

# **Tipología y ciclo de vida de los datos**

## **PRA1: ¿Cómo podemos capturar los datos de la web?**

**Miguel Ángel Quesada Fernández**

## Tabla de contenido

1. Contexto .....	3
2. Título .....	3
3. Descripción del dataset .....	3
4. Representación gráfica .....	4
5. Contenido .....	5
6. Propietario .....	6
7. Inspiración .....	7
8. Licencia .....	8
9. Código .....	8
9.1. Elementos clave .....	8
9.2. Estructura .....	9
9.3. Principales dificultades .....	10
10. Dataset .....	11
11. Video .....	11
12. Contribuciones .....	11

## 1. Contexto

En 2022, Mercadona, la mayor cadena de supermercados de España, obtuvo un beneficio de 718 millones de euros, un 5% más que el año anterior. La facturación de Mercadona aumentó un 11%, aunque la escalada de precios de la cesta de compra afectó a los márgenes de ganancias. El aumento en la facturación fue motivado en gran medida por el incremento de los precios, ya que el volumen de mercancía aumentó solo un 1%.

Juan Roig, presidente de Mercadona, ha reconocido que la inflación ha tenido un fuerte impacto en sus establecimientos, lo que ha llevado a la compañía a subir los precios un 10%, por debajo del 12% del incremento de los costes de Mercadona. Roig aseguró que, si el precio no se hubiese incrementado, la cadena de suministros se hubiese visto muy afectada. El aumento de los precios ha provocado que Mercadona actuara como un "dique de contención". (Fuente: [Eleconomista](#))

En este contexto de subida de precios, se plantea este proyecto que tiene como fin la obtención automatizada de los productos ofertados por Mercadona (descripciones y precios). El listado de productos se podrá obtener en cualquier momento, lo que permitiría también poder realizar estudios comparativos del precio de los productos a lo largo del tiempo y verificar el incremento o decremento de los precios realizado por Mercadona.

Para la realización del proyecto se ha optado por obtener los productos directamente desde la propia página web de Mercadona (<https://tienda.mercadona.es>), suponiendo una localización en el código postal 46001 (Valencia). La página web de la tienda de Mercadona es una web moderna y avanzada, que hace uso de técnicas para la carga de productos de forma asíncrona.

## 2. Título

El título elegido para el dataset es "Productos Mercadona", título que describe la esencia del contenido del dataset.

