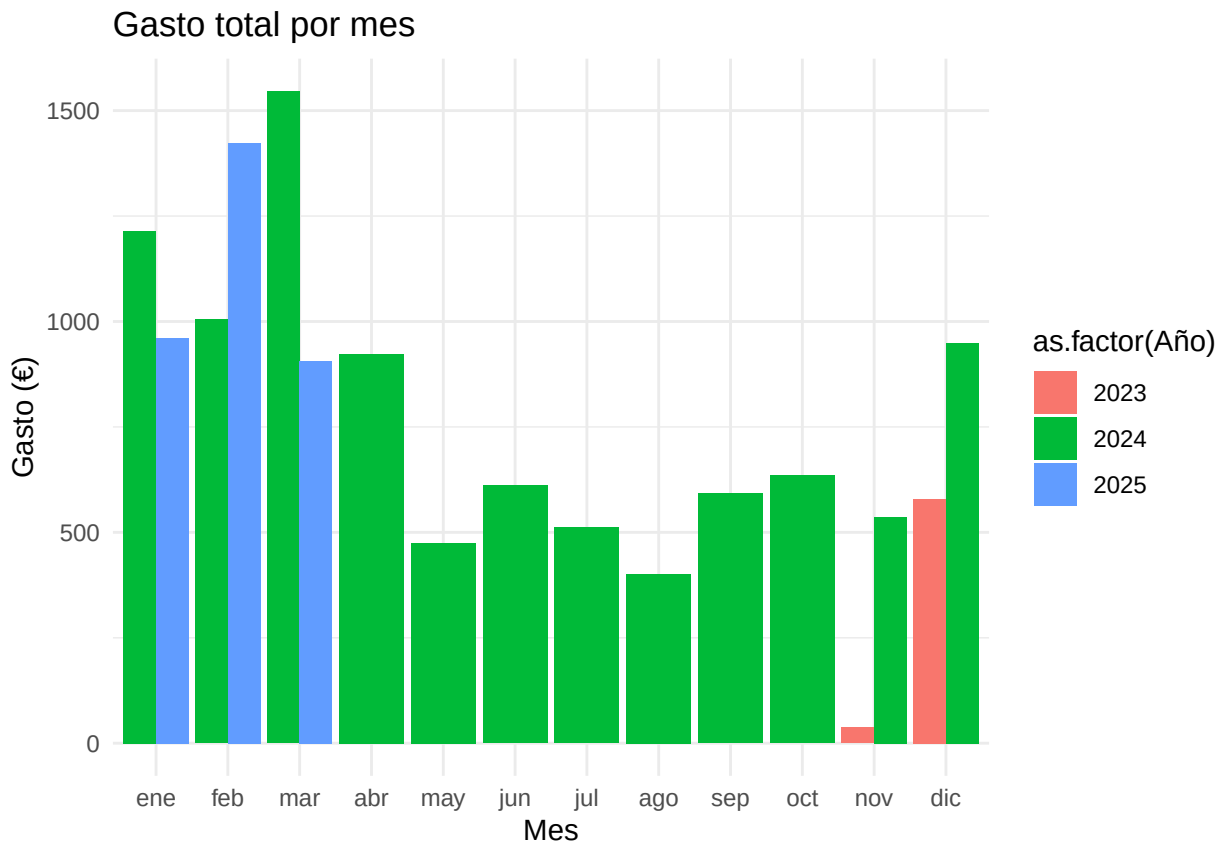

Figure 4: Gasto Mensual Total

Si nos fijamos en los tickets del año 2023 (color rosa) el mes con más gastos es diciembre. Fijandonos en el año 2024 (verde), el período con más gastos es en enero (siendo el único año con tickets todos los meses). En el año 2025 (azul) es febrero el mes con más gastos. Por lo que el período con más gastos no es en la época de Navidad.

Pregunta 12 : ¿Qué productos suelen comprarse juntos?

