PROYECTO TD 2024

Marta Molina, Iñaki Martín, Nerea Galera, Sergio Mut, Paula Sigüenza

2024-05-13

Índice

1 Introducción al trabajo.

Este proyecto trata de desarrollar un programa que permita analizar una serie de tickets de compras en un supermercado. El objetivo es enfrentarse a un problema real detratamiento de datos, realizando un seguimiento de (evolución de precios, compras más habituales, productosmás consumidos, supermercado habitual, hora de compra,etc). El proyecto ha sido realizado con control de cambios GIT. Tiene un repositorio compartido entre las miembros del grupo llamado ProyectoTD2024 en la plataforma GitHub y en el cual se ven los cambios y desarrollo del trabajo. Se dispone de varios tickets para empezar a trabajar y un júpiter notebook de python (TicketPDF2TXT.ipynb) que transforma todos los ficheros con extensión pdf presentes en la carpeta data en ficheros de texto con toda la información.

1.1 Carga de librerías y datos necesarios para el análisis:

En primer lugar, hemos cargado todas las librerías necesarias en las diferentes fases del proyecto. Haciendo esto de una forma eficiente, comprobando si el usuario tiene instalados los paquetes necesarios y descargandolos en caso necesario. A continuación, se realiza la carga del conjunto de datos que provienen de los tickets de un supermercado, en este caso, del Mercadona. Estos tickets se encuentran en dos tipos de formato, algunos están en .pdf y otros en .txt almacenados en la carpeta 'data' incluida en el repositorio del proyecto. Debido a la mezcla de formatos, para analizar nuestros datos necesitamos tener todos los archivos en formato .txt, para así poder analizar los correctamente, por lo que usaremos el archivo "TicketPDF2TXT.ipynb" que nos ha proporcionado el profesor de la asignatura, contenido también en nuestro repertorio, que automáticamente

transforma los tickets en formato .pdf a formato .txt. Para poder leer la información de forma adecuada y clarificadora creamos distintos dataframes con la distinta información que nos puede aportar cada uno de los tickets.

Para ello, hemos de crear distintos dataframes, ya que, un mismo ticket no presenta siempre la misma estructura. Es decir, en un ticket podemos encontrar frutas y verduras (que tienen un formato distinto, el precio no es por unidad, si no por peso), pescados (que también siguen otro tipo de formato, el precio también va por peso, y, además, encontraremos un string que indicará la sección 'Pescados'), productos restantes, una serie de datos sobre cada Mercadona, información sobre el servicio Parking. De esta forma, uniendo los diferentes dataframes creados(mediante un XXX_join()), obtenemos un dataframe final, con todos los datos de cada ticket en un formato correcto(df_tickets). Sin embargo, observando todos los tickets recabados, el inicio es siempre igual, por lo que creamos una variable en la que guardamos las primeras ocho líneas. Luego almacenamos los productos en distintas listas, almacenando finalmente cuatro listas: una total y tres separadas en productos, pescado y fruta. Además sabemos que algunos supermercados tienen parking y otros no por lo que creamos una variable para detectar aquellos tickets que presenten parking y los que no. Seguidamente, hemos creado una función que detecte donde terminan los productos y donde empieza el final del ticket, ya que cada ticket tiene un número distinto de productos, y así podemos observar las similitudes entre el final de los tickets. También utilizamos dos funciones específicas para analizar la fruta y el pescado, ya que tiene su propio formato. En este momento de la importación, realizamos un data.frame conjunto, teniendo en cuenta el identificador de cada ticket, y separando con la fruta, el pescado y los productos resultantes.

Finalmente obtenemos un data.frame con absolutamente toda la información contenida en el ticket pero de forma concisa, estructurada, ordenada y correcta.

1.2 Características generales de los datos:

Es posible hacerse una idea rápida de cuáles son los datos que contiene el data.frame, 'df_tickets', haciendo uso de la función glimpse, de la librería dplyr. En resumen, nuestro dataframe cuenta con 16 variables, con las cuales vamos a poder responder a nuestras preguntas ya que se corresponden con el contenido del ticket ordenado.

Variable	Descripción
Productos	Nombre y cantidad del producto
PUnidad	Precio por unidad del producto
Cantidad	Cantidad del producto
Nombre	Nombre del producto
Importe	Precio del producto
OP	Identificador de la operación
Kgs	Peso (en kilogramos) del producto
Eur_el_kg	Precio del producto por kg
Info	Información del producto
Factura_simpl	Factura de la compra
Total_euros	Importe total de la compra
Calle	Calle del supermercado
Municipio	Municipio donde se encuentra el supermercado
Fecha	Fecha de la compra
nombre_super	Información del supermercado
Parking	Si el cliente ha utilizado el parking o no

1.3 Análisis de missing data en nuestro conjunto de interés:

En este subapartado pasaremos a detectar datos o valores anómalos, ya sean NA's (valores perdidos) u outliers (valores'alejados' del resto). En primer lugar, respecto a los NA's, podemos ver que solamente existen en 4 columnas. Existen NA´s en las columnas (vbles.) PUnidad y en Total_euoros cuando se hace referencia a las frutas y veduras y a pescados. Esto tiene sentido, ya que, ambos dos tienen un formatodiferente al resto de productos, como ya hemos comentado anteriormente. Además, existen NA´s en las columnas (vbles.) Kgs, Eur_el_kg y en Info cuando se hace referencia al resto de productos que no sean ni frutas ni verduras ni pescados. Esto también tiene sentido, ya que, complementa a lo anterior, es decir, el resto de productos no constan de ese tipo de formato. En segundo lugar, para la detección de outliers sólo hace falta visualizar la gráfica (boxplot) dónde constan los 4 métodos vistos en la asignatura de Tratamiento de los Datos:

`reglasigma(x)`, `reglahampel(x)`, `reglaboxplot(x)` y`reglapercentil(x)`.

Ahora, se hará una breve introducción a cada método:

• Regla sigma 3 (reglasigma(x)): Este método asume que los datos siguen una distribución gaussiana, es decir, una forma de campana. Utiliza la media y la desviación estándar para caracterizar los datos. Según esta regla, los valores que están más allá de 3 desviaciones estándar de la media se consideran outliers.