## 3. Descripción del dataset

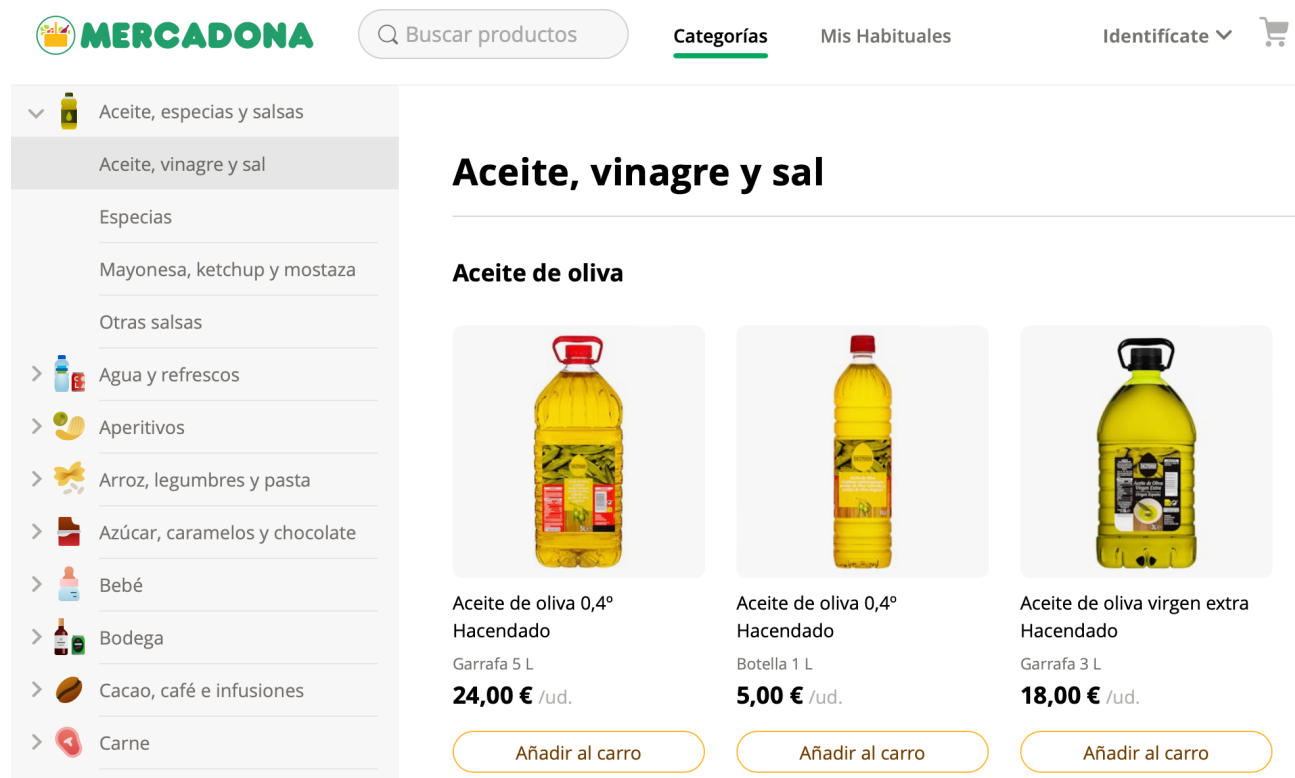
El dataset está constituido por 5420 registros, más un registro de cabecera. Cada registro representa un producto descrito por 10 variables.

El dataset contiene los productos disponibles el día 15 de abril de 2023, según el catálogo de Mercadona para el código postal 46001 (Valencia).

**El idioma del dataset es español.**

## 4. Representación gráfica

El Dataset contiene el conjunto de productos de la página web de Mercadona:



**MERCADONA** Buscar productos Categorías Mis Habituales Identifícate

**Aceite, vinagre y sal**

**Aceite de oliva**

Producto	Formato	Precio	Unidad
Aceite de oliva 0,4º Hacendado	Garrafa 5 L	24,00 €	/ud.
Aceite de oliva 0,4º Hacendado	Botella 1 L	5,00 €	/ud.
Aceite de oliva virgen extra Hacendado	Garrafa 3 L	18,00 €	/ud.

Por ejemplo, los tres productos anteriores se representarían en los siguientes registros:

categoria_principal	categoria_secundaria	subcategoria	descripcion	formato	tamano	precio	moneda	unidad_venta
Aceite, especias y salsas	Aceite, vinagre y sal	Aceite de oliva	Aceite de oliva 0,4º Hacendado	Garrafa	5 L	24,00	€	/ud.
Aceite, especias y salsas	Aceite, vinagre y sal	Aceite de oliva	Aceite de oliva 0,4º Hacendado	Botella	1 L	5,00	€	/ud.
Aceite, especias y salsas	Aceite, vinagre y sal	Aceite de oliva	Aceite de oliva virgen extra Hacendado	Garrafa	3 L	18,00	€	/ud.

A partir del dataset, podemos realizar distintos análisis, un análisis simple sería obtener las subcategorías con mayor número de productos, dónde destaca la subcategoría de “Laca de uñas” con un total de 55 productos.

