PRÁCTICA 1. WEB SCRAPING

Juan Alonso Franco Blanco
Juan Prieto Pena

9 DE NOVIEMBRE DE 2020
UNIVERSITAT OBERTA DE CATALUNYA
ASIGNATURA DE TIPOLOGÍA Y CICLO DE VIDA DE LOS DATOS

Contenido

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en una web. Para su realización, se deben cumplir los siguientes puntos:

EJERCICIO 1

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

El punto de partida de esta PEC ha sido la obtención de conocimiento respecto a los precios de bienes inmuebles en todo el mundo. Existen páginas especializadas (como Idealista) con un conjunto de datos muy interesante, pero que no ofrece información exhaustiva de precios en algunos países. Tras una búsqueda por internet se ha encontrado el sitio preciosmundi[1], que contiene precios de referencia para las principales ciudades de la mayoría de países del mundo.

Se ha recolectado la información de los precios promedio de la vivienda para cada país, clasificado de la siguiente forma:

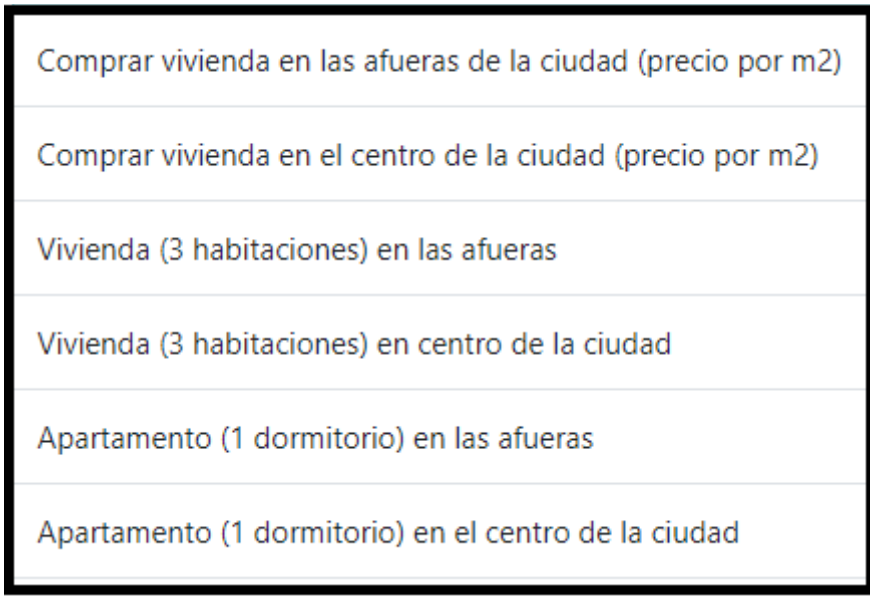

Figura 1: Información disponible en el sitio web preciosmundi.com sobre precios de vivienda en diferentes países del mundo.

De esta forma obtendremos una matriz de precios de la anterior clasificación, obteniendo una fila por país estudiado y 6 columnas con los diferentes ítems recolectados.

El modelo de negocio de esta web es el cobro de publicidad, para ello necesita visitas a su sitio. A cambio proporciona al viajero un precio de referencia de utilidad para que los interesados puedan conocer cuánto tienen que pagar en los diferentes países del mundo. El estudio de precios, en este caso de bienes inmuebles, resulta de interés para estudiar el nivel de vida de un país y su riqueza.

[1] www.preciosmundi.com

Por último, se ha verificado la veracidad de los precios obtenidos del activo que hemos extraído comparándolos con el índice de precios de Idealista para Madrid. Se ha utilizado este portal porque es el sitio más utilizado en España para la búsqueda de vivienda, y se ha escogido Madrid por ser la capital de España (y, como es lógico, una de sus principales ciudades).

Precios de alquiler o compra de vivienda en España

Producto	Dólar ($)	Euro (€)
Comprar vivienda en las afueras de la ciudad (precio por m2)	2335,50$	2000,00€
Comprar vivienda en el centro de la ciudad (precio por m2)	3620,02$	3100,00€
Vivienda (3 habitaciones) en las afueras	899,17$	770,00€
Vivienda (3 habitaciones) en centro de la ciudad	1167,75$	1000,00€
Apartamento (1 dormitorio) en las afueras	595,55$	510,00€
Apartamento (1 dormitorio) en el centro de la ciudad	782,39$	670,00€

