COMPARACIÓN DE LOS PRECIOS EN LOS SUPERMERCADOS ESPAÑOLES



Mercadona, Caprabo y Bonpreu







Lucia Blanc Velázquez Sergio Sanchez Romero

Noviembre 2023

Tipología de datos Máster en Ciencia de Datos



1. Contexto

En esta actividad se ha llevado a cabo la recopilación de datos con el propósito de comparar los precios de los productos en tres supermercados diferentes: Mercadona, Caprabo y Bonpreu.

Sitios web (enlaces):

Mercadona: https://tienda.mercadona.es/Caprabo: https://www.capraboacasa.com/es

- Bonpreu: https://www.compraonline.bonpreuesclat.cat/products

Esta iniciativa se ha desarrollado en respuesta a la notable subida de precios de diferentes productos, como puede ser el aceite de oliva, un hecho que ha impactado en la economía de los consumidores, y por consiguiente ha generado un creciente interés de identificar las opciones de compra más económicas.

2. Título: Comparación de precios en los supermercados españoles.

El título del dataset "products.csv" incluye la fecha de captura de los datos, la palabra "products" la cual hace referencia a la obtención de cada uno de los productos que encontramos en los supermercados de estudio. Es importante remarcar la importancia de documentar la extracción de los datos, debido al hecho que los precios de los productos de los supermercados pueden ir cambiando. La solución del conjunto de datos es tabular, ya que, como se puede observar se usa la extensión *.CSV.

3. Descripción del dataset

El dataset "products.csv" presenta la información de todos los productos disponibles en las páginas webs del supermercado Mercadona, Caprabo y Bonpreu, donde se adjunta las características principales como categoría, subcategoría, producto e información del producto, permitiendo la comparación de precios de productos similares.



Figura 1. Distribución de los campos del conjunto de datos "products.csv" creado en el scraping.

4. Representación gráfica

Todos los supermercados incluidos en este trabajo tienen un apartado de tienda online, por lo que, para realizar el scraping correspondiente se ha seguido la estructura de cada una de las páginas web, siguiendo las categorías y subcategorías, y por mediante de la captura de las URLs correspondiente a cada producto. Una vez obtenemos la información completa de los productos se importa a un aplicación "flask" que funciona para buscar y ayudar a clasificar por precios los productos de cada supermercado.

En las siguientes imágenes se pueden ver dos fases el proceso seguido en formato gráfico para la obtención del csv final y la correspondiente aplicación creada.

<u>Fase 1:</u> Web scraping de los productos de los supermercados y creación de un csv conjunto ("products.csv").

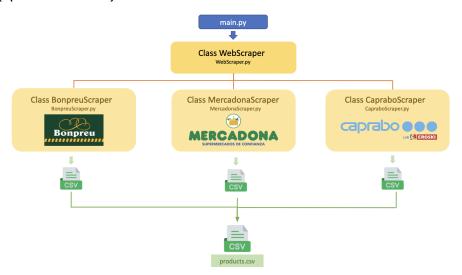


Figura 2. Procedimiento del webscraper del supermercado Mercadona, Caprabo y Bonpreu.

<u>Fase 2:</u> Importación del csv en una app buscador, donde se visualizan los productos ordenados según el precio.



Figura 3. Procedimiento de importación del csv generado a una app buscador.

5. Contenido

A continuación se presentan los campos del dataset generado en una tabla:

Campo		Descripción		
Supermarket	supermarket	Nombre del supermercado donde está disponible el producto	String	
Main Category	id_cat_principal	Identificador de la categoría principal del producto	Int	
	name_cat_principal	Nombre de la categoría principal del producto	String	
	url_cat_principal	URL donde se puede encontrar el la categoría principal	String	
Category	id_category	Identificador de la categoría del producto	Int	
	name_category	Nombre de la categoría del producto	String	
	url_category	URL donde se puede encontrar el la categoría	String	
Subcategory	id_subcategory	Identificador de la subcategoría del producto	Int	

	name_sub_category	Nombre de la subcategoría del producto	String	
	url_sub_category	URL donde se puede encontrar el la subcategoría	String	
Product	id_product	Identificador del producto	Int	
	name_product	Nombre del producto	String	
	url_product	URL donde se puede encontrar el producto online	String	
Product information	thumbnail	URL a una imagen en miniatura del producto	String	
	packaging	Tipo de embalaje utilizado para empaquetar el producto	String	
	unit_size	Tamaño o cantidad del producto	Numeric	
	unit_price	El precio del producto	Numeric	
	size_format	Formato o unidad de medida del producto ("ud": unidades)	String	
	bulk_price	Precio a granel por comprar o descuento por comprar al por mayor	Numeric	