# Categorías con mayor número de productos

```
df %>% group_by(subcategoria, .drop = FALSE) %>% count() %>% arrange(desc(n))
```

```
## # A tibble: 426 × 2
```

```
## # Groups:   subcategoria [426]
```

```
##      subcategoria          n
##      <fct>                <int>
##  1 Laca de uñas           55
##  2 Coloración color rubio  54
##  3 Marisco                53
##  4 Verdura                51
##  5 Snacks                 50
##  6 Pollo                  47
##  7 Cerveza botella y botellín 45
##  8 Frutos secos           45
##  9 Pescado                45
## 10 Perfume y colonia mujer 42
## # ... with 416 more rows
```

## 5. Contenido

Los elementos del dataset son los siguientes:

- **Categoría principal:** Variable categórica. Categoría principal a la que pertenece el producto. Existen un total de 26 posibles valores.
- **Categoría secundaria:** Variable categórica. Categoría secundaria, dentro de la categoría principal a la que pertenece el producto. Existe un total de 149 posibles valores.
- **Subcategoría:** Variable categórica. Subcategoría, dentro de la categoría secundaria y principal a la que pertenece el producto. Existe un total de 428 subcategorías.
- **Descripción:** Variable descriptiva. Es el producto ofertado. Existe un total de 4653 productos distintos, se deduce la existencia de varios registros para un mismo producto con distinto formato, tamaño y precio.
- **Formato:** Variable categórica. Representa el formato de presentación del producto. Existe un total de 281 formatos (lata, botella, paquete, tarro...)
- **Tamaño:** Tamaño del producto atendiendo a la naturaleza del producto puede estar expresado en litros, mililitros, gramos, kilogramos... incluye la unidad de medida.
- **Precio:** Precio del producto atendiendo al resto de categorías: formato, tamaño.
- **Moneda:** Siempre expresado en €
- **Unidad venta:** Si el precio del producto es por unidad, lote...
- **Fecha:** Fecha de extracción de la información

```
## 'data.frame':   5420 obs. of  10 variables:
##  $ categoria_principal : Factor w/ 26 levels "Aceite, especias y salsas",...: 1..
##  $ categoria_secundaria: Factor w/ 149 levels "Aceite, vinagre y sal",...: 1 1 ..
```

```
## $ subcategoria      : Factor w/ 426 levels "Absorbeolores y antihumedad",....
## $ descripcion      : chr  "Aceite de oliva 0,4º Hacendado" "Aceite de oli"..
## $ formato          : Factor w/ 280 levels "1 kg","1 kg aprox.",...: 264 258..
## $ tamano           : Factor w/ 823 levels " ","1 bandeja (500 g)",...: 574 ..
## $ precio           : num  24 5 18 6.05 5.15 24 5 16.2 5.55 2.4 ...
## $ moneda           : Factor w/ 1 level "€": 1 1 1 1 1 1 1 1 1 ...
## $ unidad_venta     : Factor w/ 4 levels "/150 g","/400 g",...: 4 4 4 4 4 4 ..
## $ fecha            : chr  "2023-04-15" "2023-04-15" "2023-04-15" "2023-04"..
```

El significado de los términos de categoría principal, categoría secundaria, y subcategoría se representa en la siguiente imagen. Cada categoría principal (azul) contiene una serie de categorías secundarias (naranja), y estas a su vez, contiene unas subcategorías (rojo).



**MERCADONA**  **Categorías** Mis Habituales Identifícate 

**Categoría principal** (Azul): Aceite, especias y salsas

**Categoría secundaria** (Naranja): Aceite, vinagre y sal

**Subcategoría** (Rojo): Aceite de oliva

Producto	Descripción	Unidad	Precio	Acción
	Aceite de oliva 0,4º Hacendado	Garrafa 5 L	24,00 € /ud.	<a href="#">Añadir al carro</a>
	Aceite de oliva 0,4º Hacendado	Botella 1 L	5,00 € /ud.	<a href="#">Añadir al carro</a>
	Aceite de oliva virgen extra Hacendado	Garrafa 3 L	18,00 € /ud.	<a href="#">Añadir al carro</a>

## 6. Propietario

El propietario de los datos es Mercadona, es quién proporciona los datos con fines operativos para la compra de productos a través de su página web.