Table 10: Producto Comprados Juntos

Descripcion.x	Descripcion.y	Frecuencia
ALMENDRA NATURAL	CACAHUETE CHOCOLATE	16
ALMENDRA NATURAL	PEPINO	16
CACAHUETE CHOCOLATE	ALMENDRA NATURAL	16
CACAHUETE CHOCOLATE	PEPINO	16
PEPINO	ALMENDRA NATURAL	16
PEPINO	CACAHUETE CHOCOLATE	16
ALMENDRA NATURAL	BARRITA MUESLI CHOCO	14
ATUN CLARO OLIVA	PAN SEMILLAS	14
BARRITA MUESLI CHOCO	ALMENDRA NATURAL	14
BARRITA MUESLI CHOCO	CACAHUETE CHOCOLATE	14

Se observa que “almendra natural” se suele comprar con “cacahuete chocolate”, “pepino” o “barrita muesli choco” y el “pan semillas” con “atun claro oliva”. Esto también ocurre a la viceversa.

Pregunta 13 : ¿Cuál es el producto más caro registrado en los tickets?

Table 11: Producto más Caro En Tiquets

	index	Cantidad	Descripcion	Precio	Importe	Peso	Tipo
2421	52	1	ALISTADO MEDIANO	31.33	31.33	1	

El producto más caro registrado es el “alistado mediano”, que tiene un precio de 31.33 euros.

Pregunta 14 : ¿Hay una cantidad media de productos por tickets?

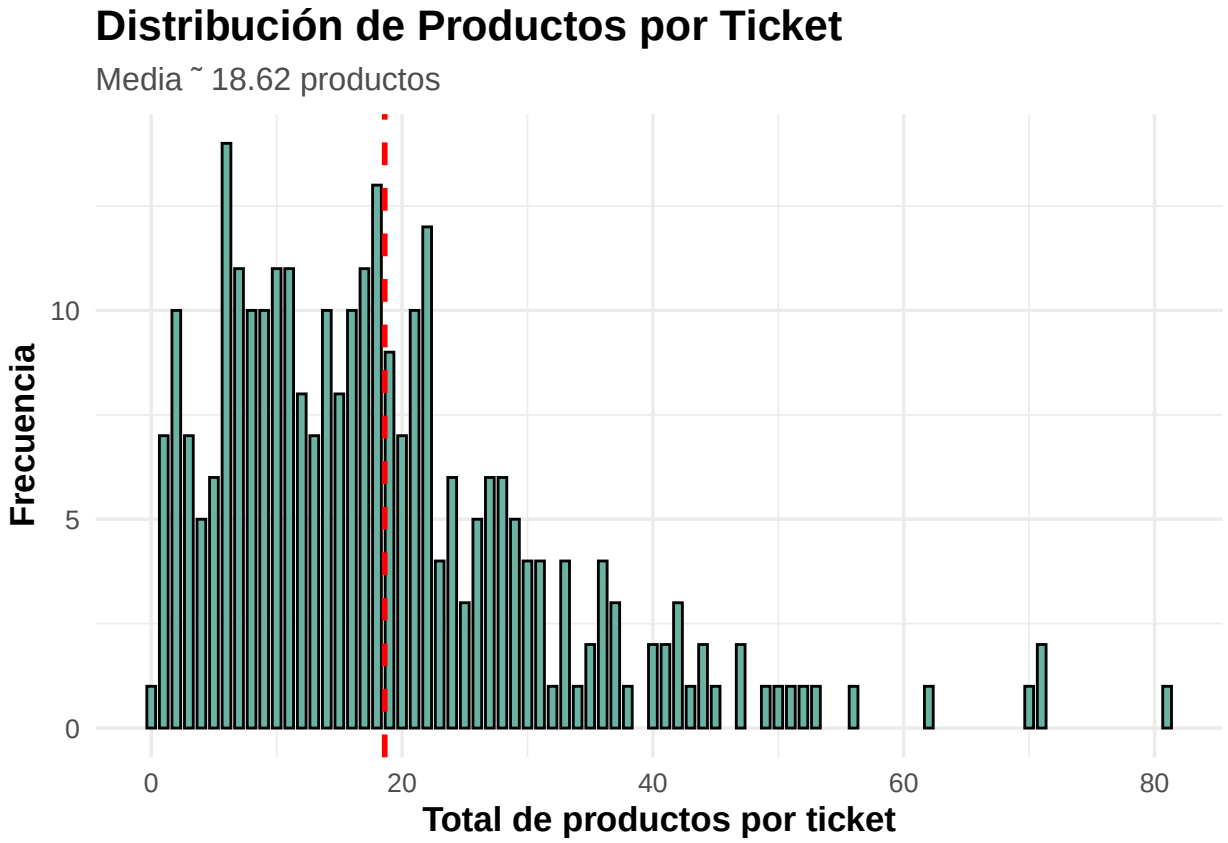

Figure 5: Cantidad Media Porductos en Tiquet

La cantidad media de productos por ticket es de 18.61856 productos.