Tabla 1. Definición de los campos de los que se compone el dataset

supermarket	id_cat_principal	name_cat_principal	url_cat_principa	id_category	name_category	url_category	id_subcategory	name_sub_category	url_sub_category	id_product	name_product	url_product	thumbnail	packaging	unit_size	unit_price	size_format	bulk_price
Bonpreu	c49d1ef2-bf51-44a	Alimentación	https://www.com	9e43d76d-91a	Azúcares y edulc	https://www.c	0198dc10-ba42-	Edulcorante comprimid	https://www.comp	20501	IFA Edulcorante	https://www.co	https://www	comprimidos	650	0.79	envase	0.00
Bonpreu	b2c9fc2f-ddb1-40	Cuidado personal	https://www.com	45fb23c6-1f2d	Cosmética facial	https://www.c	7e6ef471-5ccd-4	Limpieza facial	https://www.comp	2656	BONPREU Basto	https://www.co	mpraonline.b	caja	200	0.90	envase	0.00
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	8e881a6d-5e9	Complementos d	https://www.c	0438a070-f457-4	Menaje del hogar	https://www.comp	17946	BONPREU Cerill	https://www.co	mpraonline.b	lote	300	0.89	envase	0.00
Mercadona	23	Fitoterapia y parafarm	nacia	214	Parafarmacia	https://tienda.	686	Botiquín		79007	Bastoncillos de a	https://tienda.n	https://prod	Caja	200.0	0.90	ud	0.00
Mercadona	26	Limpieza y hogar		243	Menaje y conserv	https://tienda.	772	Cubiertos, vajilla y man	el	49319	Palillos redondos	https://tienda.n	https://prod	Bote	500.0	1.25	ud	0.00
Caprabo	2059806	Alimentación	https://www.capr	5000365	Productos de diel	https://www.c	5000380	Fitoterapia	https://www.capra	13436316	Cola de caballo	https://www.ca	https://www	caja	80.0	0,01	uds.	0,00
Caprabo	2059806	Alimentación	https://www.capr	5000365	Productos de diel	https://www.c	5000380	Fitoterapia	https://www.capra	18195040	L-carnitina en ar	https://www.ca	https://www	caja	10.0	0,01	uds	0,00
Caprabo	2060401	Higiene y belleza	https://www.capr	2060433	Afeitado y depilad	https://www.c	2060444	Depilación otros	https://www.capra	17110560	Bandas Cera De	https://www.ca	https://www	.capraboacas	a.com/ima	0,01		0,00
Caprabo	2060401	Higiene y belleza	https://www.capr	2060456	Cuidado corporal	https://www.c	2060459	Gel masculino	https://www.capra	15793110	Gel Apollo AXE,	https://www.ca	https://www	spray	0.4	0,01	1	0,00
Caprabo	2060401	Higiene y belleza	https://www.capr	2060523	Parafarmacia	https://www.c	5000128	Botiquín	https://www.capra	23130446	Bastoncillos ERO	https://www.ca	https://www	caja	200.0	0,90	uds	0,00
Bonpreu	c49d1ef2-bf51-44a	Alimentación	https://www.com	9e43d76d-91a	Azúcares y edulci	https://www.c	0198dc10-ba42-	Edulcorante comprimid	https://www.comp	27316	NATREEN Edulo	https://www.co	https://www	envase	400	3.95	envase	0.01
Bonpreu	c49d1ef2-bf51-44a	Alimentación	https://www.com	284d3af0-1cc	Panadería y bolle	https://www.c	4bbcc56c-2e99-4	Bollería y pastelería	https://www.comp	81392	PEIRON Surtido	https://www.co	https://www	bandeja	225	2.69	kg	0.01
Bonpreu	b2c9fc2f-ddb1-40	Cuidado personal	https://www.com	960c61c1-ffe6	Baño	https://www.c	771d9416-9380	Gel	https://www.comp	57607	TOPICREM Gel	https://www.co	mpraonline.b	botella	1000	9.95	L	0.01
Bonpreu	b2c9fc2f-ddb1-40	Cuidado personal	https://www.com	176ca415-f9di	Cabello	https://www.c	dcbbe427-b6d6-	Champús y tratamiento	https://www.comp	84070	H&S Champú Ci	https://www.co	https://www	botella	1000	9.39	L	0.01
Bonpreu	b2c9fc2f-ddb1-40	Cuidado personal	https://www.com	aadca00b-4d9	Resto del cuerpo	https://www.c	342af3d2-d110-4	Depilación y afeitado	https://www.comp	9198	LA TOJA Espum	https://www.co	https://www	bote	300	2.95	L	0.01
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	b628057c-f9at	Papel	https://www.c	fa77441a-a4ab-4	Servilletas de papel	https://www.comp	84459	BONPREU Servi	https://www.co	https://www	paquete	90	1.35	envase	0.01
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	b628057c-f9at	Papel	https://www.c	fa77441a-a4ab-4	Servilletas de papel	https://www.comp	68635	BONPREU Servi	https://www.co	https://www	paquete	200	1.30	envase	0.01
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	b628057c-f9at	Papel	https://www.c	fa77441a-a4ab-4	Servilletas de papel	https://www.comp	79671	SUAO Servilleta:	https://www.co	https://www	paquete	160	1.85	envase	0.01
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	b628057c-f9at	Papel	https://www.c	9e654165-7ffd-4	Pañuelos de papel	https://www.comp	7800	BONPREU Pañu	https://www.co	https://www	caja	150	1.05	envase	0.01
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	b628057c-f9at	Papel	https://www.c	9e654165-7ffd-4a	Pañuelos de papel	https://www.comp	44647	BONPREU Pañu	https://www.co	https://www	caja	100	1.30	envase	0.01
Bonpreu	5b850132-84f1-4f4	Limpieza y hogar	https://www.com	8e881a6d-5e9	Complementos d	https://www.c	0438a070-f457-4	Menaje del hogar	https://www.comp	55144	PAPSTAR Remo	https://www.co	https://www	bolsa	100	1.15	envase	0.01