Si bien, no he sido capaz de encontrar otros proyectos de web-scraping sobre el sitio web de Mercadona, sí que se ha encontrado algún trabajo que ha ayudado a comprender y encaminar el proyecto:

- **Nuez, Eduardo. Sistema de información para la recopilación y centralización de información sobre productos alimenticios.**

En su trabajo, Eduardo aborda un proyecto para implementar una aplicación donde los usuarios puedan obtener información nutricional de productos de alimentación con los precios de los supermercados españoles y además, proveer de un sistema de comercio donde los usuarios puedan crear tiendas y vender sus productos. En concreto, uno de los módulos que implementa es un módulo de web Scraping para la obtención de productos.

## 7. Inspiración

El conjunto de datos de productos de Mercadona puede resultar interesante por los siguientes motivos:

- **Análisis de precios:** Si se dispone de la información de los precios en una tendencia temporal, por ejemplo, se extraen los productos con periodicidad semanal, se podría realizar un análisis de la variación de los precios. Este análisis de precios se podría enfocar de varias formas:
  - Por un lado, descubrir las variaciones porcentuales de los productos, y descubrir patrones. ¿se producen subidas de precios en fechas próximas a navidad? ¿los productos de determinadas categorías sufren incrementos o decrementos en determinadas épocas del año?
  - Si se dispone de información de otros supermercados, se podrían realizar análisis para determinar si los precios de Mercadona están en línea con los precios de otros supermercados o si hay una diferencia significativa en los precios de los productos vendidos en diferentes categorías y formatos.
- **Análisis de oferta:** Al disponer de la información de la categoría y subcategoría de cada producto se pueden realizar análisis en relación con la variedad de productos. Por ejemplo, se puede determinar que categorías de productos tienen un mayor número de productos.

En definitiva, podríamos responder a las siguientes preguntas:

- ¿Cuántos productos se ofertan en una determinada categoría o subcategoría?
- ¿Qué categorías o subcategorías son las que tienen más productos?
- ¿Qué productos son los más económicos y los más caros dentro de cada categoría?

Si se dispusiera de información temporal, se podría responder a preguntas del siguiente tipo:

- ¿Cómo ha evolucionado con el tiempo el precio de un determinado producto?

- ¿Cómo ha evolucionado con el tiempo los productos ofertados en una categoría o subcategoría?

En definitiva, hay información valiosa que se puede extraer de los productos ofertados por Mercadona, además, esta información se puede enriquecer notablemente si se extrae la información correspondiente a distintos días para observar la evolución temporal.

## 8. Licencia

Para la publicación del dataset se ha considerado el uso de la licencia CC BY-NC-SA 4.0 License. (Creative Commons Attribution – Non Commercial – Share Alike 4.0)

Esta licencia permite a los usuarios compartir, adaptar y desarrollar a partir el trabajo original, siempre y cuando se de crédito al autor original, no se utilice con fines comerciales y se comparta bajo los mismos términos de licencia.

De esta forma, otras personas pueden beneficiarse del trabajo que se ha realizado, siempre y cuando se respeten los derechos de autor y se mantengan los mismos términos de licencia, además, se evitaría que alguien pueda utilizar el trabajo con fines comerciales. En definitiva, se sigue promoviendo la creatividad, el intercambio de conocimiento y se reconoce la autoría y el esfuerzo realizado.

