Dataset de precios de productos de Supermercado:

un conjunto de datos para el análisis de evolución de precios por estacionalidad y/o eventos

Nombres integrantes	Albert Gallego Jiménez David Roldán Puig
Web elegida	https://www.compraonline.bonpreuesclat.cat/
Repositorio Github	https://github.com/AlbertGallegoJimenez/Bonpreu-scraper
Dataset en Zenodo	https://doi.org/10.5281/zenodo.14036298
Carpeta Drive	https://drive.google.com/drive/u/4/folders/19JpqvSEcpfBSay44XaXJpWqWIow4QmoD
Video práctica	https://drive.google.com/file/d/1uRg6gWCUGmiks3O1khBawmw_bqcXqgG3/view?usp=drive_link

1. Contexto

La mayoría de las familias dependen de una planificación financiera cuidadosa para cubrir sus necesidades, y el éxito de esta planificación puede variar según la situación de cada hogar. Una de las partidas más importantes del presupuesto familiar es la dedicada a la alimentación. Sin embargo, este gasto no es siempre un coste fijo, ya que los precios de los productos esenciales pueden fluctuar a lo largo del año, afectando directamente el presupuesto familiar y cualquier planificación de éste. Estos cambios de precios están influenciados por factores como la estacionalidad, por ejemplo, en frutas y verduras, eventos macroeconómicos o por festividades como la Navidad que afectan a los precios por la alta demanda de ciertos productos.

El análisis que se propone cubre dos etapas: una descarga inicial de datos de precios de productos básicos de un supermercado mediante técnicas de *web scraping*; y un posterior análisis enfocado a la variabilidad de los precios en diferentes épocas del año. Cabe destacar que la segunda etapa de este análisis no se contempla en la realización de esta práctica y que solamente se propone como trabajo futuro. Este análisis permitirá identificar periodos de gasto elevado, lo cual ayudará a los consumidores a anticipar esos momentos, planificar mejor su presupuesto y tomar decisiones de compra informadas que optimicen sus recursos financieros.

La elección del sitio web para la recolección de datos es pertinente, ya que se trata de un supermercado real y reconocido cuya información publicada es precisa y actualizada, destinada a consumidores reales. Los supermercados mantienen sus precios al día para reflejar el coste de sus productos, lo cual asegura que los datos obtenidos para el análisis sean representativos. Además, al basarse en un sitio público y accesible, se garantiza que las fluctuaciones de mercado se vean reflejadas en los precios que el consumidor debe afrontar. Esto hace que el sitio web sea una fuente adecuada y confiable para el propósito de este estudio.

En este caso hemos seleccionado el Supermercado BonPreuEsclat, en su dirección de compra online.

2. Título

El título del dataset refleja el objetivo que se plantea para este estudio. En primer lugar, se realiza la recolección sistemática de precios de productos de supermercado en diferentes periodos del año, lo cual permite observar y documentar su evolución en el tiempo.

En segundo lugar, se analiza la variabilidad de los precios a lo largo del año, con el fin de determinar si factores como la estacionalidad, eventos puntuales o períodos festivos tienen un impacto significativo en el coste de ciertos productos. Así, se pretende explorar si existen patrones que indiquen aumentos o disminuciones de precios vinculados a la época del año o a eventos específicos.

Por ello, el título con el que se titula el dataset es: **“Dataset de precios de productos de Supermercado: un conjunto de datos para el análisis de evolución de precios por estacionalidad y/o eventos”.**

3. Descripción del dataset

El conjunto de datos contiene información detallada sobre todos los productos de supermercado. Se recogen datos de los productos organizados por categorías en la fecha de la captura de los datos. Para cada producto se recupera el nombre, precio, peso o cantidad, y el enlace directo al producto en el sitio web del supermercado. A cada uno de los productos se le añade como información el árbol de categorías y subcategorías que el supermercado añade, hasta cuatro niveles.