Figura 4. Vista previa del dataset "products.csv".

6. Propietario

6.1. Propietario de los datos

- Mercadona: El propietario del sitio web es MERCADONA, S.A., con domicilio en Tavernes Blanques (Valencia), C/ Valencia nº 5. Su NIF es A-46.103.834.
- Caprabo: El propietario del sitio web es CAPRABO, S.A. Av. Estany de la Messeguera, 40-44 El Prat de Llobregat (Barcelona) CIF: A08115032.
- Bonpreu: El propietario del sitio web es Bon Preu, SAU, con domicilio en Les Masies de Voltregà, Ctra. C-17, Km 73 y CIF A-08665838.

6.2. Acceso ético

Durante el proceso de recopilación de los datos , se han seguido los principios éticos y legales, así como las directrices establecidas en los archivos robots.txt -presentados en el siguiente punto- de cada supermercado con el fin de evitar acceso a aquellas áreas no autorizadas de las páginas web.

El acceso se limitó a las secciones públicas relacionadas con los productos y los precios correspondientes.

6.3. Robots.txt

En relación a protecciones detectades, a continuación se específica el contenido del archivo robots.txt:

 Mercadona: El archivo indica que se permite el acceso a la página raíz (/), al archivo sitemap (/sitemap.xml) y a la sección de productos (/product), pero no a otras partes del sitio web. Es decir, se permite el raspado web a las áreas públicas relacionadas con los productos y los precios.

User-agent: *

Allow: /\$

Allow: /sitemap.xml Allow: /product Disallow: / Disallow: /api

Disallow: /legal

• Caprabo: Según lo que aparece en el archivo robots está prohibido el acceso a páginas con el siguiente URL "*/search/results/?q=" y "*/?zipCode" con el fin de evitar acceso a resultados de búsqueda específicos ni a páginas relacionadas con códigos postales o ubicaciones.

User-agent: *

Disallow: */search/results/?q=

Disallow: */?zipCode

 Bonpreu: El archivo de robots prohíbe que los agentes de usuario accedan a la URL o al directorio /sso-login del sitio web, es decir, el rastreador web o el bot de búsqueda no deben acceder ni indexar páginas o contenido bajo el directorio especificado en el sitio web.