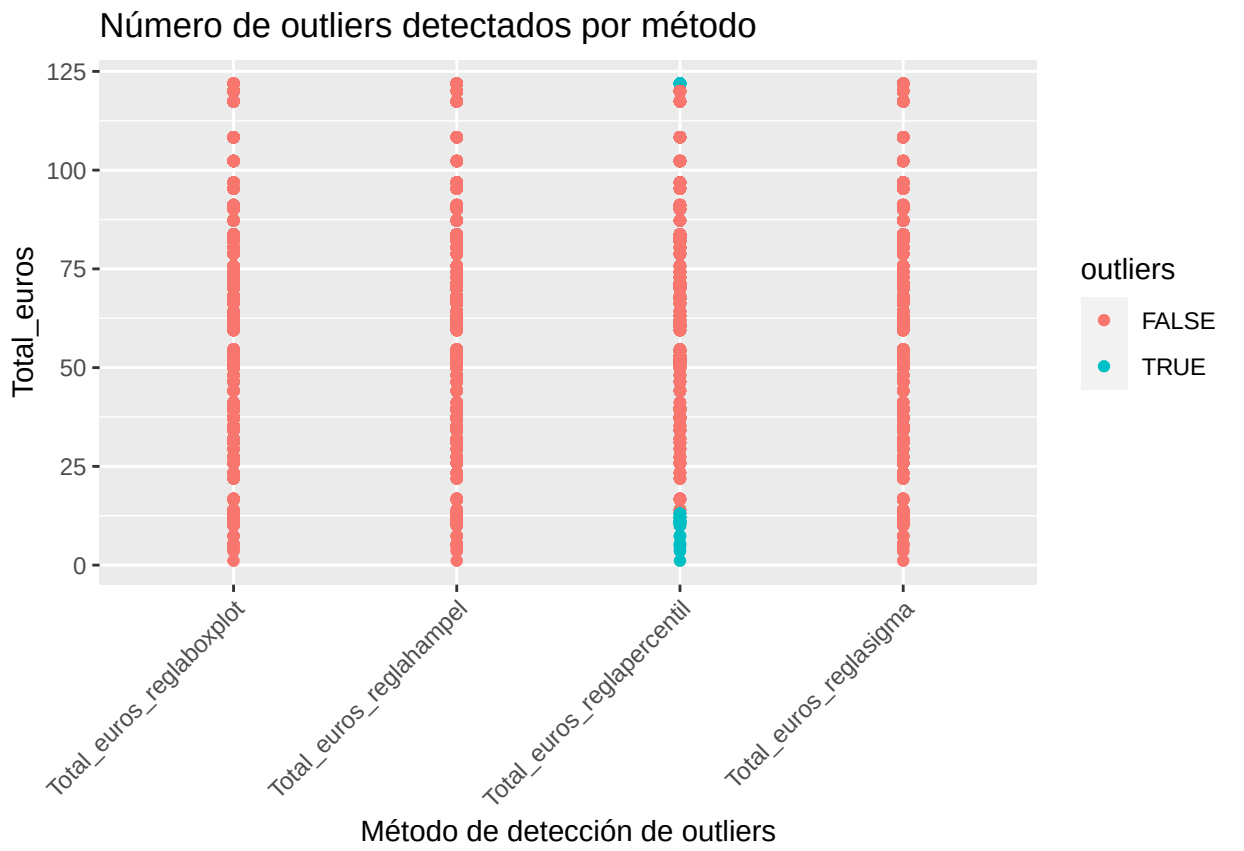
• Identificador Hampel (reglahampel(x)): A diferencia de la regla sigma, este método no asume que los datos sigan una distribución gaussiana. En su lugar, utiliza estimadores robustos como la mediana y la desviación absoluta mediana(MADM) para caracterizar los datos. Los valores que están más allá de 3 veces la MADM se consideran outliers. • Regla boxplot (reglaboxplot(x)): Este método utiliza un diagrama de caja (boxplot) para visualizar la distribución de los datos. Define los valores típicos superiores e inferiores hasta el cuartil 75% y 25% respectivamente, y la dispersión se calcula como el rango intercuartílico (IQR). Los valores que están más allá de 1.5 veces el IQR por encima del cuartil 75% o por debajo del cuartil 25% se consideran outliers. • Percentiles (reglapercentil(x)): Este método considera que cualquier valor que esté fuera del rango del percentil 5% y 95% se puede considerar como atípico. Es decir, los valores que están por debajo del percentil 5% o por encima del percentil 95% se consideran outliers.

```
## numeric(0)
```

```
## numeric(0)
```

```
## numeric(0)
```

```
## numeric(0)
```




```
## Saving 6.5 x 4.5 in image
```

FIGURA 1: Gráfica de outliers

Como se puede observar en el gráfico de la figura 1,en cada columna encontramos un método distinto para encontrar outliers. Vemos que en el método de percentiles tenemos outliers, podemos apreciarlo por los puntos de otro color azul. Además, nuestra gráfica está compuesta por líneas con puntos, ya que cada una de estas es un ticket diferente, por lo que los outliers se presentan en unos tickets concretos no en todos. Si nos fijamos bien, sólo aparecen outliers en la regla percentil ya que, este es el método más sensible, es decir detecta los outliers con mayor precisión, como también vimos en la Práctica 5 de la asignatura.

2 Exploración / visualización.

Una vez hemos asegurado que nuestros datos están correctamente en el dataframe, y, observando que tienen los valores correctamente etiquetados y están almacenados con el tipo correcto, además de conocer el origen de los NA, podemos empezar a buscar posibles patrones entre las variables de los tickets. Tras observar nuestros datos, nos surgen una serie depreguntas:

-¿Influye la hora de compra con el número de tickets enesos intervalos de tiempo?, y por consiguiente, ¿influye en el precio total?

-¿Qué productos tienen la mayor variabilidad en los precios? ¿Los precios de los productos se mantienen en todos los supermercados?

-¿Qué frutas y pescados han sido los más consumidos por los clientes? ¿Y qué productos han sido los más vendidos?

-¿Existe diferencia en el tamaño del ticket de compra entre aquellos clientes que utilizan el servicio de parking del supermercado y aquellos que no lo utilizan?