Pregunta 15 : ¿Cuando se hacen más compras en las diferentes partes del día: mañana, tarde y noche?

Para esta pregunta separaremos el día en mañana (9:00 - 13:59), tarde (14:00 - 18:59) y noche (19:00 - 21:45). Para ello crearemos una nueva columna.

Table 12: Compras Durante el mañana, tarde y noche

parte_dia	Compras
noche	106
tarde	104
mañana	81

Podemos observar que durante la noche y la tarde se hacen muchas más compras que en comparación por la mañana, ya que las compras entre semana se hacen más usualmente por la tarde, ya que por la mañana las personas normalmente trabajan o estudían. Por esto queremos ir más allá y vamos a separarlo entre dias de la semana y dias del fin de semana

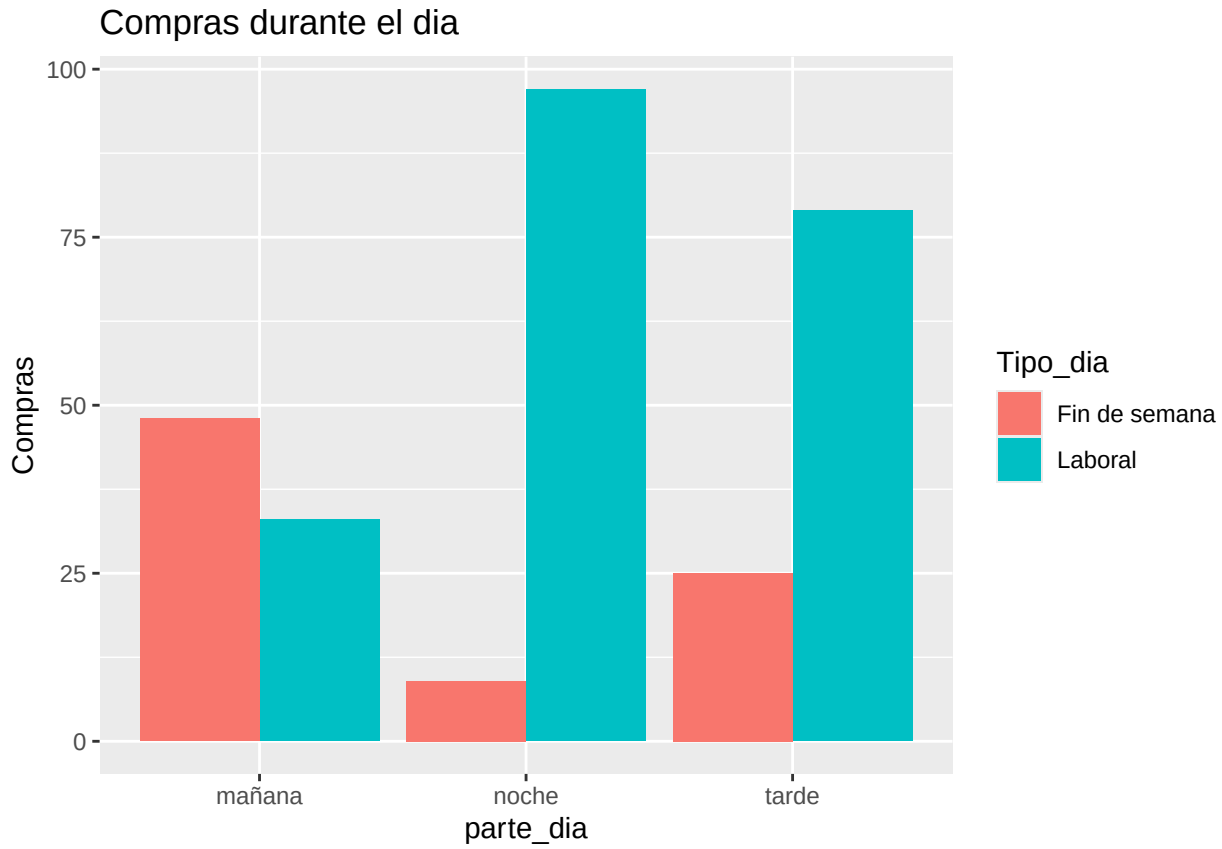

Figure 6: (#fig:figura-comp_M_T)Comparativa Compra en Distintos Momentos del Día

Se confirma lo que pensabamos se va más por la mañana en el fin de semana en cambio durante los dias laborales por la tarde y la noche se va mucho más a comprar

Pregunta 16 : ¿Que productos son los más comprados durante las diferentes estaciones, verano (junio - agosto), invierno(diciembre - marzo)... ?

Para ello haremos los mismo que hemos hecho previamente pero esta vez dividiendo por estaciones: verano (junio - agosto), invierno(diciembre - febrero), otoño (septiembre - noviembre) y primavera (marzo - mayo)

