M2.851 - Tipología y ciclo de vida de los datos

Xuan Zheng, Albert Casanova

1. Contexto. Explicar en qué contexto se ha recolectado la información. Explicar por qué el sitio web elegido proporciona dicha información.

En este proyecto hemos decidido explorar el portal inmobiliario Idealista. Los servicios que Idealista ofrece son la compra o alquiler de inmuebles. La mayoría de los anunciantes son empresas del sector inmobiliario, aunque también se permite a los particulares anunciar sus propios inmuebles. Idealista opera en España, Portugal e Italia, y podemos encontrar viviendas, habitaciones, garajes, trasteros, oficinas, locales o naves y terrenos. Además de las opciones comentadas anteriormente, podemos escoger la zona de la que nos interese buscar un inmueble introduciendo la ciudad o barrio que queramos.

La información que se ofrece de cada inmueble pasa por características como el precio, el barrio o distrito donde está ubicado con la posibilidad de localizarlo en un mapa estilo Google Maps, los metros cuadrados, número de habitaciones y baños, la planta en la que se ubica y si tiene o no ascensor, estado general de la vivienda, certificaciones energéticas de consumo y emisiones, etc. También aparecen imágenes y videos de los inmuebles, y algunos incluso ofrecen la posibilidad de realizar un tour virtual en el que nosotros decidimos cómo movernos dentro del inmueble y explorarlo a nuestro gusto.

Hemos decidido investigar directamente la página web de idealista, https://www.idealista.com/ y dentro de España, las viviendas a la venta en la ciudad de Barcelona, concretamente en los distritos de Nou Barris, Sant Andreu y Sant Martí.

2. Título. Definir un título que sea descriptivo para el dataset.

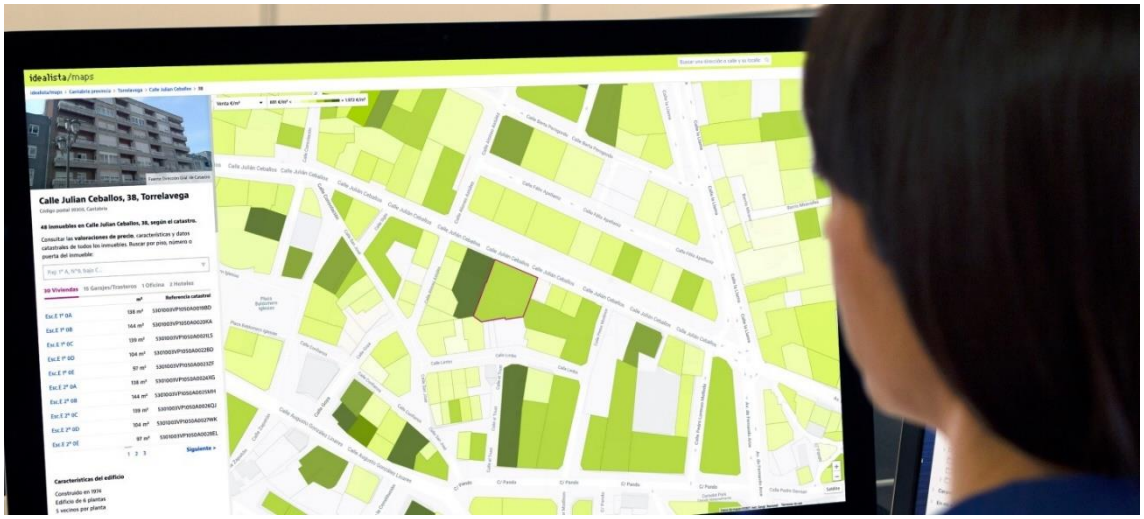
Características de las viviendas a la venta en la ciudad de Barcelona, distritos de Nou Barris, Sant Andreu y Sant Martí.

3. Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído. Es necesario que esta descripción tenga sentido con el título elegido.

El conjunto de datos generado para la Práctica 1 de la asignatura Tipología y ciclo de vida de los datos del Máster Universitario en Ciencia de Datos de la UOC contiene diversas características sobre las viviendas a la venta en la ciudad de Barcelona, y concretamente en los distritos de Nou Barris, Sant Andreu y Sant Martí disponibles en el portal inmobiliario Idealista.

El conjunto de datos extrae la información de la plataforma en las zonas mencionadas anteriormente para las 5 primeras páginas de cada barrio. Los datos resultantes no han pasado un proceso de limpieza de datos, por lo que el resultado no es el óptimo para aplicar un análisis directamente. El fichero resultante es un fichero CSV que permite manejar gran cantidad de datos utilizando poco espacio de almacenamiento, y facilita el tratamiento para realizar posteriores análisis.

4. Representación gráfica. Dibujar un esquema o diagrama que identifique el dataset visualmente y el proyecto elegido.



5. Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se han recogido.

Las características extraídas de la plataforma pertenecen cómo máximo a las 5 primeras páginas que aparecen en el buscador al introducir como target los barrios de las áreas de Nou Barris, Sant Andreu y Sant Martí, dentro de la ciudad de Barcelona. Son las siguientes:

- **Price**: Precio en euros de la vivienda
- **Room**: Número de habitaciones de la vivienda
- **Space**: Superficie interior de la vivienda, expresada en metros cuadrados
- **Name**: Título del anuncio
- **Link**: Hiperenlace que nos dirige a la página web donde encontramos las características de la vivienda
- **City**: Ciudad donde se ubica la vivienda
- **Area**: Distrito donde se ubica la vivienda
- **Subarea**: Barrio del distrito donde se ubica la vivienda.
- **Page**: Número de página donde se encuentra el anuncio de la vivienda.

