Práctica 1: ¿Cómo podemos capturar los datos de la web?

Tipología y ciclo de vida de los datos
Raquel Calvo Núñez

Noviembre 2025

Metadatos de la práctica

- **Asignatura:** M2.851 - Tipología y ciclo de vida de los datos
- **Estudiante:** Raquel Calvo Núñez
- **Sitio web elegido:** Portal inmobiliario Idealista
  URL: `https://www.idealista.com`
- **Dataset publicado en Zenodo (DOI):**
  `https://doi.org/10.5281/zenodo.17583890`
- **Vídeo de presentación de la práctica:**
  `https://drive.google.com/drive/folders/1pleHt5NFUljfsgqsXm1mNWnFKfgr57t3?usp=sharing`

1. Contexto

Esta práctica 1 de la asignatura *Tipología y ciclo de vida de los datos* del Máster Universitario en Ciencia de Datos de la UOC tiene como objetivo la creación de un dataset a partir de datos obtenidos de la web, lo que se conoce como *web scraping*.

En este caso, he seleccionado el portal **Idealista** (`https://www.idealista.com`), que es uno de los principales sitios web de oferta inmobiliaria en España. Idealista funciona como un punto de encuentro entre inmobiliarias, propietarios y personas que buscan comprar o alquilar, y permite visualizar fácilmente los anuncios publicados de viviendas en distintas zonas. La plataforma destaca porque actualiza con frecuencia su base de datos y recopila muchísimos anuncios de todo el país, lo que la hace una fuente muy fiable para observar el mercado inmobiliario y cómo se comporta en cada zona y ciudad.

En el contexto de esta práctica, me he centrado en la ciudad de Madrid, especialmente en los pisos en venta próximos a **Plaza de Castilla**, con precios hasta 400.000 € y con un número de habitaciones entre dos y cinco. El objetivo principal es construir un dataset estructurado que represente la oferta de vivienda actual en esa zona y que sirva de base para hacer análisis exploratorios o comparativos sobre precios, metros cuadrados y zonas.

Los datos se han recogido de una sección concreta de Idealista que muestra los pisos en venta en la zona centro y norte de Madrid, dentro de ese rango de precio y habitaciones. La dirección exacta de la página usada como fuente es:

`https://www.idealista.com/areas/venta-viviendas/con-precio-hasta_400000,de-dos-dormitorios`
`de-tres-dormitorios,de-cuatro-cinco-habitaciones-o-mas/?shape=%28%28%28wn%7EuFvhsUkI%`
`3FwBo%40cJO%7BDmAwBmAmC%7BCuBiFkBkCeDqNa%40iFoCqP%3FsLb%40eJn%40wHn%40%7BCVkD%3FeK%`
`7E%40iEwAl%40o%40%7DAuB%3F_B%7D%40aM%7C%40%7DDNoJjCmC%7CBuBl%40eDhE_BzCoGjE_BjC%3F%`
`7C%40aCbMqAxEqDtJ_BlBgEjC_B%5EqDOqDmBiByDaCwHo%40%7BDyCsK_BkDa%40uKl%40%7BBnCkDpAm%`
`40lFOrEkCbCkDhBiEdA%7BE%7EAyD%7CA%7BBxCmBjBmBfHyDfAmAzDmB%60FiD%7CD%7DAnJ%3FdD_%40bGiFjB%`
`7D%40pD%3FtBl%40hE%3FtBn%40tE%3F%7EH%7C%40%60FlAdDlBfHvG%7CDzBdDzDhEbLhBfIz%40fIpAdIdDbOfAxDd`
`7EA%7E%40pAjBVzDWzCuBhDsAxFuBpOqAjEeDxEgAzC%7DDdIaC%7CBeD%7CA%29%29%29`

Considero que Idealista es una fuente bastante pertinente y fiable porque:

- Los datos son públicos y actualizados, vienen de anuncios reales de propietarios y agencias inmobiliarias.
- Es una plataforma muy conocida y usada en España para consultar precios y tipos de viviendas.
- La estructura de los datos es clara y homogénea, con campos bien definidos (precio, superficie, habitaciones, ubicación).

2. Título del dataset

Título del dataset: **ventaPisosMadridNorte400k_idealista_noviembre_2025**

Este título representa de forma clara el contenido principal del dataset. Se trata de un conjunto de datos sobre pisos en venta en la zona norte de Madrid, con un precio máximo de 400.000 €.