Table 13: 5 Productos Más Vendidos por Temporada

estacion	Descripcion	Frecuencia
invierno	PLATANO	30
invierno	PAN SEMILLAS	17
invierno	KIWI VERDE	15
invierno	QUESO LONCHAS CABRA	15
invierno	FILETE PECHUGA	14
primavera	TOMATE TRITURADO	12
primavera	UVA BLANCA S/SEM	9
primavera	BANANA	8
primavera	COPOS DE AVENA	8
primavera	NARANJA 5 KG.	8
verano	QUESO LONCHAS CABRA	10
verano	PAN SEMILLAS	9
verano	CANÓNIGOS	8
verano	FILETE PECHUGA	8
verano	FRESA	8
otoño	PLATANO	12
otoño	QUESO LONCHAS CABRA	10
otoño	12 HUEVOS GRANDES-L	9
otoño	FILETE PECHUGA	9
otoño	ACT O% NAT ED 8	8

Una de las cosas destacables ya ha sido comentado previamente en una de las preguntas, ya que en invierno y en otoño se puede ver que el platano es el producto más comprado cuando luego en la tabla de verano y primavera no aparece y en su defecto aparece la Banana. Esto es debido a que durante las estaciones de primavera y verano el precio del platano es mucho más superior al estar fuera de temporada. Igualmente con otras frutas que salen en el top únicamente en una estación o en dos, podríamos atribuir este hecho a que cada fruta tiene una temporada.

Pregunta 17 : ¿Cuánto supone económicamente de media el IVA en las compras?