La organización del dataset en niveles de categorías y subcategorías facilita un análisis detallado por tipo de producto, mientras que la presencia de campos de cantidad y URL permite realizar un seguimiento preciso y verificado de los productos a lo largo del tiempo. Como la captura de los datos y por lo tanto del precio del producto se obtiene cada vez que se ejecuta el proceso, se decide añadir el campo fecha al dataset de salida.

Es un proceso que requiere ejecutarse diferentes veces durante un año para poder capturar la variabilidad del precio según las estaciones del año y las respectivas fiestas. Como se ha mencionado en el apartado anterior, en esta primera práctica el foco recae en la extracción de los datos, en la generación del dataset con los campos deseados y en la exportación a CSV dándole el nombre pertinente identificado por la categoría y fecha.

De esta manera, teniendo los ficheros separados por fecha y categoría, eventualmente se pueden combinar para realizar análisis por categorías independientes de productos. Para la entrega de la práctica, esta combinación de archivos divididos por categorías ya se ha llevado a cabo.

El análisis futuro que se propone con este dataset es clave para entender la evolución de los precios en productos de consumo cotidiano y su impacto potencial en la planificación financiera de las familias.

4. Representación gráfica.

En la Figura 1 se muestra el diagrama de flujo conteniendo los principales pasos a seguir para la obtención del dataset, el cual se describe en detalle en el siguiente apartado.