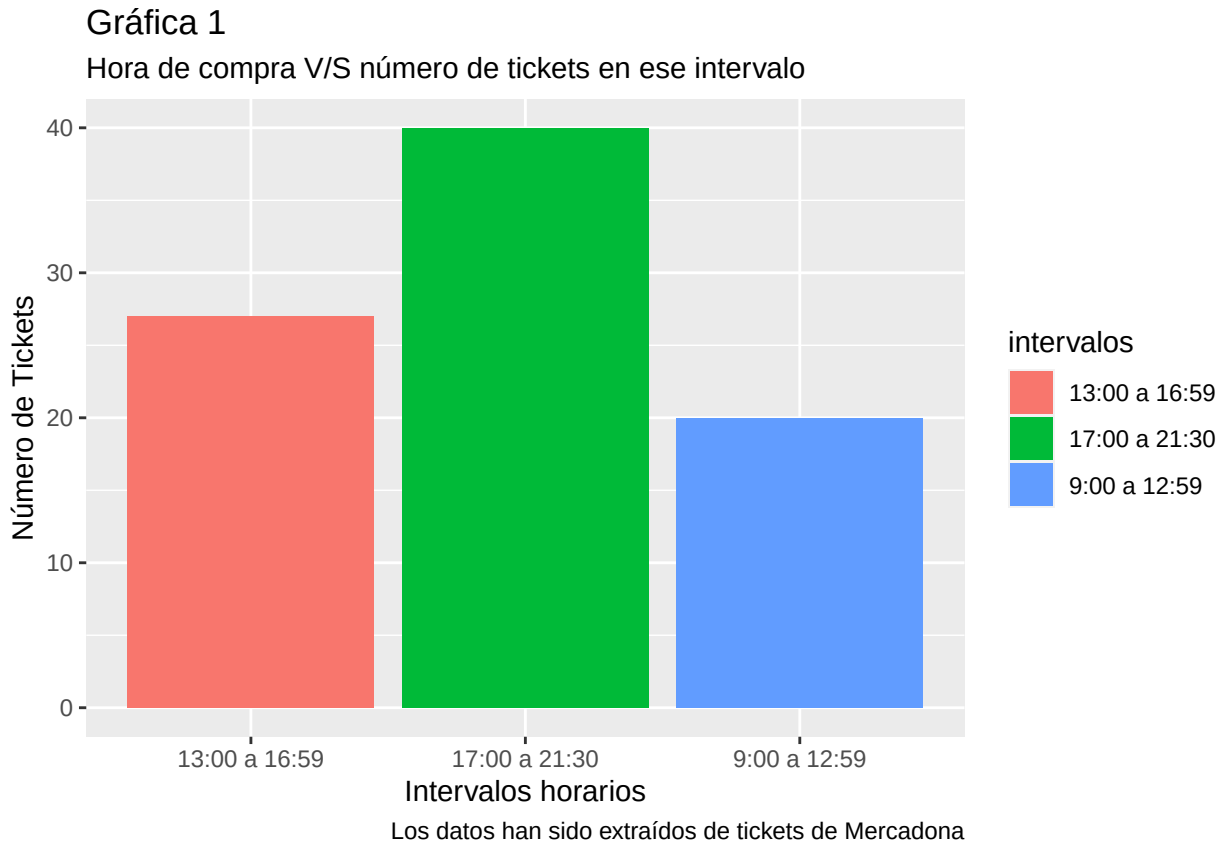
-¿Cuántos tickets están registrados en cada población?

Todas estas preguntas se intentarán responder en el presente documento, mostrando las respectivas representaciones gráficas con los tipos de datos correctos, ya que el tipo de gráfica empleada será correspondida con un tipo de variables u otras.

2.1 Análisis univariante:

Distinguiremos el análisis entre las variables de tipo numérico, las de tipo categórica y las tipo factor, que son una forma específica de representar variables categóricas. Este análisis es crucial, ya que, ciertos estadísticos descriptivos (como por ejemplo la media) carecen de sentido en las variables categóricas.

2.2 Análisis bivariante:

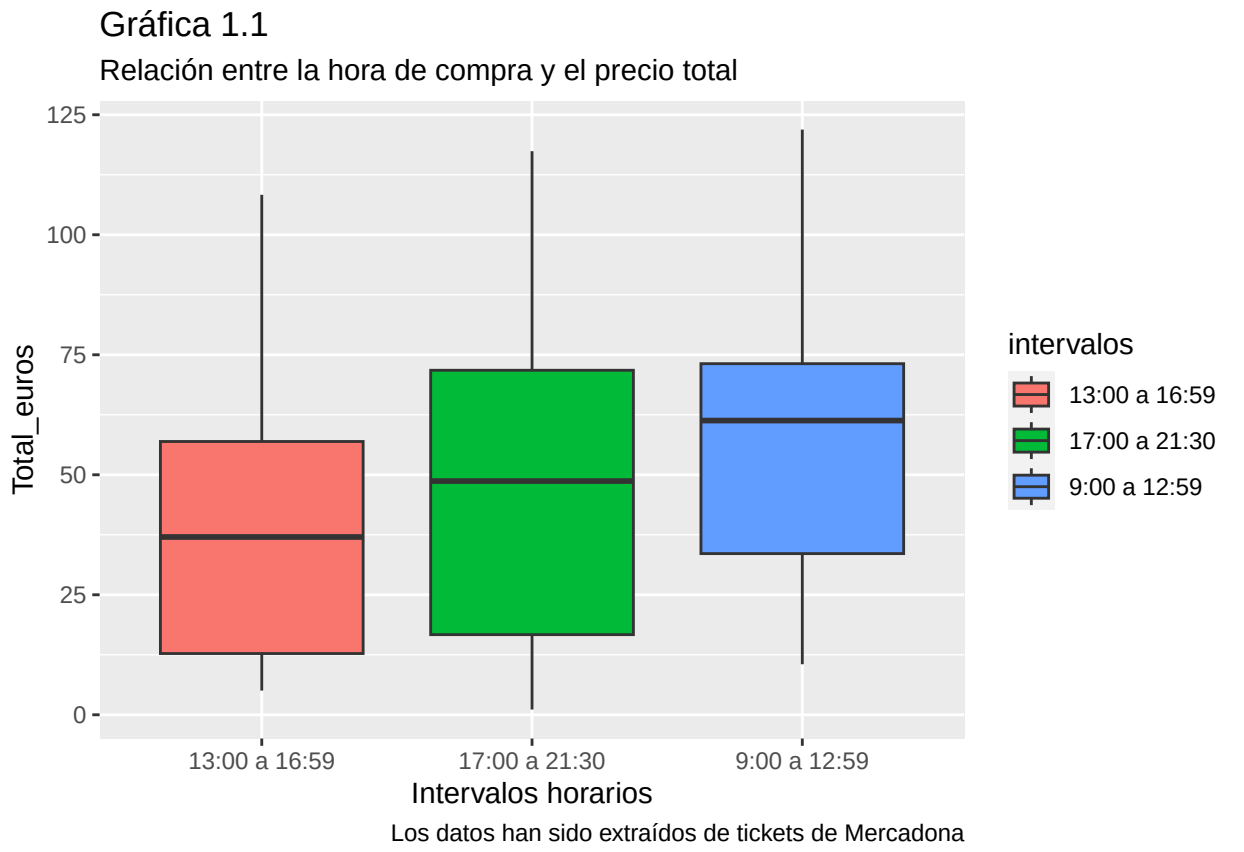



```
## Saving 6.5 x 4.5 in image
```

FIGURA 2: ¿Influye la hora de compra con el número de tickets en esos intervalos de tiempo?