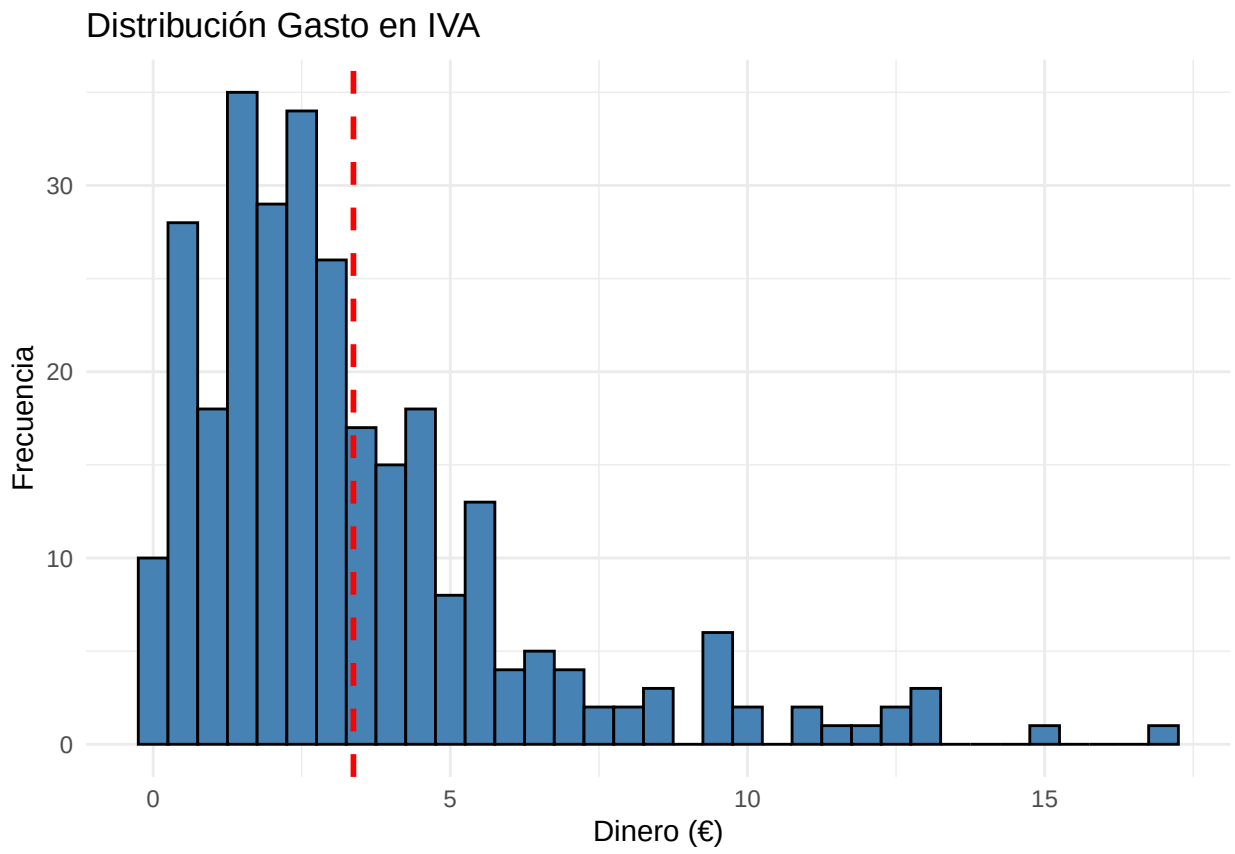

Figure 7: Peso IVA

En cada tiquet tenemos que dependiendo del producto se aplica un tipo de iva u otro, si sumamos la cantidad de dinero que esto representan para cada tiquet, podemos saber cual es la media.

Al realizar el calculo, podemos ver que da una medía de 3'37€ en IVA por tiquet respecto a 46€ de media de gasto por cada tiquet. A su vez mediante el gráfico 7 se puede ver que la distribución tiende a un bajo coste en IVA

Pregunta 18 :¿Como ha variado el precio del aceite, y relacionados, a lo largo del tiempo?