El dataset recoge información de anuncios de viviendas en venta publicados en el portal Idealista, centrado en la parte norte de Madrid, dentro del área de la M-30. Es una recopilación estructurada que incluye campos como el precio, los metros cuadrados, el número de habitaciones y la zona del piso.

El objetivo del título es reflejar el contenido y el contexto de los datos, destacando que se trata de información obtenida con Idealista durante el mes de noviembre de 2025 y con el objetivo de estudiar la oferta inmobiliaria actual.

3. Descripción del dataset

El dataset **ventaPisosMadridNorte400k_idealista_noviembre_2025** tiene la información de anuncios de viviendas en venta publicadas del portal **Idealista**, centradas en la zona norte de Madrid y conun precio maximo de 400.000euros. Los datos se obtuvieron a partir de páginas HTML descargadas manualmente del sitio web, que luego se procesaron con un script en Python.

Cada registro dentro del dataset representa un anuncio individual, un piso, y contiene las variables básicas que describen ese inmueble: el título del anuncio, el precio de venta, la superficie en metros cuadrados, el número de habitaciones, el barrio o zona donde se encuentra la vivienda y el enlace directo al anuncio original.

El conjunto final cuenta con unas 60 entradas, todas correspondientes a los pisos. El fdataset final en CSV nos permite analizar fácilmente los datos. En conjunto, el dataset refleja una instantánea del mercado inmobiliario en la zona hablada del norte de Madrid durante el mes de noviembre de 2025, mostrando las características más imporatntes de la oferta disponible en ese momento.

4. Representación gráfica

En la Figura 1 se muestra una vista previa del dataset final generado. En ella se pueden ver las columnas principales extraídas del portal Idealista y los primeros registros del proceso de scraping.

	Título	Precio	Metros_cuadrados	Habitaciones	Barrio	url_detalle
0	Piso en Calle de San Dacio, 9, Tres Olivos - V...	300.000€	60	3 hab.	Tres Olivos - Valverde	https://www.idealista.com/inmueble/109393744/
1	Piso en Valdeacederas, Madrid	389.000€	67	3 hab.	Piso en Valdeacederas	https://www.idealista.com/inmueble/109531024/
2	Piso en Calle de las Azucenas, Valdeacederas, ...	285.000€	93	2 hab.	Valdeacederas	https://www.idealista.com/inmueble/109794236/
3	Piso en Tres Olivos - Valverde, Madrid	339.000€	105	3 hab.	Piso en Tres Olivos - Valverde	https://www.idealista.com/inmueble/109725735/
4	Piso en Calle de Sandalio López, Tres Olivos -...	204.900€	55	3 hab.	Tres Olivos - Valverde	https://www.idealista.com/inmueble/109629837/
5	Piso en Calle del Plátano, Valdeacederas, Madrid	342.500€	81	3 hab.	Valdeacederas	https://www.idealista.com/inmueble/109290181/
6	Piso en Ventilla-Almenara, Madrid	350.000€	103	3 hab.	Piso en Ventilla-Almenara	https://www.idealista.com/inmueble/108174078/
7	Piso en Calle Manuela Mínguez, Valdeacederas, ...	239.900€	54	2 hab.	Valdeacederas	https://www.idealista.com/inmueble/109783444/
8	Piso en Berruguete, Madrid	305.000€	62	2 hab.	Piso en Berruguete	https://www.idealista.com/inmueble/109458251/
9	Piso en Tres Olivos - Valverde, Madrid	320.000€	65	3 hab.	Piso en Tres Olivos - Valverde	https://www.idealista.com/inmueble/108746991/

Figura 1: Vista previa del dataset `ventaPisosMadridNorte400k_idealista_noviembre_2025`.

En la Figura 2 se puede ver el diagrama de flujo del trabajo realizado. Cada caja simboliza una fase del proceso para obtener el dataset.