User-agent: *
Disallow: /sso-login

6.4. Documentación Legal y Política de Privacidad

En cada uno de los sitios webs scrapeados y visitados, se indica un uso respetado de los documentos y se permite el acceso a toda aquella información que se encuentra disponible de cara al público.

Para más información específica, se permite visitar los documentos de Términos y Condiciones de Uso de cada uno de los sitios web:

- Mercadona: https://tienda.mercadona.es/legal/terms/es/

- Caprabo: https://www.caprabo.com/es/home/aviso-legal/

- Bonpreu:

https://www.compraonline.bonpreuesclat.cat/content/condicions-dus

7. Inspiración

El presente conjunto de datos se podría explotar y aprovechar de diversas maneras, tanto a nivel de la empresa como la de los clientes.

7.1 Nivel de la Empresa

Uno de los posibles usos de los datos de "products.csv" en la aplicación Flask a nivel de empresa, podría ser la de optimizar el inventario y la estrategia de compra, usando los datos de cada venta y las preferencias que tiene cada cliente para estudiar y analizar las tendencias de compra, la demanda de los productos y las fluctuaciones. De esta manera, se optimizaría el inventario, asegurando que los productos estén disponibles cuando los clientes los quieran sin producir excesos en el stock, al igual que se podrían identificar los productos populares y promoverlos en momentos estratégicos con el fin de aumentar las ventas de estos.

7.2 Nivel del Cliente

Siguiendo alguno de los ejemplos encontrados en otras páginas web de comparador de productos de supermercados, una de las maneras que consideramos más importante de aplicar sería el hecho de que los clientes pudieran usar estos datos

generados, mediante el uso de una aplicación móvil, para comparar la calidad y el precio de los productos entre varios supermercados, y así tomar mejores decisiones y ahorrar dinero. La aplicación podría permitir que el cliente tuviera la posibilidad de crear listas de compra personalizadas y estudiar donde saldría más económico y factible comprar los productos.

Los datos recopilados de los supermercados de estudio, podrían usar tanto por cada una de las empresas para mejorar la eficiencia y aumentar la fidelización de los clientes, como para los clientes para mejores tomar decisiones ahorrando tiempo y dinero, es decir, la explotación de estos datos conlleva beneficios significativos para ambas partes.

8. Licencia

La licencia seleccionada para el dataset resultante es: Licence Commons Attribution-Non-Commercial-ShareAlike 4.0 International (CC BY-NC-SA).

Esta licencia es la más adecuada para este dataset al combinar información de diferentes fuentes, y por lo tanto se necesita que respete los derechos de autor y los términos de uso de las páginas web de donde hemos obtenido la información.

Las características de esta licencia permiten que otros usuarios utilicen y compartan el conjunto de datos siempre y cuando den crédito al creador por la recopilación y creación de los datos, es decir, se hace un reconocimiento del trabajo y esfuerzo en la obtención. Por otro lado, es NonCommercial, es decir que no se pueden usar los datos con un objetivo comercial sin el permiso del creador. La licencia CC BY-NC-SA está aceptada en la comunidad de datos abiertos, permitiendo crear obras derivadas bajo la misma licencia, hecho que fomenta la colaboración y el enriquecimiento de los datos. Por último, el uso de esta licencia fomenta la transparencia en la recopilación y la comparación de los precios de los productos de los diferentes supermercados, lo que puede generar un beneficio directo a los consumidores y otros investigadores.

9. Código

El código fuente de extracción y el dataset creado son accesibles mediante la plataforma GitHub. Para acceder se puede visitar el siguiente enlace: https://github.com/ssanchezromer/supers

El programa usado para realizar la codificación de esta práctica ha sido PyCharm Community, donde se han usado scripts con el formato .py.

Los principales paquetes que se han usado durante toda la práctica han sido los siguientes:

- BeautifulSoap
- Selenium
- CSV
- Requests
- Time
- Pandas

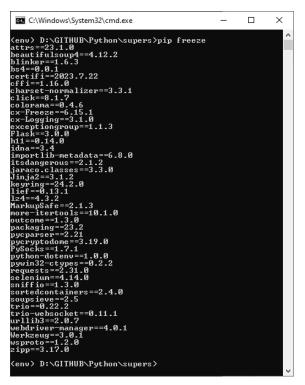


Figura 5. Visualización de paquetes y versiones de las librerías del entorno con pip freeze.