Analizamos esta pregunta mediante una gráfica de barras. En primer lugar hemos creado esta gráfica con 'geom_bar' debido a que se trata de variables categóricas, ya que se trata de un intervalo de tiempo y no se puede medir directamente en términos numéricos continuos. Para crearla hemos separado los tickets según el intervalo de horas en el que se realizó la compra. Debido a la falta de tickets de 21:00 a 21:30 la

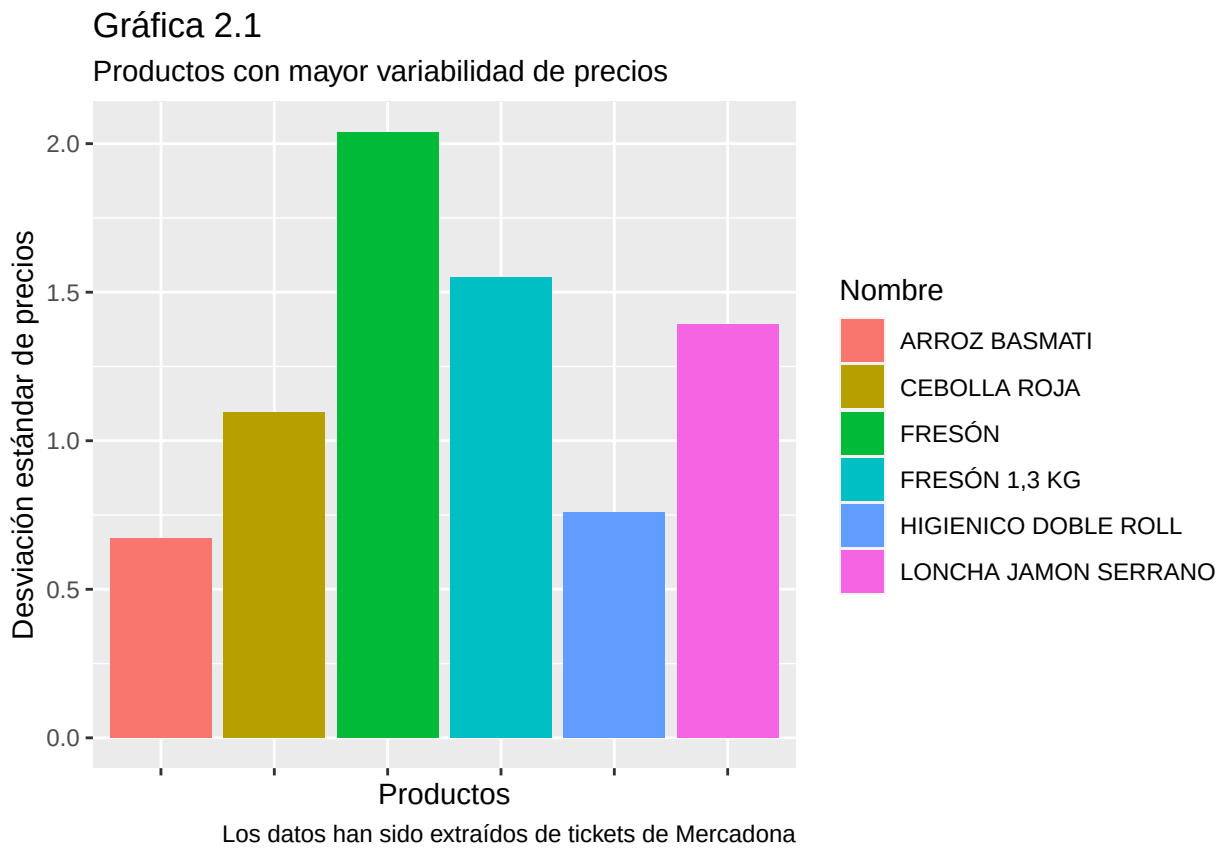
gráfica contenía un intervalo insignificante, por lo que lo unimos con el anterior intervalo. En nuestra gráfica podemos observar que, con diferencia, la hora en la que más se compra es por la tarde de 17:00 a 21:30, esto puede deberse a que la gran mayoría de gente trabaja por la mañana además de que lo niños van a clase, por ello, es más común que vayan a comprar por la tarde. Vemos también que el intervalo con menos conumo es el de 9:00 a 13:00, lo cual puede deberse al mismo motivo. Además, los dos intervalos más comunes también se encuentran cerca de las comidas más abundantes del día como son la comida y la cena.




```
## Saving 6.5 x 4.5 in image
```

FIGURA 3: ¿Influye en el precio total la hora en la que se compra?

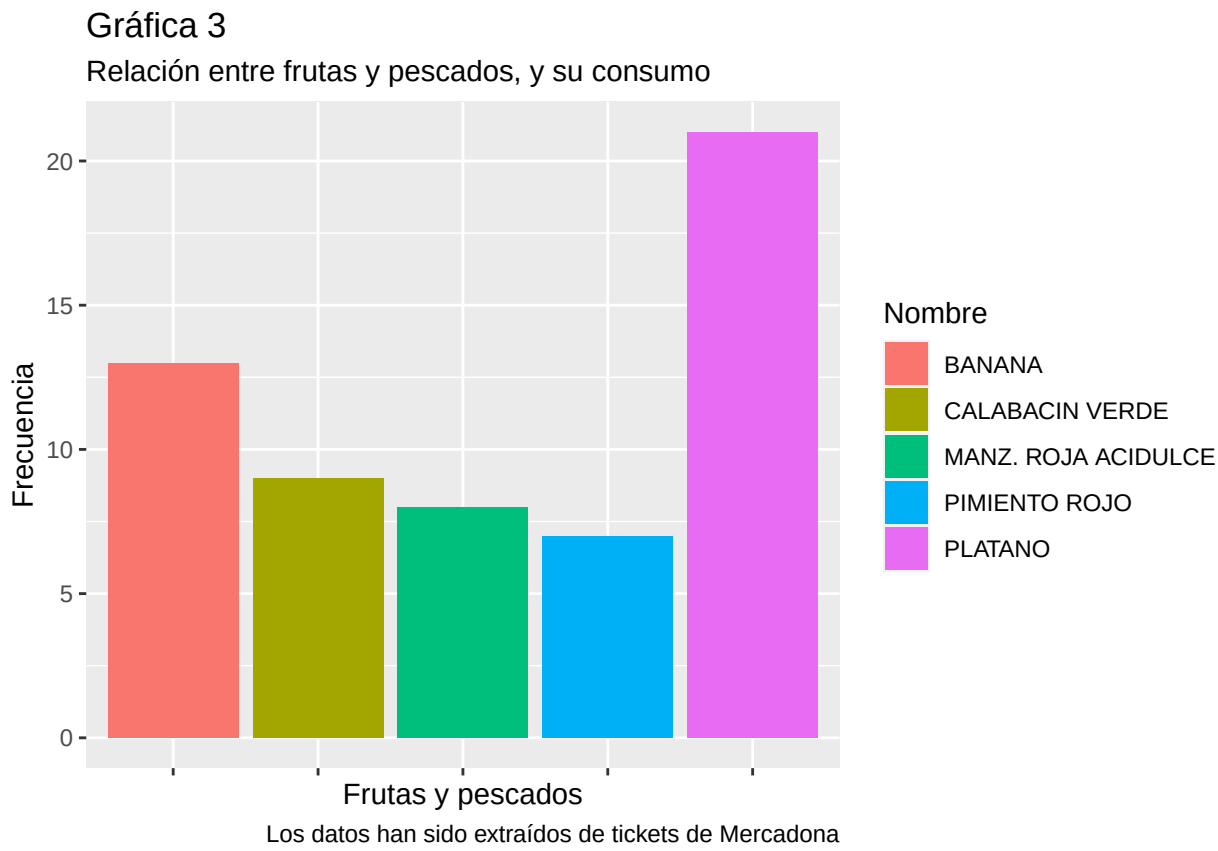
Al igual que en la figura anterior, se trata de variables categóricas por lo que para responder a esta pregunta utilizamos un boxplot. Para crear esta gráfica, analizamos el precio total respecto de la hora de compra de ese ticket. Podemos observar que el precio más elevado es a primera hora (bigotes diagrama de cajas), es decir, la hora en la que menos gente acude y el más bajo es a medio día. Por lo que observamos que la gente gasta más dinero cuando compra por la mañana y esto puede deberse a que compran para todo el día y al comprar por la tarde compran para ese momento. Además, podemos ver que cada intervalo tiene asociada una mediana: el primer intervalo de compra tiene una mediana de unos 65€ aproximadamente, el segundo intervalo de compra tiene una mediana sobre unos 35€ y, el último, tiene una mediana de 50€. Por último, destacar que el intervalo de 17:00 a 21:30 tiene una mayor variabilidad en el precio total de cada ticket.