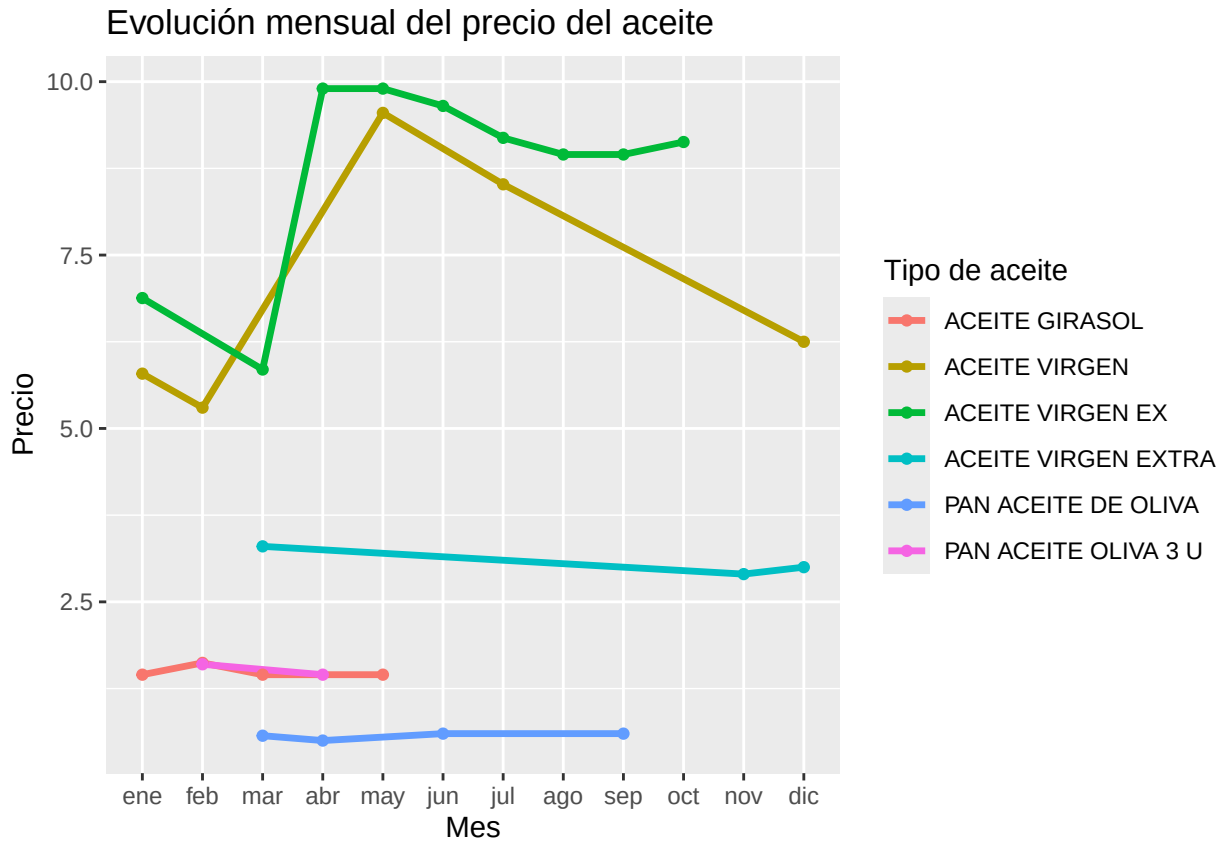

Figure 8: Variación Precio Aceite

En el gráfico @ref(fig: figura-precaceite) se puede observar que a lo largo del año las variaciones en el aceite no han sido apenas apreciables excepto en los casos del aceite virgen extra y el aceite virgen, que toman un augmento considerable de casi el doble de su precio prebio en marzo o febrero.

Pregunta 19 : ¿En que se ha gastado(€) más en pescado o en otros productos que van por peso? ¿Es equivalente al número de veces que se han comprado unidades de estos?

Table 14: Venta Pescado Fruta y Verdura

Tipo	Gastado	Compras
Fruta o Verdura	915.50	374
Pescado	243.03	27

En la tabla 14 podemos ver represenado el gasto en euros total tanto en pescado como en fruta o verduta, a su vez tambien se muestra la cantidad comprada de los mismos.

Se puede ver con claridad que tanto el producto más comprado es la fruta o verdura respecto a los de tipo pescado, con esto se observa debido a la gran diferencia de compra que en la fruta y verdura se ha gastado más.

Pregunta 20 :¿Hay una relación entre la cantidad de productos que se compra y si ha utilizado el parking o no?

Table 15: Relación Compra y Uso Parking

usado_parking	mediana_prod	total_prod
FALSE	13	262
TRUE	17	29

Como se ve en la tabla 15, cuando no se utiliza el parking, el número de productos es más bajo (mediana = 13). Cuando se utiliza el parking, el número de productos es más alto (mediana = 19).

Por lo que podemos concluir, que cuando se utiliza el parking es más probable que se compren más productos que si no se utiliza.

Pregunta 21 : ¿Cual es el tiempo medio que tardan en hacer la compra las personas que han utilizado el parking?