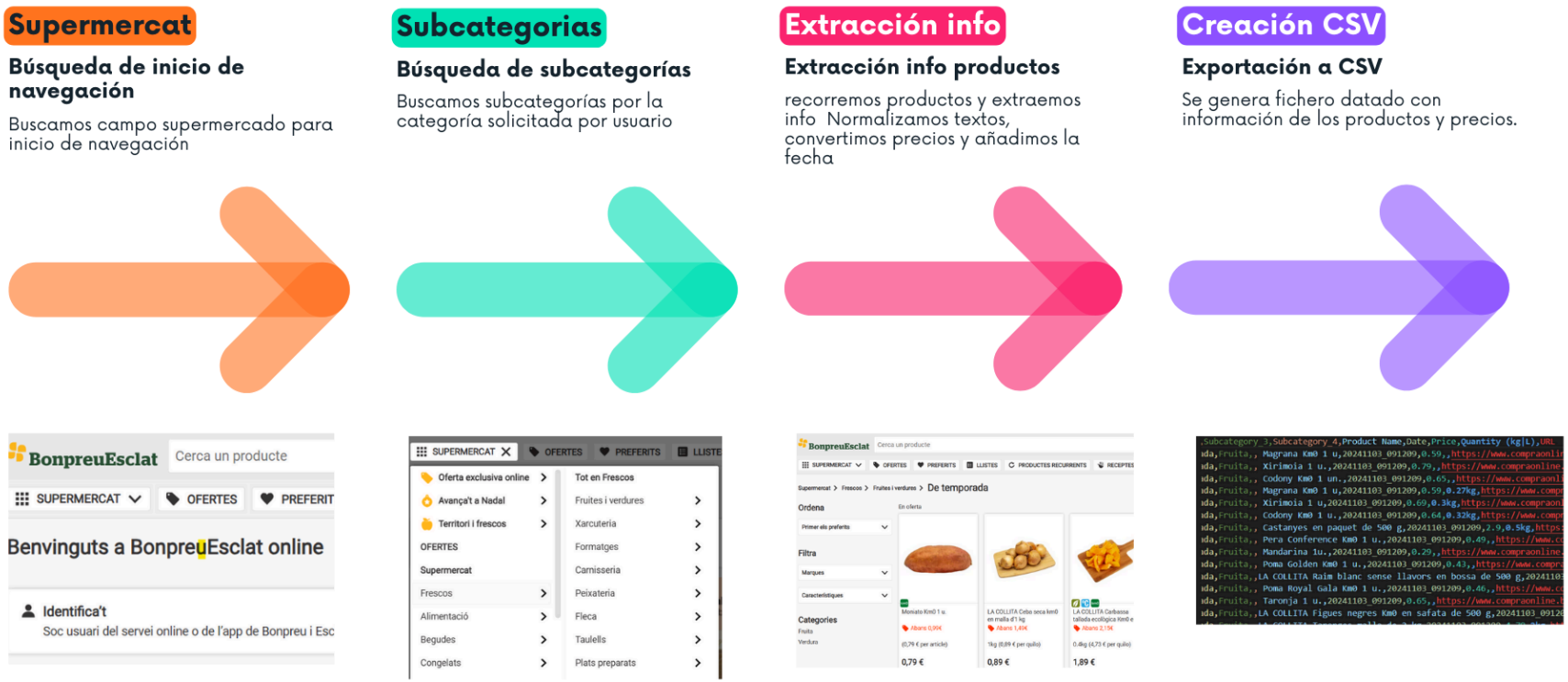

Figura 1. Diagrama de flujo para la obtención del dataset.

En la Figura 2 se muestra un ejemplo del CSV generado, los detalles de cada campo se abordan con más detalle en el apartado 5. Contenido.

```
Category,Subcategory_1,Subcategory_2,Subcategory_3,Subcategory_4,Product Name,Date,Price,Quantity (kg|L),URL
Frescos,Fruites i verdures,De temporada,Fruita,, Magrana Km0 1 u,20241103,0.59,,https://www.compraonline.bonpreuesclat.cat/products/magrana-km0-1-u/82289
Frescos,Fruites i verdures,De temporada,Fruita,, Xirimoia 1 u.,20241103,0.79,,https://www.compraonline.bonpreuesclat.cat/products/xirimoia-1-u/82272
Frescos,Fruites i verdures,De temporada,Fruita,, Codony Km0 1 un.,20241103,0.65,,https://www.compraonline.bonpreuesclat.cat/products/codony-km0-1-un/51566
Frescos,Fruites i verdures,De temporada,Fruita,, Magrana Km0 1 u,20241103,0.59,0.27kg,https://www.compraonline.bonpreuesclat.cat/products/magrana-km0-1-u/07390
Frescos,Fruites i verdures,De temporada,Fruita,, Xirimoia 1 u,20241103,0.69,0.3kg,https://www.compraonline.bonpreuesclat.cat/products/xirimoia-1-u/07456
```

Figura 2. Ejemplo de fichero CSV exportado.

5. Contenido

Para describir en detalle los contenidos del dataset exportado, se decide replicar parte de un CSV de prueba exportado (véase Tabla 1).

Tabla 1. Estructura del dataset exportado a fichero CSV.

Nombre Campo	**Descripción campo**	**Ejemplo**
Category	Categoría a la que pertenece el producto	Frescos
Subcategory_1	Nivel 1 de subcategoría	Fruites i verdures
Subcategory_2	Nivel 2 de subcategoría	De temporada
Subcategory_3	Nivel 3 de subcategoría	Fruita
Subcategory_4	Nivel 4 de subcategoría	–
Product name	Nombre de producto	Magrana Km0 1 u
Date	Fecha en día de captura del dato en formato YYmmdd	20241103
Price	Precio en euros	0.59
Quantity (Kg|L)	Cantidad en Kg o en Litros del envase del producto.	0.27 Kg
URL	URL informacional y específica del producto seleccionado.	https://www.compraonline.bonpreuesclat.cat/products/magrana-km0-1-u/07390

Como las búsquedas se pueden optimizar para buscar sólo por categoría seleccionada, el fichero CSV generado se guarda con un nombre formado por la combinación *category* + *timestamp*. Un ejemplo de nombre de fichero sería “Frescos_20240901_120000.csv”.

De este modo se pueden ir obteniendo diferentes ficheros a lo largo de un periodo de tiempo, evitando la sobrescritura del archivo resultante. El hecho de separar los ficheros por categoría presenta dos ventajas: por un lado, se contribuye a un proceso de scraping más responsable, evitando sobresaturar el servidor donde se aloja la web, buscando así únicamente la información necesaria; y por otro lado, se facilita un posterior procesado de los datos con la finalidad del análisis de evolución planteado. Adicionalmente, se habilita la opción de fusionar de forma automatizada diferentes datasets extraídos de varias categorías en un único fichero CSV.