Los archivos .py creados para está práctica son:

- WebScraper.py: Proporciona métodos y funciones para realizar web scraping en las páginas de Mercadona, Caprabo y Bonpreu, y administrar los datos extraídos
- *MercadonaScraper.py:* Recopila datos de los productos del supermercado "Mercadona" a través de la API web.
- *CapraboScraper.py:* Recopilan datos de los productos del supermercado "Caprabo".
- BonpreuScraper.py: Recopilan datos de los productos del supermercado "Bonpreu".
- main.py y run.py: Con el fin de juntar todos los procesos de scraping de los tres supermercados, y cargarlo en la aplicación de Flask creada para realizar las búsquedas de productos pertinentes.

_

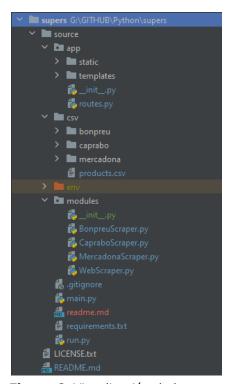


Figura 6. Visualización de la estructura de los ficheros y scripts en Pycharm.

10. Dataset

El dataset generado en está práctica se encuentra almacenado en el repositorio público Zenodo, accesible de manera pública sin restricciones.

El nombre y la descripción del conjunto de datos coinciden con las características especificadas anteriormente en el documento y por lo tanto, el nombre del fichero CSV es "products.csv".

El DOI del dataset publicado en Zenodo se encuentra en el siguiente enlace: https://zenodo.org/records/10086087

11. Video

El vídeo explicativo de la práctica se encuentra públicamente accesible en los siguientes enlaces:

- Repositorio público de github: https://github.com/ssanchezromer/supers,
- GoogleDrive de la UOC:
 https://drive.google.com/drive/folders/1r9utCm92l4t02Jzygo3exZg7ezeY ofz
 ?usp=share link

12. Aspectos de mejora

Como posibles mejoras en relación con el código y la práctica en general proponemos lo siguiente:

- 1. Mejorar el código para cargar toda la información de detalle de los productos, así como la obtención de cada una de las imágenes de los .
- 2. Mejorar la ejecución código del supermercado de Bonpreu, con el fin de reducir el tiempo de ejecución.
- 3. Posible implementación en el código de un batch de ejecución durante la noche, con el objetivo de actualizar los productos de los supermercados.
- 4. Posible implementación de un botón en la aplicación de Flask, mediante código JavaScript, el cual sirva para actualizar el precio de manera directa.
- 5. Limpieza del dataset, para unificar los valores de las variables, como por ejemplo las unidades del packaging y las unidades de medida.

Contribuciones

Contribuciones	Firma						
	Sergi Sánchez Romero	Lucia Blanc Velázquez					
Investigación previa							
Redacción de las respuestas							
Desarrollo del código							
Participación en el vídeo							

Referencias

- 1. Lawson, R. (2015). *Web Scraping with Python*. Packt Publishing Ltd. Chapter 2. Scraping the Data.
- 2. Mitchel, R. (2015). Web Scraping with Python: Collecting Data from the Modern Web. O'Reilly Media, Inc. Chapter 1. Your First Web Scraper.
- 3. Open Data Commons. https://creativecommons.org/licenses/by-nc-sa/4.0/
- 4. Soysuper. (s.f.). Aceite de oliva. https://soysuper.com/search?q=Aceite%20de%20oliva&p=2E26F150-7DA6-11EE-821B-B23648A4FFC4

ANEXO I. Headers User-Agent

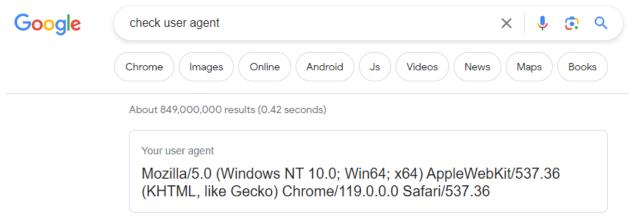


Figura A1. Cabecera User-Agent navegador.

```
C\Windows\System32\cmde \times + \times - \times \times \

\begin{array}{c} \ccolor \c
```

Figura A2. Cabecera User-Agent código scripts source.

Figura A3. Customización de Headers para la librería requests.