En la primera parte del esquema aparece la fuente de información, **Idealista**, ofrece los anuncios públicos de viviendas en venta tanto de propietarios como de inmobiliarias. A partir de ahí, en el segundo bloque se indica la **descarga manual de las páginas HTML**, que se tuvo que hacer para utilizar el modo más seguro y ético de obtener los datos sin incumplir las reglas de acceso de la web. Después, la siguiente caja representa el **script desarrollado en Python**, el cual se encarga de leer los archivos HTML guardados y extraer los valores más importantes de cada anuncio: el título, el precio, los metros cuadrados, el número de habitaciones, el barrio y la URL al detalle del piso. Esta fase es donde se produce la transformación real de los datos. Por último, la parte final del flujo corresponde al **dataset final en formato CSV**, llamado `ventaPisosMadridNorte400k_idealista_noviembre_2025`, donde toda la información queda guardada y lista para analizar. El esquema visualiza de manera clara la secuencia de step del proyecto.

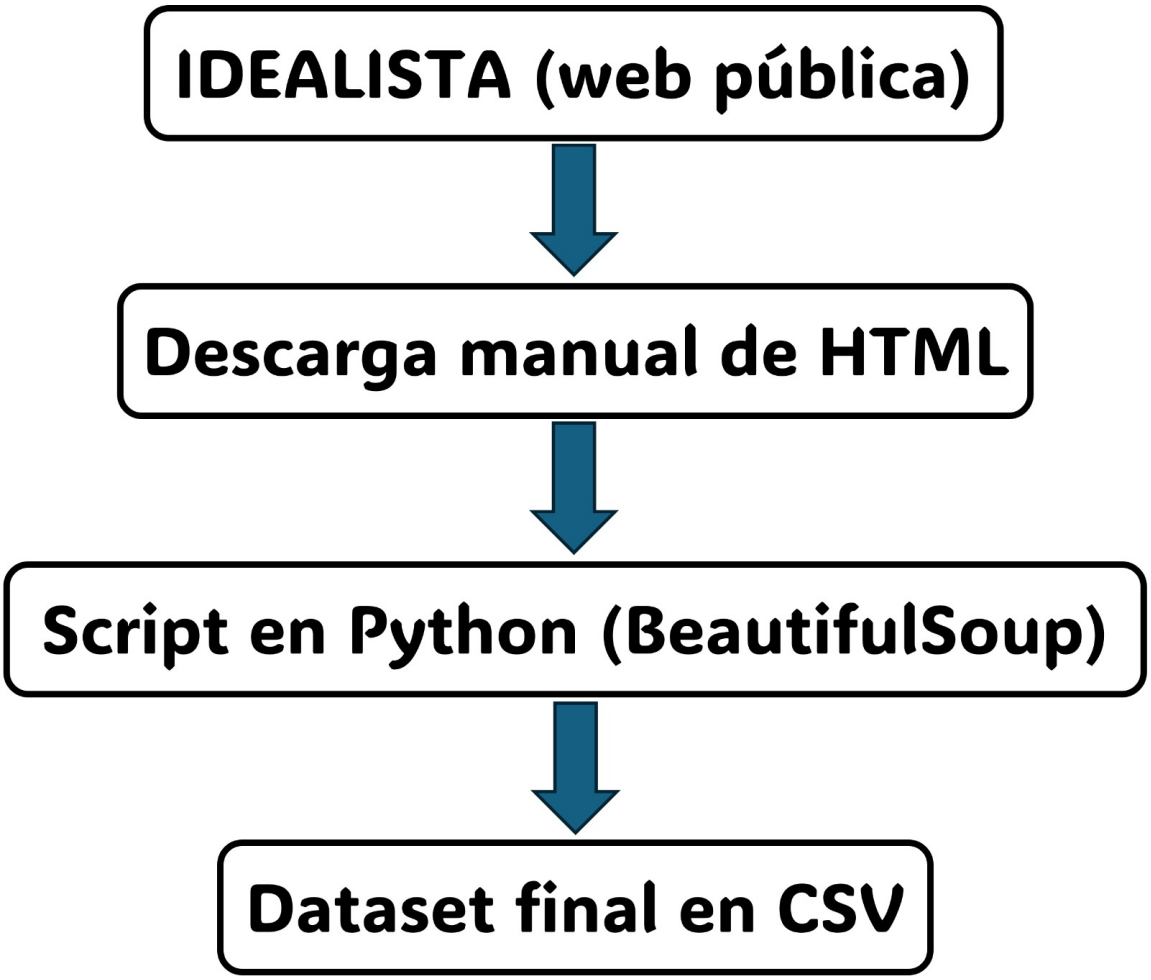

Figura 2: Diagrama de flujo.