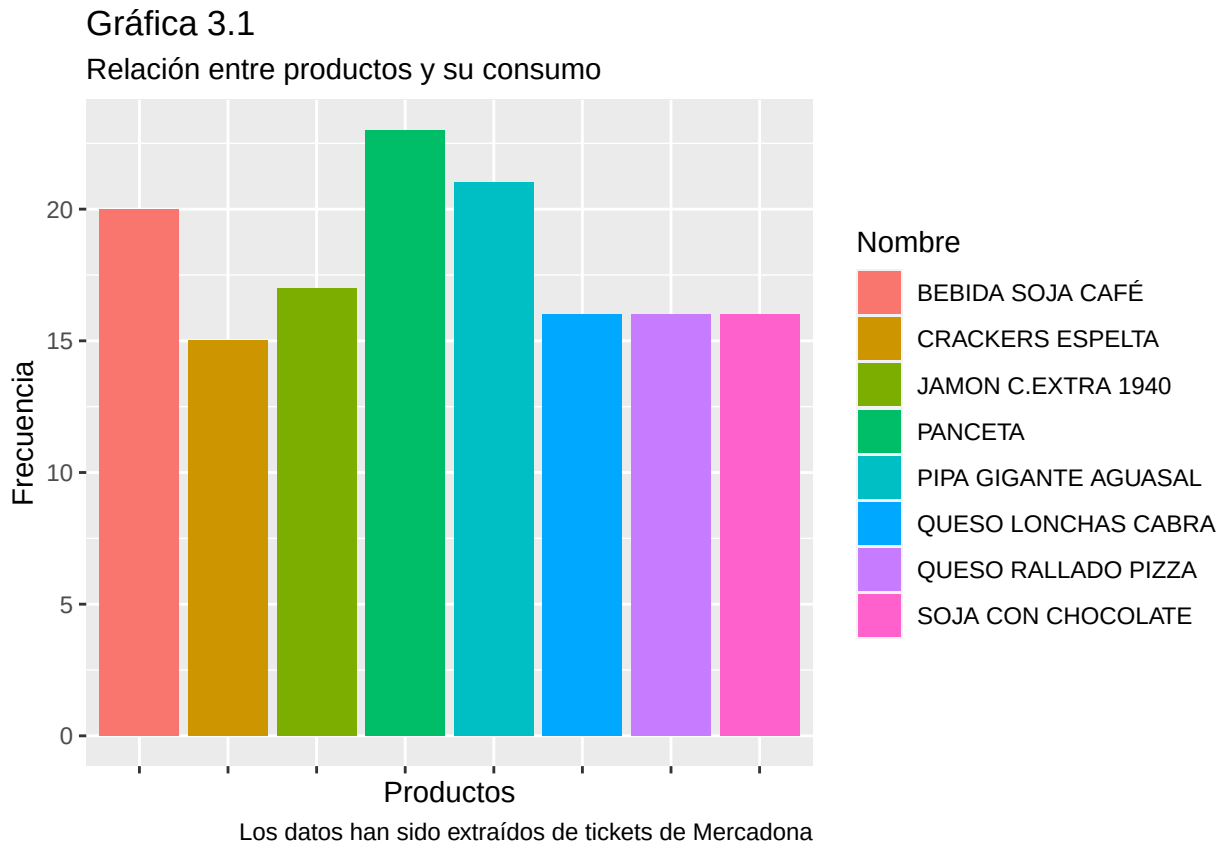

```
## Saving 6.5 x 4.5 in image
```

FIGURA 4: ¿Qué productos tienen la mayor variabilidad en los precios?

Para poder responder correctamente a esta pregunta, realizamos un gráfico de barras, debido a que las variables son de tipo categórico, ya que representan diferentes nombres de productos. Las variables categóricas son aquellas que toman valores de una categoría o grupo específico, en lugar de valores numéricos continuos. En este caso, los nombres de los productos (como "arroz basmati", "cebolla roja", etc.) son categorías distintas, y la gráfica muestra cómo varía la desviación estándar de precios para cada uno de estos productos. En esta gráfica mostramos los 6 productos en los que su precio es el más variable. El papel higiénico se encuentra entre éstos debido a que este está compuesto por celulosa, un componente que antes era de gran escasez, pero ahora se está haciendo más abundante, por lo que el precio de este está bajando. Por otro lado, también observamos que el fresón es otro de los elementos cuyo precio varía en abundancia, debido a que esta fruta es mucho más cara cuando no está en temporada, pero su precio disminuye en temporada. Por último, destacamos también el arroz. Últimamente su precio se ha visto incrementado debido a que la India, el mayor exportador de arroz a España ha decidido prohibir la venta de arroz al extranjero, por lo que debido a esta escasez, su precio ha aumentado.




```
## Saving 6.5 x 4.5 in image
```