Figura 2: Precios de vivienda en España según preciosmundi.com

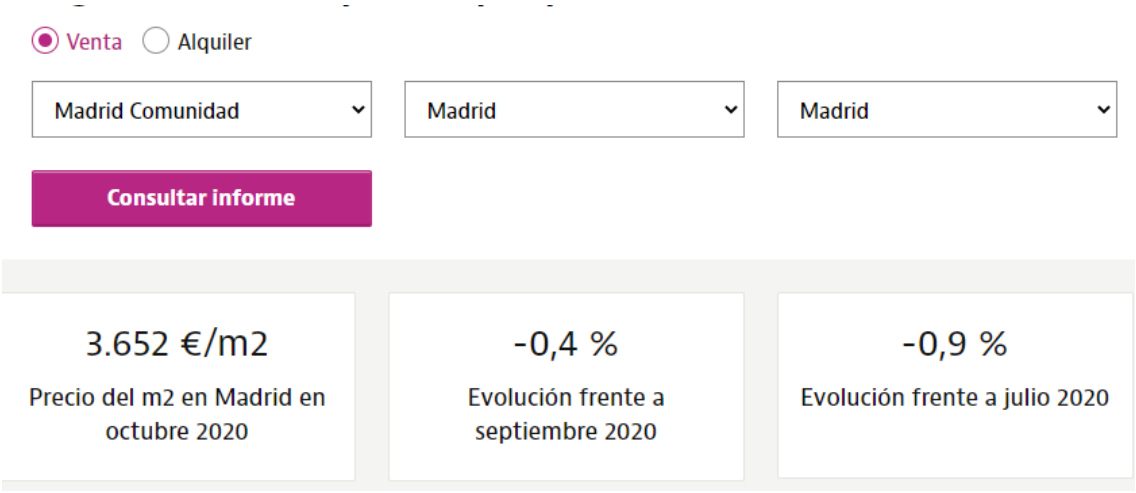

Figura 3: Precios de vivienda en Madrid según Idealista.com

La divisa de referencia de Idealista es el euro, costando el metro cuadrado en Madrid 3652€. Si lo comparamos con la información de preciosmundi (3100 €), podemos ver que ambos precios se encuentran en el mismo rango. La verificación de la información para la validez del conjunto de datos nos sirve para asegurarnos que las estimaciones de los precios que necesitamos son correctas.

EJERCICIO 2

Definir un título para el dataset. Elegir un título que sea descriptivo.

En nuestro ejemplo ha sido fácil la elección porque el nombre incluye el activo y la segmentación geográfica, el nombre elegido es *PreciosViviendaPorPais.csv*

EJERCICIO 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).

En nuestro dataset tenemos una estrecha relación entre el nombre elegido y el juego de datos. En el nombre del dataset tenemos:

- Los indicadores del conjunto de datos, en este caso el precio, con la palabra “precios”.
- La dimensión del juego de datos, en este caso país con la palabra “país”.
- El activo que estamos evaluando, en este caso viviendas, con la palabra vivienda.

Por tanto, tan sólo hay que incluir en el juego de datos la clasificación o segunda dimensión para poder consultar los indicadores, esta clasificación es la siguiente:

- Vivienda a las afueras, compra (USD/m2)
- Vivienda en el centro de la ciudad, compra (USD/m2)
- Vivienda (3 habitaciones) a las afueras, alquiler (USD)
- Vivienda (3 habitaciones) en el centro de la ciudad, alquiler (USD)
- Apartamento (1 dormitorio) en las afueras, alquiler (USD)
- Apartamento (1 dormitorio) en el centro de la ciudad, alquiler (USD)

Se ha decidido extraer sólo la información de precios en dólares, dado que es la moneda global de referencia.

EJERCICIO 4

Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente

En la imagen se puede apreciar las dos dimensiones con todos sus indicadores, en el eje X está la clasificación del activo (vivienda), mientras que en el eje Y tenemos la segmentación por países:

	Vivienda a las afueras, compra (USD/m2)	Vivienda en el centro de la ciudad, compra (USD/m2)	Vivienda (3 habitaciones) a las afueras, alquiler (USD)	Vivienda (3 habitaciones) en el centro de la ciudad, alquiler (USD)	Apartamento (1 dormitorio) en las afueras, alquiler (USD)	Apartamento (1 dormitorio) en el centro de la ciudad, alquiler (USD)
Afganistan	374.75	788.54	153.55	266.75	71.57	132.73
Albania	779.89	1519.38	314.78	526.19	184.17	298.8
Alemania	4087.12	5955.52	1284.52	1634.85	653.94	864.13
Andorra	3269.7	3853.57	980.91	1401.3	572.2	829.1
Angola	3999.96	5999.95	9000.07	9979.32	88.35	171.6
Arabia-saudita	693.29	986.6	506.63	666.62	266.65	373.31

Figura 4: Esquema de los datos presentes en nuestro dataset, tomados de preciosmundi.com

EJERCICIO 5

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

Los campos que incluye el dataset sobre el precio de la vivienda en 140 países diferentes son los siguientes:

- Vivienda a las afueras, compra (USD/m2)
- Vivienda en el centro de la ciudad, compra (USD/m2)
- Vivienda (3 habitaciones) a las afueras, alquiler (USD)
- Vivienda (3 habitaciones) en el centro de la ciudad, alquiler (USD)
- Apartamento (1 dormitorio) en las afueras, alquiler (USD)
- Apartamento (1 dormitorio) en el centro de la ciudad, alquiler (USD)