5. Contenido

Como se ha visto en el apartado anterior en la Figura 1, el dataset final generado contiene seis columnas principales, que recogen la información más importante de cada anuncio de piso o vivienda. Estas columnas son las siguientes:

- **Título**: descripción principal del anuncio tal como aparece en el portal Idealista, puede incluir el tipo de inmueble y su localización.
- **Precio**: valor total de venta del piso en euros.
- **Metros cuadrados**: superficie del piso medida en metros cuadrados.
- **Habitaciones**: número de dormitorios de la vivienda.
- **Barrio**: zona o distrito dentro de la ciudad de Madrid donde se ubica el piso.
- **url_detalle**: enlace al anuncio original en la web de Idealista.

Cada fila del dataset representa un piso distinto, por lo que cada registro contiene la información completa de un anuncio único. En total se han obtenido sesenta registros válidos. El período temporal de los datos corresponde al mes de **noviembre de 2025**, momento en el que se realizó la descarga manual de las páginas HTML desde el portal Idealista.

6. Propietario

El propietario original de los datos es el portal **Idealista** (`https://www.idealista.com`), que como hemos dicho es una de las plataformas inmobiliarias más conocidas en España y donde se publican a diario miles de anuncios tanto de particulares como de agencias. Idealista actúa como un punto de encuentro entre propietarios y las personas que buscan vivienda. Y por eso tiene una cantidad enorme de información sobre el mercado actual.

En este proyecto no se ha podido acceder directamente a la información mediante web-scraping, ya que el sitio bloquea las peticiones con código y devuelve un error 403. Por eso, y también para respetar las normas del portal, entonces, los datos se obtuvieron de forma manual, descargando las páginas HTML del navegador y trabajando sobreesos archivos locales. Este método no ha sido tan automatico, pero seguro y ético.

No existe un dataset abierto de Idealista que esté disponible para su descarga, aunque si que hay muchos análisis y estudios en Internet que usan sus datos para fines de investigarción comola evolución de los precios o la oferta de viviendas por zonas. En este caso, la idea no era imitar esos trabajos, sino crear un conjunto de datos propio con un objetivo puramente académico y práctico.

Durante todo el proceso se ha tenido cuidado en eguir principios éticos y legales: solo se han usado datos de acceso público, no se ha accedido a información privada. El trabajo se ha limitado a recoger y estructurar la información visible en la web de forma responsable, con un fin educativo. De esta manera, esteel proyecto no infringe derechos de autor ni vulnera la propiedad intelectual del portal.

7. Inspiración

La inspiración principal de este trabajo viene de una situación cotidiana de buscar piso en Madrid. En los últimos años los precios de la vivienda, tanto en venta como en alquiler, han subido muchísimo y cada vez van a más, y especialmente en zonas como el norte de la ciudad, donde se está desarrollando el proyecto de Madrid Nuevo Norte. Ver cómo cambian los precios y qué zonas son más caras o más asequibles se ha vuelto algo casi necesario para cualquiera que quiera mudarse o invertir en una vivienda.

A partir de esa idea surgió el interés por crear un *dataset propio* con datos de Idealista, para poder analizar la oferta actual y ver si se puede identificar qué zonas son las más baratas dentro de un rango de precios. Además, este tipo de conjunto de datos podría servir a futuro para hacer análisis más amplios sobre el mercado inmobiliario en Madrid o incluso compararlo con otras ciudades como Barcelona, donde también los precios están subiendo muy rápido. El objetivo final es aprender una herramienta que permita manejar la información de manera más clara, y que ayude a cer qué zonas ofrecen mejores oportunidades según lo que busque cada persona.

8. Licencia

Este apartado he intentado resumir lo que significa cada lincencia mencionada en el enunciado una para poder decidir cuál se ajusta mejor a este proyecto.

La licencia **CC0 (Public Domain License)** dejá los datos totalmente libres, sin ningún tipo de restricción. El autor renuncia a todos los derechos y cualquiera pudiera usarlos sin mencionar la fuente, modificarlos o incluso venderlos. Está bien para datos totalmente públicos, pero no me parecía adecuada aquí porque el contenido viene de una fuente externa (Idealista) y no son datos creados completamente por mí.