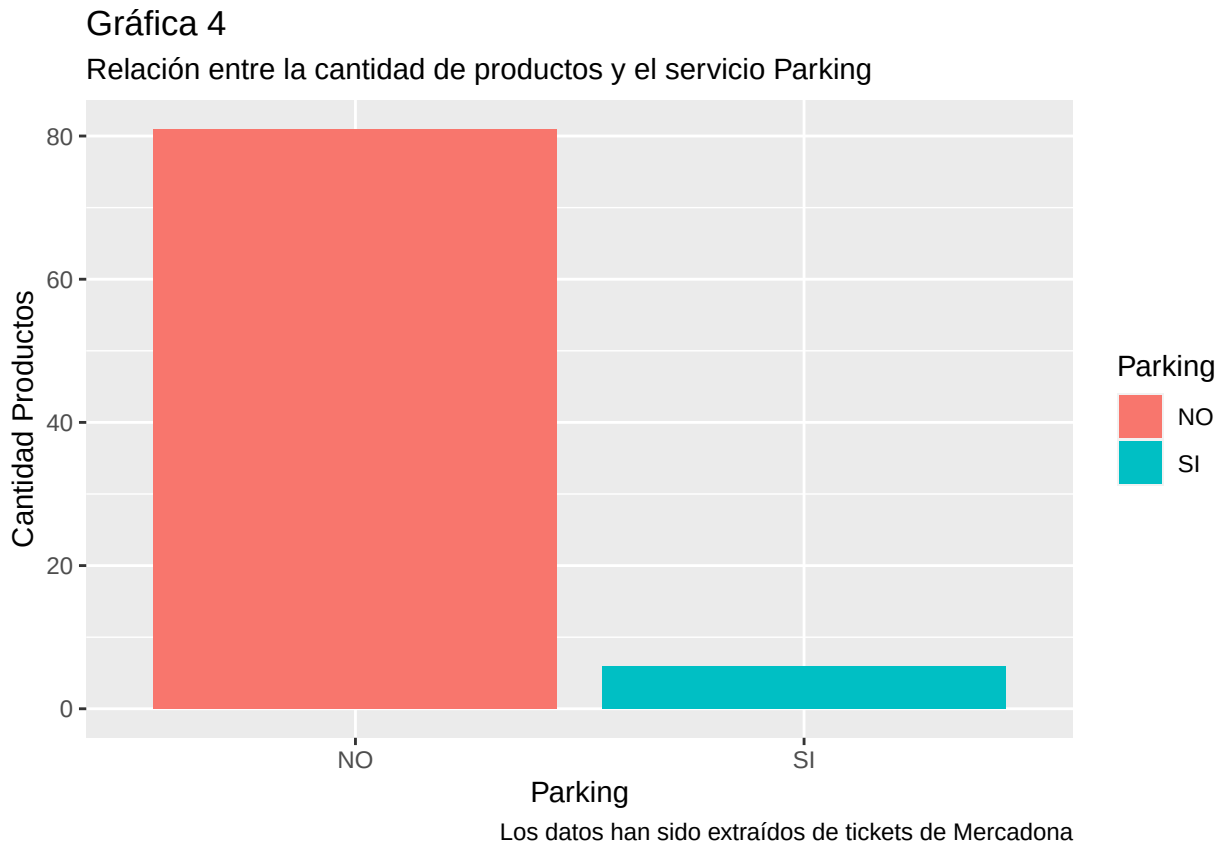

```
## Saving 6.5 x 4.5 in image
```

FIGURA 6: ¿Qué frutas y pescados han sido los más consumidospor los clientes?

Para responder correctamente a esta pregunta, creamos un gráfico de barras. Utilizamos un gráfico de barras porque las variables que estamos analizando son categóricas, es decir,representan diferentes nombres de productos. En lugar de valores numéricos continuos, como precios exactos, estamos interesados en cómo varía la desviación estándar de cada uno de estos productos. Por lo tanto, la gráfica nos muestra visualmente esa variabilidad. En concreto, en esta gráfica aparecen las frutas y verduras más compradas por los consumidores. Vemos que aparecen tanto el plátano como la banana y esto es debido a que es una fruta cuyo precio es muy asequible en el mercado, además, destacar que el plátano se consume más porque la fuente de dónde proviene es Canarias y, es requerido para múltiples recetas. Otra fruta muy consumida es la manzana. Esto se debe a que son muy baratas, versátiles culinariamente y prácticas para consumir a lo largo del día.

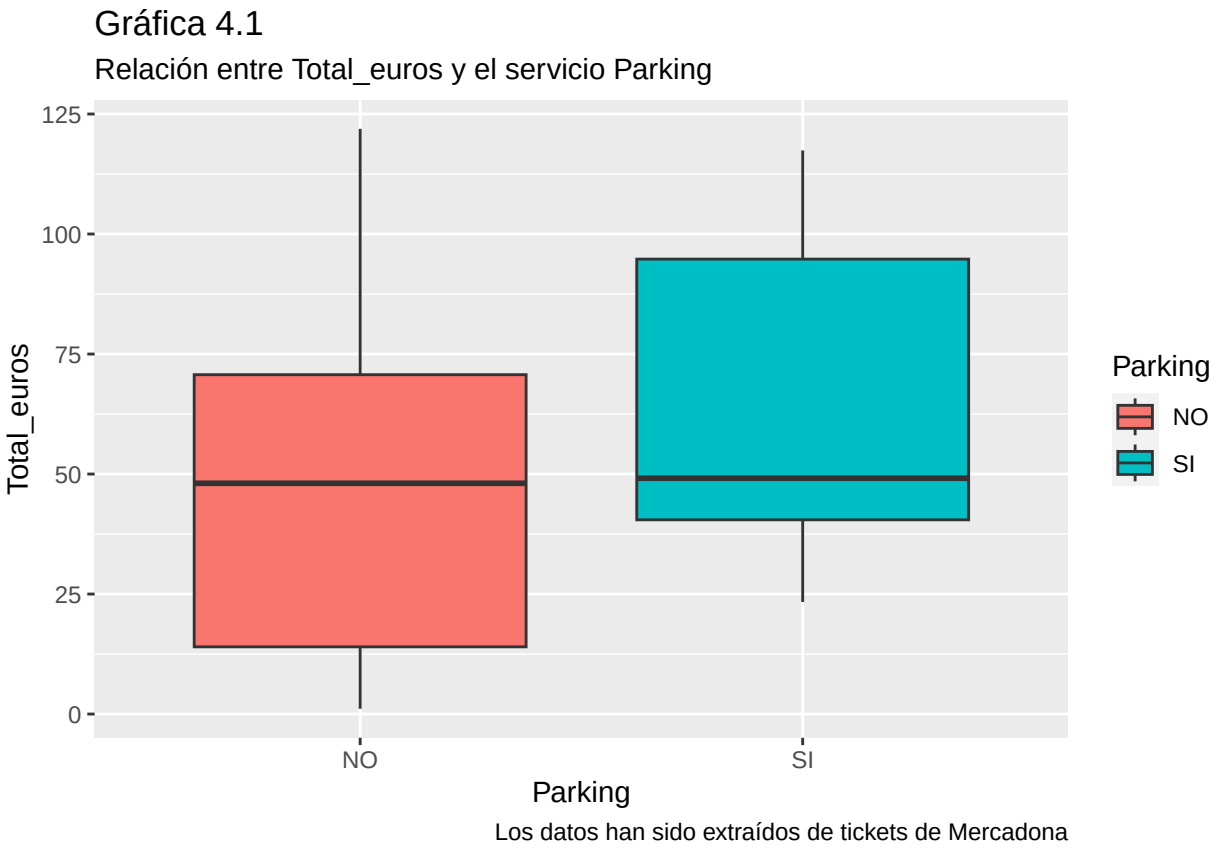
FIGURA 7: ¿Qué productos han sido los más vendidos?