6. Propietario

Para el proyecto de análisis de precios en supermercados, el propietario del conjunto de datos es el supermercado Bonpreu Esclat, que publica la información de sus productos y precios a través de su página web de comercio en línea. Por lo tanto, la información de productos, precios, categorías y disponibilidad es de acceso público, ya que está destinada al uso y la compra de productos por parte de sus clientes.

Este proyecto de scraping tiene el propósito de capturar los precios en diversas fechas del año para la evolución de precios en función de factores como la estacionalidad y eventos especiales (como las festividades), con el fin de proporcionar una herramienta informativa que pueda beneficiar a los consumidores en su planificación de gastos.

En referencia a estudios previos, existen análisis de precios en cadenas de supermercados, como los realizados por organismos como la OCU (Organización de Consumidores y Usuarios en España), que realizan comparaciones de precios entre diferentes establecimientos para identificar patrones de inflación y la variabilidad estacional[1]. El mismo organismo también pone a disposición una aplicación móvil[2]. Igualmente, también existen otras aplicaciones móviles de terceros con el mismo fin, como pueden ser Soysuper, Ofertia, FindItApp, Tiendeo, etc.

Estos estudios y desarrollos evidencian la alta demanda de la sociedad por este tipo de análisis, subrayando así la relevancia y la necesidad del análisis propuesto en este documento. El caso propuesto tiene como objetivo generar una visión detallada de la variabilidad de precios en un único supermercado, lo cual se alinea con el derecho a la información y la transparencia en el mercado, al permitir a los consumidores hacer elecciones informadas basadas en datos concretos.

Para un uso más ético de las consultas en la web, el sistema ofrece la posibilidad de filtrar hasta el nivel de subcategoría la consulta a realizar. En definitiva, permite realizar una búsqueda completa de todos los datos, o bien filtrada por categoría o por subcategoría dentro de su categoría. Cada una de ellas exporta la información en un fichero con el nombre específico de la consulta y la fecha[3]. Aunque la consulta se haga por la globalidad de todas las categorías, se acaba generando uno por cada una de ellas ya que se considera mucho más eficiente no solo por la manera de acceder a la información de la web sino el poder luego agrupar de manera más específica sin mezclar categorías.

Tal y como se comentará más adelante, se ofrece un script para fusionar todas las exportaciones de cada categoría en un solo fichero con la globalidad de los datos.

[1] Disponible en: https://www.ocu.org/consumo-familia/supermercados
[2] Disponible en: https://play.google.com/store/apps/details?id=org.ocu.market&pli=1
[3] En el fichero README se especifica la manera de ejecutar cada una de las respectivas consultas.

7. Inspiración

¿De dónde viene el concepto de producto de temporada? ¿Afecta solo a la calidad de los productos o también al precio para adquirirlos?

Definimos como productos de temporada aquellos que se producen naturalmente en una época específica del año, como frutas, verduras, y ciertos productos frescos, y que se encuentran en su punto de calidad, por sabor, madurez y cantidad de nutrientes que nos aportan por no haber sido conservados mucho tiempo. Estos productos también tienden a ser más económicos ya que la oferta es más alta y no se necesitan costes adicionales de almacenamiento o transporte. Esto implica que consumir productos de temporada suele ser más accesible y, en general, más económico para el consumidor.

Como ejemplos claros de productos de temporada hay: las naranjas, mandarinas, kiwis en invierno; las fresas en primavera; la sandía o el melón en verano; y las calabazas, castañas y boniatos en otoño. La disponibilidad natural de estos productos permite que su precio sea más estable y bajo cuando están en temporada, ya que la oferta es suficiente para satisfacer la demanda sin recurrir a importaciones o a métodos de cultivo en invernaderos, que encarecen el producto.