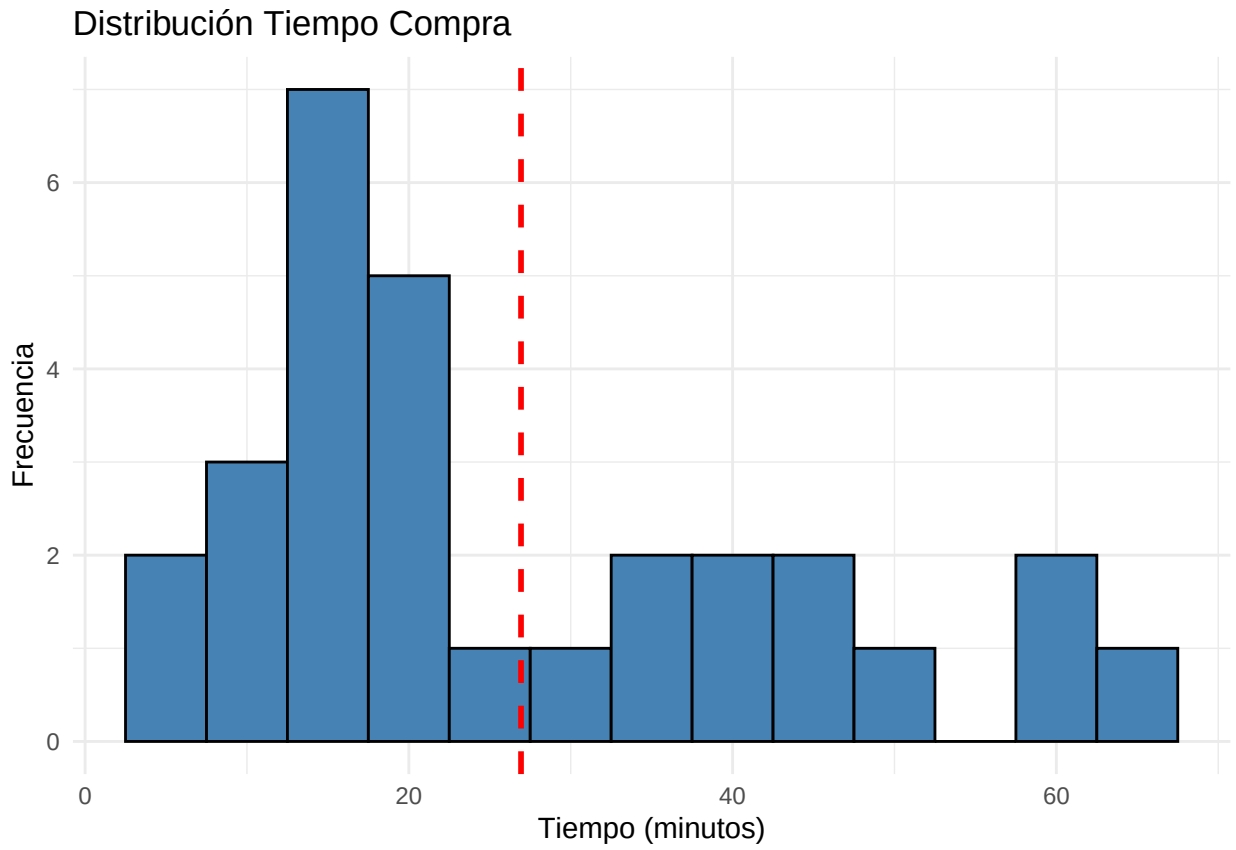

Figure 9: Tiempo medio Compra

El tiempo medio que tardan en hacer la compra las personas que utilizan el parking es de 27 minutos aproximadamente.

En el gráfico 9 se puede ver como se distribuyen los tiempos de compra mediante un histograma marcando con una línea roja el valor de la media mencionado.

3 Outlier

Un outlier (o valor atípico) es un dato que se aleja significativamente del resto de los valores en un conjunto de datos. No parece consiciente con el resto de los datos. Pueden ser datos con gran variabilidad en la media, errores experimentales, errores en la introducción de datos, fallos en el sistema de adquisición, cambios en las unidades... La presencia de estos valores atípicos puede producir una influencia en la media o la desviación típica.

4 Conclusión

Este proyecto nos ha permitido enfrentarnos a un problema real de tratamiento y análisis de datos, trabajando tickets (en PDF) y convirtiéndola en datos analizables. Hemos desarrollado un sistema capaz de analizar y limpiar automáticamente los tickets electrónicos de Mercadona, descartando el resto de tickets posibles. Siendo capaces de extraer diferentes variables como productos, cantidades, fechas, localizaciones y precios.

Durante el proyecto hemos sido capaces de identificar patrones de comportamiento del consumidor y diferencias según ciudades o días de la semana. Con este trabajo hemos realizado un análisis completo de datos en un entorno realista, pudiendo desarrollar habilidades técnicas y competencias necesarias para el trabajo en equipo.