Las variables categóricas son aquellas que representan diferentes categorías o grupos específicos, en lugar de valores numéricos continuos. En el contexto de la gráfica que estamos analizando, las variables son categóricas porque se refieren a nombres de productos. Interpretando la gráfica, podemos ver que en función de los tickets obtenidos, el producto más vendido es la panceta. Esto puede deberse a su gran versatilidad con distintos productos, ya que puede combinarse con gran cantidad de productos. Los demás se consumen más o menos con la misma frecuencia. En cuanto al jamón podemos deducir la misma conclusión que la panceta además de ser un plato típico que se basta de sí mismo para disgustarse. Por otro lado encontramos el queso. Un lácteo muy común en la mayoría de las neveras de nuestros hogares debido, de nuevo, a su fácil combinación con otros elementos. También encontramos productos como la soja con chocolate, las pipas o los crackers de espelta que son consumidos normalmente como snack hacen que cualquier consumidor quiera tener uno de estos productos en su despensa para consumirlos en cualquier momento.



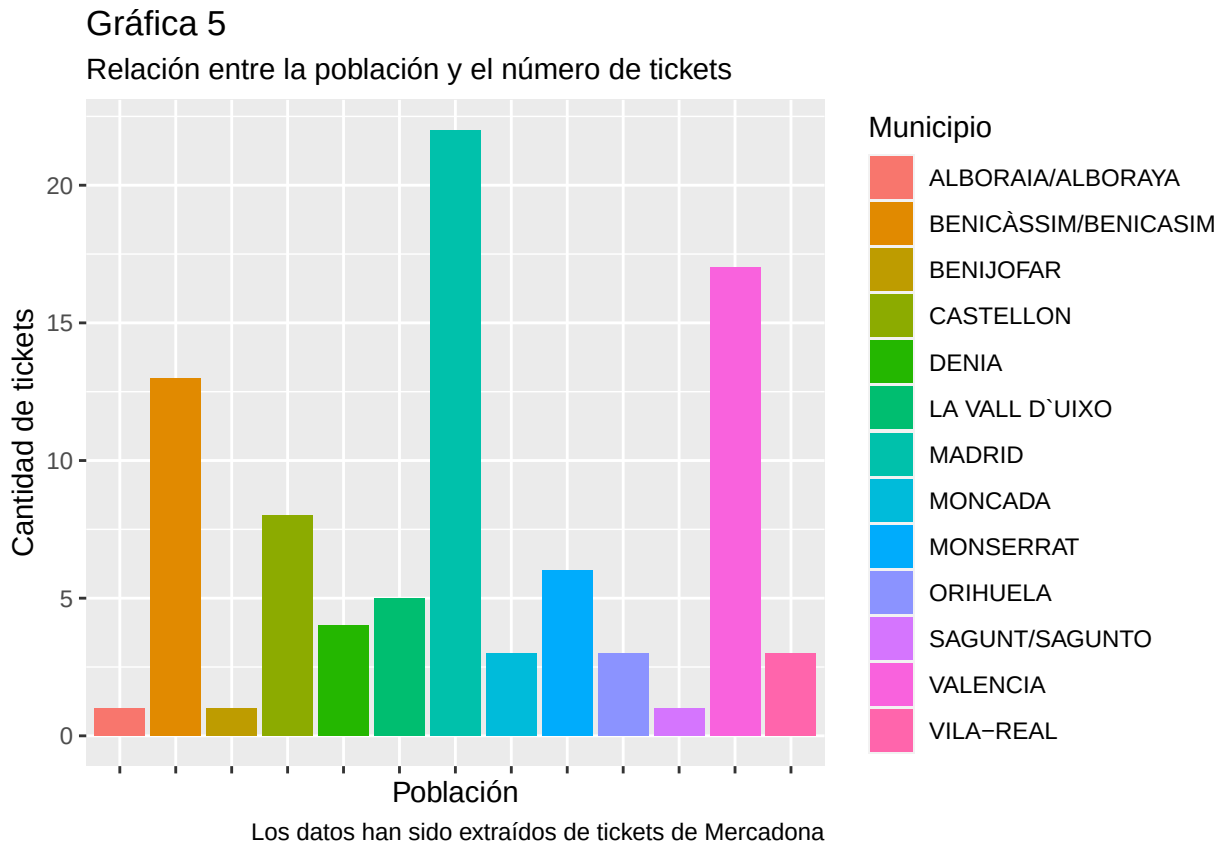

```
## Saving 6.5 x 4.5 in image
```

FIGURA 8: ¿Existe diferencia en el tamaño del ticket de compra entre aquellos clientes que utilizan el servicio de parking del supermercado y aquellos que no lo utilizan? Como podemos observar, nos encontramos frente a otro gráfico de barras, representado con un bar_plot y responde a nuestra pregunta con un rotundo no. Esto se explica gracias a que el tamaño de la compra no está relacionado con si el cliente ha cogido parking o no, ya que se pueden dar factores como que haya aparcado su vehículo relativamente cerca del supermercado, viva cerca del supermercado, etc.




```
## Saving 6.5 x 4.5 in image
```

FIGURA 9: ¿Existe diferencia entre el precio total de un ticket y el uso del servicio Parking? En este diagrama en caja, realizado con la función boxplot() (cargado en el paquete base de R), podemos observar como claramente el servicio parking está relacionado con el total en euros del ticket. Esto se puede dar gracias a que el cliente puede hacer una compra elevada en cuanto a productos por la facilidad de tener parking pero dado que en la gráfica de antes hemos visto que no necesariamente debe estar relacionado el tamaño del ticket con el uso del servicio parking podemos deducir que el parking también tiene un coste y como consecuencia ese coste se añade al ticket por lo que acabaría influyendo en el total de euros del coste. En cuanto al diagrama en caja en sí, podemos decir que cada caja representa la distribución de los gastos totales para cada grupo (Sí y No). La línea central de cada caja indica la mediana de los gastos, destacar que ambas cajas tienen la mimsa mediana (50€). El rango intercuartílico (la altura de la caja)muestra la variabilidad y dispersión de los datos alrededor de la mediana. Las líneas que se extienden desdela caja (bigotes) indican la variabilidad fuera del cuartil superior e inferior, pero dentro de un rango típico de datos.