## 9. Código

Enlace GitHub: <https://github.com/maquesadaf/scraping-mercadona>

### 9.1. Elementos clave

Los elementos clave del proyecto son los siguientes:

- **Se ha utilizado una tecnología avanzada como Selenium** para la navegación y el **descubrimiento de enlaces y navegación autónoma**. A través del uso de Selenium se ha realizado una navegación basada en el menú de categorías del sitio web. Esta navegación es robusta e independiente del número de categorías principales, secundarias y subcategorías. Es decir, si en un futuro se añaden nueva categorías, subcategorías o productos, el script seguirá funcionando correctamente.
- No se ha visto necesario implementar mecanismos de espera entre las distintas llamadas de forma explícita, como se ha observa en el video, **la espera introducida a través de Selenium necesaria para la carga dinámica de los productos se observa suficiente para evitar saturar el servidor**.



- Se comprueba al inicio de cada ejecución que el User-Agent que se está utilizando es *User-agent: Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_15\_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/112.0.0.0 Safari/537.36*.
- No ha sido necesario la utilización de APIs, toda la información ha sido obtenida en base al descubrimiento de enlaces y navegación autónoma.
- En relación con las **librerías y versiones utilizadas**, el listado completo sería:
 

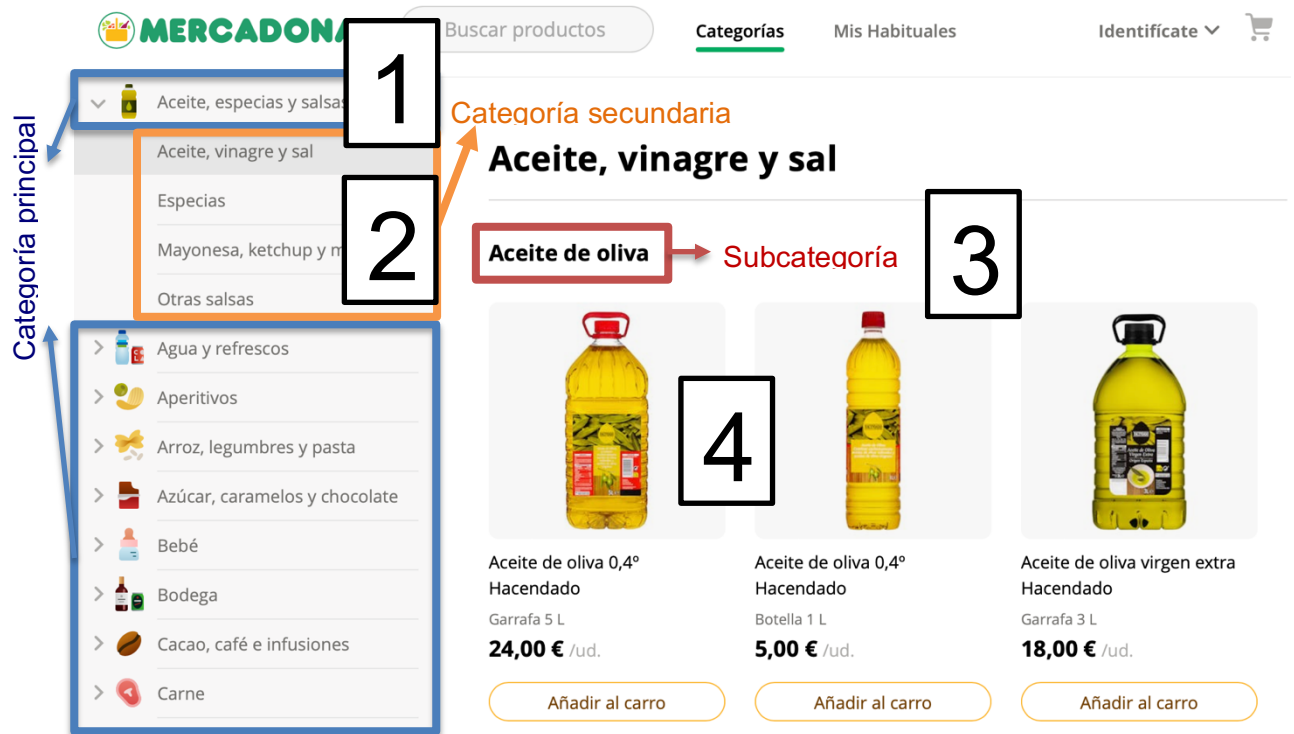
```
selenium~=4.8.3
pandas~=2.0.0
```

## 9.2. Estructura

En referencia a la estructura del código y la funcionalidad, a continuación se describe de forma resumida las acciones que se realizan en el módulo principal del proyecto **ScraperMercadona.py**:

- Se establecen las opciones de navegación, se inicia de forma maximizada el navegador y se deshabilitan las extensiones para que no se produzca ningún conflicto.
- Se inicia la navegación en la web de Mercadona (<https://tienda.mercadona.es/categories/>)
- Se realiza una serie de acciones iniciales descritas en el módulo **ScraperAccionesIniciales.py**, estas acciones son las siguientes:
  - Se espera a que aparezca el pop-up de cookies y aceptamos.
  - Se espera que aparezca el pop-up para introducir el código postal, se introduce el código postal de Valencia (46001) para poder continuar con la navegación.
- Se obtiene el listado de categorías principales del menú. Para cada categoría principal que se ha obtenido, se procede a procesar las categorías secundarias, para ello se ejecuta el módulo **ScraperCategoría.py**, que realiza las siguientes acciones:
  - Se hace clic en la categoría principal
  - Se obtienen el listado de categorías secundarias de la categoría principal.
  - Para cada categoría secundaria obtenida, se hace clic sobre la misma, y se espera a que se haya cargado el contenido.
  - Una vez cargado el contenido, se procesa las subcategorías de la categoría secundaria en el módulo **ScraperSubcategoría.py**:
    - Se obtiene el listado de subcategorías.

- Para cada subcategoría se obtienen los productos, se genera el correspondiente producto (**Producto.py**) y se añade a una lista de productos.
  - Cada vez que se procesa una categoría principal se guarda una versión del Dataframe con el conjunto de productos obtenidos hasta el momento.
- El flujo puede resumirse en el siguiente gráfico:



1. Se recorren las categorías principales
2. Para cada categoría principal, se recorren las categorías secundarias
3. Para cada categoría secundaria, se recorren las subcategorías
4. Finalmente, para cada subcategoría, se obtienen los productos.

## 9.3. Principales dificultades

Las **principales dificultades** del proyecto han sido las siguientes:

- El proyecto presenta varios **pop-ups que hay que controlar**, así como introducir el código postal en uno de ellos. Teniendo esto en cuenta, y la navegación por la que se ha optado, se ha decidido hacer uso de Selenium.
- La **información es cargada de forma asíncrona y dinámica** (se utilizan ficheros en **formato XHR** que transfieren datos), por lo que en el momento cero es posible que no estén los productos cargados, esto se ha resuelto metiendo esperas de tiempo que introducen retardos hasta que se hayan cargado los elementos necesarios.

```
# Se espera a que se haya cargado el contenido
productos_categoria_secundaria = WebDriverWait(driver, 100) \
    .until(EC.presence_of_element_located((By.CSS_SELECTOR,
    'div.category-detail__content')))
```

## 10. Dataset

El dataset ha sido publicado en Zenodo:

<https://doi.org/10.5281/zenodo.7838540>

## 11. Video

[https://drive.google.com/file/d/1EejwLPnk3GdZu8EYH2s1NwuWcZsQnCh/view?usp=share\\_link](https://drive.google.com/file/d/1EejwLPnk3GdZu8EYH2s1NwuWcZsQnCh/view?usp=share_link)

## 12. Contribuciones

Contribuciones	Firma
Investigación previa	Miguel Ángel Quesada Fernández
Redacción de las respuestas	Miguel Ángel Quesada Fernández
Desarrollo del código	Miguel Ángel Quesada Fernández
Participación en el vídeo	Miguel Ángel Quesada Fernández