En definitiva, se genera un ciclo de precios predecible por el cliente tendiendo a la baja durante su temporada alta y subiendo en las épocas en que se vuelven menos accesibles.

Sin embargo, la estacionalidad no es el único factor que influye en los precios. Se pueden considerar desastres naturales, como sequías o inundaciones que reduzcan la cantidad de un producto. Si se mantiene la demanda del producto, la consecuencia directa es el aumento de su precio. Otro ejemplo es la alta demanda de ciertos productos durante festividades específicas provocando incrementos en sus precios debido a la relación directa entre demanda estacional y oferta limitada. Este fenómeno es común en eventos como las fiestas de fin de año, donde los consumidores suelen buscar productos tradicionales, y los proveedores pueden prever una mayor venta de ciertos artículos, incrementando los precios como respuesta.

El objetivo de este análisis es poder detectar estas variaciones durante un año viendo los productos que más sufren esta variabilidad y poder ajustarla a su ventana temporal.

Como se ha comentado en el apartado 6. Propietario, hay estudios que se enfocan en la variación de precios entre distintas cadenas de supermercados comparando una cesta de la compra establecida en un momento particular y que permite a los usuarios identificar la opción más económica en un mercado competitivo. Es decir, realizan una comparativa en un momento concreto, y establecen qué supermercado tiene la cesta de la compra establecida más económica.

Adicionalmente, el análisis que se propone en este trabajo no se centra en una comparación entre comercios, sino en ver cómo los precios de los productos de supermercado pueden fluctuar bajo diferentes coyunturas económicas o naturales, así como a la estacionalidad. Con este análisis se aspira a un mejor conocimiento en las situaciones que afectan a los precios y a una mejor comprensión de qué momentos del año se puede ahorrar en categorías específicas, cómo afecta el factor estacional a productos específicos, y en última instancia, poder anticiparse a incrementos en los precios.

8. Licencia

La licencia del código y dataset resultante de esta práctica es la **GNU GENERAL PUBLIC LICENSE** (GPL) para ser coherente con los objetivos de una práctica de código abierto de ciencia de datos.

En primer lugar, la GPL fomenta el acceso abierto y la transparencia. Esta licencia asegura que tanto el código como el dataset resultante puedan ser reutilizados, modificados y compartidos libremente, siempre que cualquier trabajo derivado mantenga la misma licencia. Esto significa que otros estudiantes pueden beneficiarse del código y del análisis de precios de este proyecto, realizando mejoras, adaptándolo a nuevas necesidades, o ampliándolo con otros datos o variables de su interés.

Otro punto importante es que la GPL respalda la colaboración. En este ejercicio, donde el objetivo es comprender la evolución de los precios, el permitir que otros usen el código y los datos contribuye a la mejora de esta área de estudio. Un estudio colaborativo donde cada uno puede implementar sus cambios y ofrezca una versión mejorada del trabajo es, especialmente en entornos académicos y de investigación, un beneficio para todos dado que el resultado tendrá mayor interés y calidad.

Pero a su vez, la licencia GPL protege los derechos del autor original. El creador del proyecto se asegura de que cualquier uso futuro del código y los datos respete la atribución del autor original. Esta elección de licencia garantiza que cualquier persona que emplee o quiera redistribuir el trabajo deba atribuir correctamente la autoría del mismo y se le limite el uso comercial no autorizado.

9. Código

El código desarrollado se encuentra en el repositorio Bonpreu-scraper, concretamente, en la carpeta src. El código del programa consta de tres archivos .PY, los cuales son:

1. **main.py**: Script principal para ejecutar el scraper.
2. **scraper.py**: En este archivo está definida la clase BonpreuScraper, la cual contiene las funciones de scraping.
3. **merge_csv.py**: Este script automatiza la fusión de archivos CSVs exportados resultantes de ejecutar main.py.