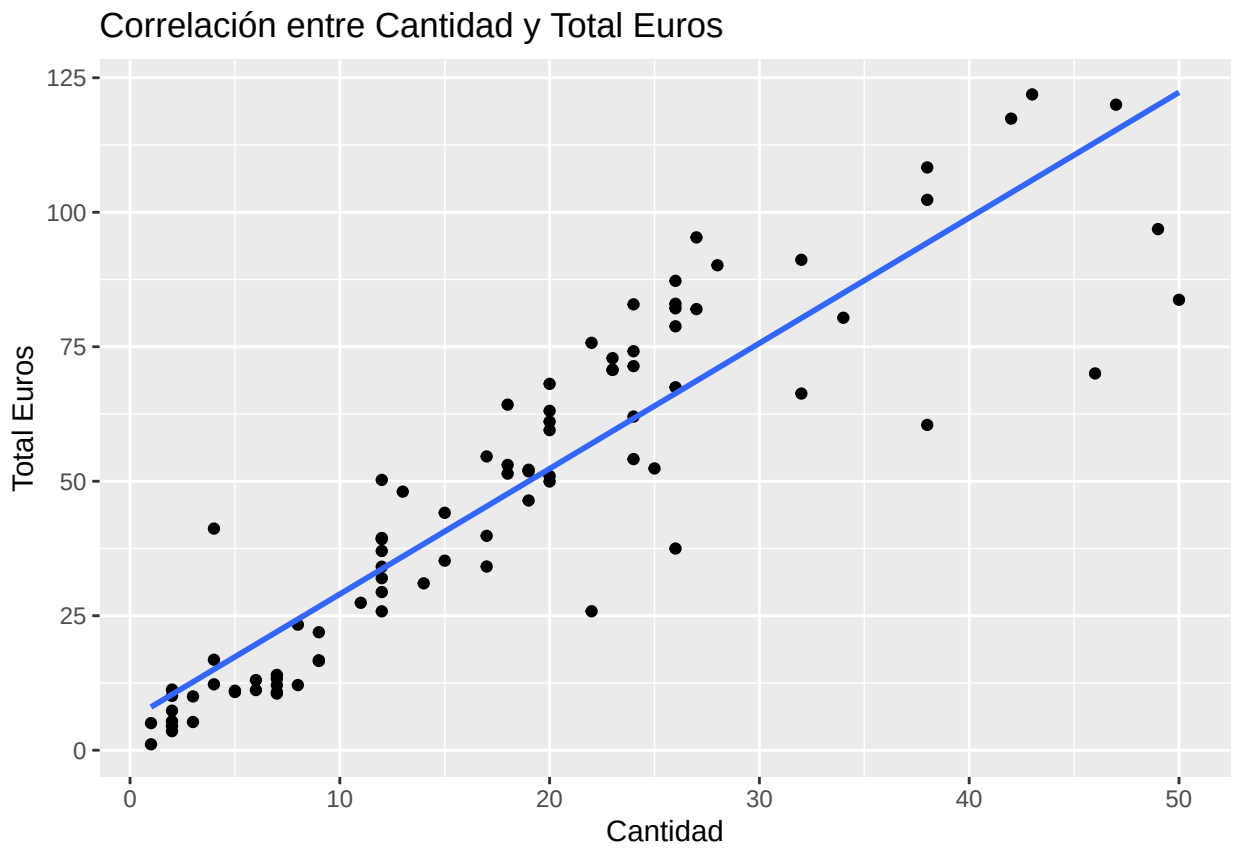

```
## Saving 6.5 x 4.5 in image
```

FIGURA 10: ¿Cuántos tickets están registrados en cada población?

En este diagrama de barras nos topamos con una gráfica realizada mediante la función geom_bar() ya que como hemos mencionado antes utilizamos este tipo de gráficos siempre y cuando se trate de variables categóricas. Interpretando el diagrama vemos que existe un número mayor de tickets en madrid por ejemplo, o valencia. Esto se puede deber a que son ciudades más grandes y con una población bastante más elevada respecto al resto de ubicaciones. También podríamos tener en cuenta que estas dos ciudades son dos de las ciudades con más estudiantes del país, lo que conlleva a una mayor compra en supermercado y obviamente, la obtención de ticket. Sabiendo esto, podemos deducir que los pueblos/ciudades más pequeños/pequeñas serán los que menos número de tickets tengan.

2.3 Gráfica Correlación:

```
## `summarise()` has grouped output by 'Factura_simpl'. You can override using the
## `.groups` argument.
## `geom_smooth()` using formula = 'y ~ x'
```




```
## Saving 6.5 x 4.5 in image
## `geom_smooth()` using formula = 'y ~ x'
```

FIGURA 11 La gráfica muestra un diagrama de dispersión con una línea de tendencia o recta de regresión, analizando la relación entre la cantidad de productos comprados y el total gastado en euros. Cada punto representa una observación individual que muestra la cantidad de productos comprados y el total gastado asociado a esa cantidad. La recta muestra la correlación entre las dos variables, sugiriendo una tendencia positiva donde a medida que aumenta la cantidad de productos, el total gastado en euros también aumenta. En cuanto a los puntos más 'alejados', podemos decir que se trata de outliers, es decir, valores atípicos o aberrantes. Estos pueden ser los clientes cuyo comportamiento de compra no se ciñe a la tendencia general, es decir, a la distribución del resto de clientes (recta).

3 CONCLUSIÓN:

En este momento del proyecto, haremos una breve conclusión para poder clarificar el objetivo de este trabajo. Una vez hemos respondido las preguntas planteadas sobre nuestro dataframe objetivo (df_tickets), ya sea el cómo se distribuye una variable numérica (Total_euros), una categórica (Productos) o una tipo factor(Parking), buscar diferentes relaciones entre nuestro entorno de variables, identificando por su tipo (numérico-numérico, categórico-numérico, categórico-categórico) es la etapa más importante de un correcto análisis. Más aún es el haber podido acceder a diversos métodos destacables y potentes como los métodos de detección de outlayers. Todo esto nos ha aportado un dominio del conocimiento de la distribución de cada ticket de nuestro dataframe sin necesidad de ser expertos en esta área de estudio. Además, en el transcurso del proyecto se ha aprendido a cómo aplicar el temario de la asignatura de Tratamiento de los Datos sobre un problema real en Ciencia de Datos. Por otro lado, también hemos aprendido a tomar decisiones a la hora de

interpretar un gráfico, tanto de manera individual como de forma conjunta, dotándonos así de competencias requeridas en nuestros trabajos del futuro.