La licencia **CC BY-SA 4.0** permite usar los datos siempre que se cite al autor original y, además, obliga a que cualquier versión modificada se comparta con la misma licencia.

La licencia **CC BY-NC-SA 4.0** es parecida a la anterior, pero añade que el uso no puede ser con fines comerciales. Es decir, cualquiera puede copiar, distribuir o adaptar los datos si cita la fuente y los comparte igual, pero sin hacer negocio con ellos. Por eso creo que es la que mejor encaja con el objetivo de esta práctica, ya que el trabajo es puramente académico y no comercial, y además reconoce la autoría.

También existe la **Open Database License (ODbL)**, que se usa mucho en proyectos de bases de datos abiertos, como los mapas de OpenStreetMap. Permite copiar, modificar y redistribuir los datos si se mantiene la misma licencia y se da crédito al autor. En mi caso no la veo tan necesaria porque el dataset es pequeño y no una base de datos completa.

Después de valorar todas, he decidido publicar el conjunto de datos bajo la licencia **CC BY-NC-SA 4.0**. Esta opción me parece la más equilibrada: permite que otras personas lo usen y aprendan de él, siempre que mencionen la fuente y que no lo usen con fines comerciales. Además, respeta el hecho de que los datos originales proceden de Idealista, que tiene sus propios derechos sobre el contenido de los anuncios.

9. Código

El código que he utilizado para hacer el dataset está hecho en **Python** y lestá guardado en la carpeta `/source`, con el nombre `PRAC1_Raquel_Calvo.ipynb`. En este script se usan principalmente tres librerías: `BeautifulSoup`, que sirve para leer y analizar el contenido HTML; `csv`, que se usa para guardar la información en un fichero separado por comas; y `dataclasses`, que me ayuda a estructurar los datos de forma ordenada, como si fuera una plantilla o una tabla con los campos de cada piso.

Idealista no permite hacer scraping automático (da error 403 si se intenta), lo que hice fue descargar manualmente dos páginas HTML desde el navegador, con los anuncios de pisos en la zona norte de Madrid. Luego, el script abre las HTML, las analiza y va buscando dentro de cada página los bloques `<article>` que contienen la información de cada anuncio. Dentro de esos bloques, se localiza el contenedor con la información del piso (`.item-info-container`) y de ahí se sacan los campos principales: el título, el precio, los metros cuadrados, el número de habitaciones, el barrio y el enlace al anuncio original en Idealista.

Para hacerlo más ordenado, se definió una clase llamada `Piso`, que contiene todas las variables que consideraba importante y quería guardar. De esta forma, cada vez que el programa encuentra un anuncio, crea un objeto nuevo con esos datos y luego lo añade a una lista general. Cuando ya se han procesado todas las páginas, esa lista se convierte en un archivo `CSV` final, que queda guardado con el nombre `ventaPisosMadridNorte400k_idealista_noviembre_2025.csv`.

El flujo del script sería algo así: primero lee los archivos HTML, después extrae la información de cada anuncio, y finalmente genera un dataset estructurado y limpio.

El código también incluye una función llamada `guardar_csv()` que se encarga de escribir los datos en el archivo final. Primero mira si hay datos válidos que guardar, luego crea la carpeta si n existe y por último escribe las filas una a una con las claves de cada objeto de tipo `Piso`.

Durante el desarrollo me encontré con algunas dificultades. La más importante fue que el sitio web de Idealista bloquea las peticiones automáticas, así que tuve que cambiar el planteamiento. En lugar de hacer peticiones con `requests`, descargué las páginas directamente y el script las analizó en local. Esto hizo el proceso menos automatico y más lento, pero limpio y ético, siguiendo las indicaciones de la asignatura. También tuve que ajustar la lectura de los datos porque no todos los anuncios tienen la misma estructura: algunos no muestran el barrio/zona o la superficie, por lo que consideré añadir condiciones que en esos casos se pusiera "None" o vacío.

Las librerías y versiones utilizadas están guardadas en un fichero llamado `requirements.txt`, generadas con el comando:

```
pip freeze > requirements.txt
```

Y las más importantes para este proyecto son:

```
beautifulsoup4==4.12.2
requests==2.31.0
pandas==1.5.3
```