La recolección de datos de los productos a la venta se realiza siguiendo el procedimiento descrito a continuación[4]:

1. **Consulta de la página principal**: Se accede al contenido HTML de la página web principal del comercio en https://www.compraonline.bonpreuesclat.cat.
2. **Procesamiento de la información**: La información obtenida se procesa con la librería BeautifulSoup[5] para identificar y extraer el menú de categorías de productos disponible en el sitio.

[4] Recordar que este aspecto también ha sido tratado, aunque con menos detalle, en la Figura 1.
[5] Documentación disponible en: https://www.crummy.com/software/BeautifulSoup/bs4/doc/.

3. **Captura de categorías y subcategorías**: De acuerdo con los parámetros especificados por el usuario (*i.e.* categorías y subcategorías), se capturan todos los niveles de subcategorías posibles junto con sus respectivas URLs.
4. **Extracción de información de productos**: Se obtiene el contenido de todas las páginas que contienen productos, extrayendo la información relevante de cada uno.
5. **Exportación de datos**: Finalmente, los datos recolectados se exportan en un archivo CSV.

En cuanto a las dificultades que planteó el sitio web y las soluciones implementadas, a continuación se destacan los puntos más relevantes:

- Para extraer información de todos los niveles de subcategorías en cada categoría, inicialmente se optó por una función recursiva. Este enfoque tenía la ventaja de ser más conciso en cuanto a líneas de código, pero presentaba un problema significativo: el tiempo de ejecución aumentaba considerablemente en comparación con el enfoque final seleccionado. La solución definitiva consistió en identificar previamente el número de niveles de subcategorías y luego programar tantos bucles anidados como niveles existen[6]. Aunque este método es menos elegante visualmente, resultó en una optimización notable del tiempo de ejecución.

- La extracción de todo el contenido de la página de productos representó otro desafío, debido a la presencia de un popup inicial de cookies y la presencia de contenido dinámico. Se abordó la solución de dos maneras diferentes usando el framework *requests* y *Selenium*[7].

  La lógica tanto para el uso de *requests* como para las funciones de *Selenium* se encuentra implementada en el método *_parse_html* del fichero scraper.py. El flag *dynamic_content* indica si se usa *Selenium* o *requests*, para valores *True* o *False*, respectivamente. De este modo se permite escoger cuál de las dos soluciones se usa para el parseo del contenido de la página.

  Para el parseo hecho con *Selenium* se ha implementado un proceso de *scroll-down* en las navegaciones, activando así la carga progresiva de los elementos, ya que se van cargando a medida que vamos se interactúa en ella.

- Aunque no parece que la página de Bonpreu tenga mecanismos de detección de bots, el scraper se ha programado de forma responsable para evitar sobrecargar el servidor. Entre las medidas implementadas se incluyen: la introducción de tiempos de espera entre solicitudes; evitar la carga de imágenes y otro contenido multimedia; o la extracción de productos de manera segmentada por categorías y subcategorías, en lugar de obtener todo el contenido de la web de una sola vez, aunque esta opción también está disponible. Además, para reducir el riesgo de detección, se han integrado prácticas como la rotación de *User-Agents* y la misma introducción de tiempos de espera comentados.

[6] Definida en la función *_extract_subcat_structure()* de la clase *BonpeuScraper* del archivo scraper.py.
[7] Documentación disponible en: https://selenium-python.readthedocs.io/

10. DataSet

Se añade a la carpeta data del repositorio una muestra de la exportación agrupada de todos los productos. El dataset también es publicado en la plataforma de Zenodo en la siguiente URL:

https://doi.org/10.5281/zenodo.14036298

11. Video

Se publica el enlace al vídeo resumen de la explicación del proyecto y de su implementación. Se encuentra en el mismo Drive de la memória o mediante el link directo:

https://drive.google.com/file/d/1uRg6gWCUGmiks3O1khBawmw_bqcXqgG3/view?usp=drive_link

12. Contribuciones

Contribuciones	Firma
Investigación previa	AGJ & DRP
Redacción de las respuestas	AGJ & DRP
Desarrollo del código	AGJ & DRP
Participación en el vídeo	AGJ & DRP