La recolección de los datos ha sido a través de web scraping utilizando el lenguaje de programación Python junto a las librerías BeautifulSoup y Selenium. Primero definimos los distritos que queremos explorar, y a continuación de manera automática entramos en los barrios que componen cada distrito, de donde recogemos para cada vivienda la información deseada.

6. Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de análisis anteriores o, en caso de no haberlas, justificar esta búsqueda con análisis similares. Justificar qué pasos se han seguido para actuar de acuerdo con los principios éticos y legales en el contexto del proyecto.

El propietario del conjunto de datos es la empresa Idealista, S.A.U. Gestionan la página web y aplicaciones, tienen sede en Madrid, en Plaza de las Cortes y están inscritas en el Registro Mercantil de Madrid con NIF A-82505660.

En Internet existen varios análisis que tratan de hacer web scraping sobre el portal inmobiliario Idealista:

- https://github.com/David-Carrasco/Scrapy-Idealista
- https://github.com/hmeleiro/idealisto

El primero de los dos utiliza como nosotros lenguaje Python pero una librería distinta, scrapy, y para que no baneen la IP a la hora de scrapear utiliza una pool pública de proxies que a día de hoy suelen estar muertos por lo que no funcionan. El segundo utiliza como lenguaje de programación R, y en su ReadMe nos avisa de que el scraper no funciona hoy en día, ya que Idealista lo ha bloqueado. Ninguno de los dos ofrece un dataset con el resultado del scraping que se realizó en su momento.

Para este proyecto, el propietario de la página web indica que no se pueden utilizar mecanismos automáticos para copiar o extraer su contenido. También se ha identificado una API que permite consultar sus datos, aunque el límite de peticiones permitidas es muy bajo por lo que no contemplamos su uso para nuestra investigación. Hemos decidido realizar web scraping ya que el sitio web expone información de forma pública, además, el uso que vamos a hacer de los datos es académico y no comercial. Por otra parte, no estamos copiando la totalidad de la página con todos los datos para replicarla y el uso que hacemos es moderado al extraer pocos datos de forma que tampoco saturamos los servidores con muchas peticiones.

7. Inspiración. Explicar por qué es interesante este conjunto de datos y qué preguntas se pretenden responder. Es necesario comparar con los análisis anteriores presentados en el apartado 6.

El conjunto de datos es interesante para conocer el estado actual del sector de la vivienda en Barcelona. Es posible obtener conclusiones sobre la tendencia del mercado analizando datos como el precio o los metros cuadrados, que en un futuro pueden volver a extraerse para realizar comparaciones y ver su evolución. Esta evolución nos permitirá conocer los patrones de comportamiento de los compradores, y qué tipo de viviendas son las que ofrecen las grandes empresas inmobiliarias, que son las que más ofertas publican en la plataforma.

Las preguntas que se pretenden responder son: ¿En qué distrito el precio medio de una vivienda es superior? ¿Cuál es el promedio de metros cuadrados de las viviendas en los distritos investigados? ¿Qué barrio tiene las viviendas más caras por metro cuadrado? ¿Y más baratas por metro cuadrado?

8. Licencia. Seleccionar una de estas licencias para el dataset resultante y justificar el motivo de su selección:

La licencia escogida ha sido la Licencia Pública General de GNU (GNU GPL). Se ha escogido una licencia pública ya que el propósito del proyecto es académico y no tiene en mente utilizarse para usos comerciales, aunque la licencia permite la comercialización del código. Utilizando la Licencia Pública General de GNU garantizamos a los usuarios finales (ya sean personas, organizaciones o compañías) poder utilizar, compartir o modificar el código, siempre teniendo en cuenta el espíritu de colaboración y cooperación que abre la posibilidad a los usuarios de ayudarse entre ellos para corregir errores e implantar mejoras. Además, la licencia exige que las siguientes versiones del código que se publiquen también sean publicadas bajo una licencia pública.

9. Código. Adjuntar en el repositorio Git el código con el que se ha generado el dataset, preferiblemente en Python o, alternativamente, en R.

El código se encuentra en el siguiente repositorio: https://github.com/acasanovago/Web-scraping-housing-barcelona

Concretamente, el código fuente está en src/idealista_scraper.py

10. Dataset. Publicar el dataset obtenido(*) en formato CSV en Zenodo con una breve descripción. Obtener y adjuntar el enlace del DOI.

Dataset publicado en Zenodo: https://zenodo.org/record/6436017#.YlMK2tNBzIV

El DOI es el siguiente: https://doi.org/10.5281/zenodo.6436017

Aunque también se encuentra en el repositorio del apartado anterior, en csv/housing-barcelona.csv

11. Vídeo. Se debe hacer entrega de un vídeo explicativo de la práctica en donde cada uno de los integrantes del grupo explique con sus propias palabras tanto las respuestas del proyecto como el código utilizado para llevar a cabo la extracción. El vídeo debe ser enviado a través de un enlace a Google Drive que deben proporcionar, junto con el enlace al repositorio Git, al momento de entregar la práctica.

El enlace del vídeo almacenado en Google Drive es el siguiente:

https://drive.google.com/file/d/1KprYSO5grH6AIldePkibqED7A_p9wvH7/view

Se tiene acceso directo con una cuenta @uoc.edu, en cambio con un correo externo a la UOC se debe pedir permiso para verlo.

Contribuciones	Firma
Investigación previa	XZ, AC
Redacción de las respuestas	XZ, AC
Desarrollo del código	XZ, AC