Los datos son una instantánea de la situación actual de los precios de vivienda en todo el mundo. La página web no almacena datos históricos. En cada una de las páginas de información podemos ver el momento de la última actualización, como se puede ver en la siguiente figura:

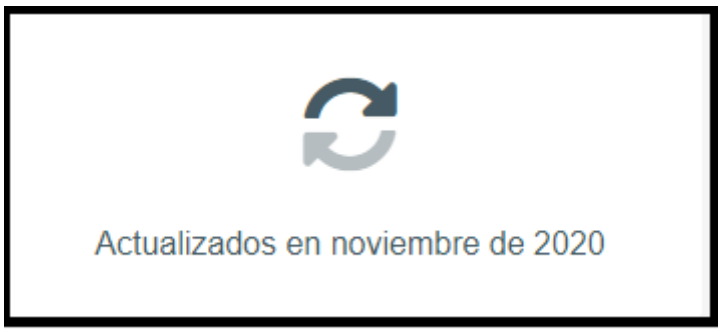

Figura 5: Imagen que muestra el último periodo de actualización de la página web preciosmundi.com

Como podemos ver en el sitio web, los datos dinámicos y actualizándose mes a mes. En esta práctica hemos obtenido los datos correspondientes al mes actual (Noviembre de 2020).

EJERCICIO 6

Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

El propietario del sitio web es Víctor Rodríguez Obensa. En ningún sitio de la web se explicita el origen de los datos o propiedad intelectual. La información legal de la página está publicada en la web[2].

EJERCICIO 7

Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

Este conjunto de datos tiene como objetivo el estudio de los precios de la vivienda en el mundo. El valor de la vivienda, para la gran mayoría de las familias, es la mayor inversión a lo largo de su vida debido a su elevado precio. En numerosas ocasiones hay una opción de inversión en una segunda vivienda, por ejemplo, para una estancia vacacional o de compra en otro país con

[2] https://preciosmundi.com/legal

el fin de realizar una inversión, y es en este caso donde este comparador de precios puede hacer ver la gran cantidad de oportunidades que hay en el mercado, generalmente limitadas al mercado interior. Se podrían encontrar precios más bajos con unas calidades más adaptadas al interés del inversor.

Hay que destacar que este tipo de inversiones suele hacerse en la capital del país o en las ciudades más importantes. Este conjunto de datos tiene en cuenta además de la capital del país las grandes ciudades.

Un ejemplo de inversión extranjera es España. En Madrid y Barcelona hay muchos inversores de América que compran en zonas con rentabilidades altas para la compra/venta o alquiler de vivienda.

Con estos datos podemos hacer un estudio de precios en el mundo, podemos tener un precio referencia de un activo con grandes rentabilidades que atraen al inversor.

En resumen, el juego de datos nos da información útil para las curiosidades del inversor. Demuestra que hay oportunidades en el mundo donde se puede obtener mayores rentabilidades ya que el precio es el indicador principal en el cálculo de rentabilidad. En este caso es una buena referencia para abrir una primera investigación en un país determinado.

EJERCICIO 8

Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

- **Released Under CC0: Public Domain License**
- **Released Under CC BY-NC-SA 4.0 License**
- **Released Under CC BY-SA 4.0 License**
- **Database released under Open Database License, individual contents under Database Contents License**
- **Other (specified above)**
- **Unknown License**

Hemos decidido emplear la licencia CC BY 4.0 debido a que es una licencia que permite el libre uso de los datos generados por parte de otras personas y con la única condición de que se deba dar atribución a los autores del dataset. Junto con su subida a un repositorio público indexado y su formato csv, el uso de este tipo de licencias permite hacer que los datos cumplan las líneas especificadas en "FAIR Guiding Principles for scientific data management and stewardship"[3]

La publicación en abierto en el repositorio Zenodo selecciona por defecto la licencia elegida[4].

EJERCICIO 9

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

El dataset ha sido generado en Python y está en el fichero *PreciosViviendaPorPais.ipynb* en el repositorio de GitHub[5]

En la pestaña wiki del repositorio está disponible la información sobre el código que ha generado el dataset y las personas que hemos realizado esta práctica. En la pestaña “code” está el script con el código utilizado y el dataset generado.

[3] The FAIR Guiding Principles for scientific data management and stewardship, Mark Wilkinson et al. Nature Scientific Data 3, 160018 (2016), en https://www.nature.com/articles/sdata201618
[4] https://creativecommons.org/licenses/by/4.0/deed.es
[5] https://github.com/jpripen/UOC_TCVD_PRAC_1

- Dataset: *PreciosViviendaPorPais.csv*
- Código: *PreciosViviendaPorPais.ipynb*

EJERCICIO 10

Dataset. Publicación del dataset en formato CSV en Zenodo (obtención del DOI) con una breve descripción.

https://doi.org/10.5281/zenodo.4256508

Tabla de contribuciones

Contribuciones	Firma
Investigación previa	JAFB, JPP
Redacción de las respuestas	JAFB, JPP
Desarrollo código	JAFB, JPP