



Tipología y ciclo de vida de los datos.
Máster Universitario en Ciencia de Datos

Práctica 1:
Scraper a una web inmobiliaria.

Autores:

- Andrea Giralt Castellano
- Manuel Fernández Álvarez

Profesor: Jose Moreira Sanchez

Lunes, 8 de noviembre de 2021

Índice

1. Contexto	3
2. Título	3
3. Descripción del <i>dataset</i>	4
4. Representación gráfica	4
5. Contenido	5
6. Agradecimientos	17
7. Inspiración	18
8. Licencia	18
9. Código	19
10. <i>Dataset</i>	20
Referencias	21

1. Contexto

En este proyecto se ha decidido extraer información de la web pisos.com. Esta plataforma, donde el producto a comercializar son inmuebles, permite unir tanto a los compradores como a los vendedores. Es importante destacar que los anunciantes pueden ser tanto particulares como inmobiliarias.

En un entorno donde conseguir un comprador es complicado, utilizar plataformas donde poder anunciarte de forma gratuita es primordial. Por esta razón, muchos particulares o inmobiliarias utilizan estas plataformas para publicitarse y darse a conocer utilizándolas, también, para poder analizar los precios de mercado con el fin de dar un precio competitivo respecto al resto de inmuebles.

Es por eso que este análisis permitirá tanto a particulares como a los profesionales de inmobiliaria a:

- Hacer estudios para identificar las zonas más caras o que están en auge.
- Detectar gangas o viviendas que están muy por encima del precio del mercado.
- Estimar el precio de una vivienda, dada sus características.
- Análisis de posibles variables estacionales.
- Análisis de un posible incremento en el precio si el anunciante es una inmobiliaria.
- Análisis de la posible influencia del paro en el precio de un inmueble.

Por otro lado, es importante destacar que nuestros datos no son de carácter personal, ya que solamente tenemos información de los inmuebles y no de las personas a las cuales pertenecen, luego cumplimos la RGPD.

Aun así, si tuviésemos datos de carácter personal, podríamos utilizarlos y cumplir con la RGPD, siempre y cuando tuviésemos el consentimiento del usuario poseedor de los datos.

Por último, con el fin de realizar un estudio de posible influencia del paro al precio de la vivienda, se procesado empleando un *spider*, el último csv disponible en la URL: datos.gob.es el cual, actualmente, recoge los datos de paro por municipio a nivel estatal hasta septiembre de 2021.

2. Título

Análisis del sector inmobiliario en Mallorca.

3. Descripción del *dataset*

El dataset, por tanto, recoge los datos más relevantes que pueden influir en el precio de un inmueble. En este dataset, se presentan los datos obtenidos de una descarga puntual de un día, de los inmuebles de Mallorca. Las unidades del precio de la vivienda son en euros. Hay que tener en cuenta que los datos no han pasado por ningún proceso de limpieza, por lo que este proceso se realizará en la práctica siguiente. El formato final del dataset es un fichero CSV.

4. Representación gráfica

Como podemos ver en la imagen, existen dos tipos de publicitantes, las personas físicas, llamadas vendedores en la imagen, o las inmobiliarias. Ambos pueden poner un anuncio de forma gratuita en pisos.com sobre un inmueble. Por otro lado, existe la posibilidad de comprar anuncios clasificados, y es de aquí de donde pisos.com obtienen sus ganancias. Estos anuncios, son destacados en la web por encima de otros e incluso son propuestos a aquellos compradores que por las características del inmueble cumplan sus requisitos.

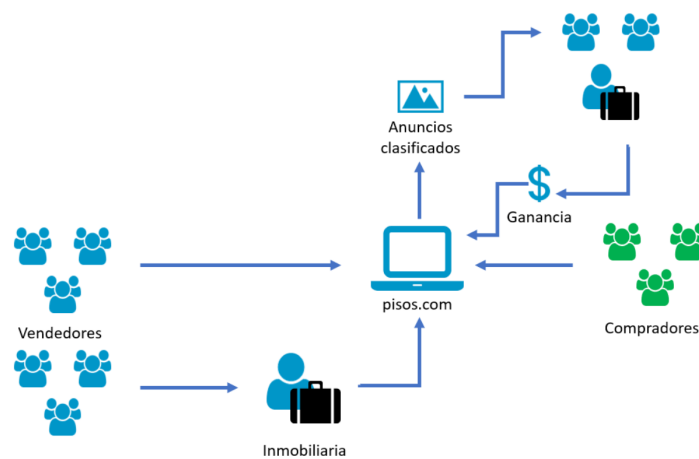


Figura 1: Representación gráfica

5. Contenido

Los datos descargados se han obtenido en un día puntual, obteniendo un total de 5736 anuncios.

- *id*: Identificador del anuncio en la web.
- *type*: Tipo de vivienda (Piso, Chalet, Apartamento...)
- *title*: Título del anuncio.
- *description*: Breve descripción del anuncio
- *town*: Municipio
- *zone*: Localidad
- *price*: Precio de la vivienda
- *surface*: Superficie en m2
- *rooms*: Número de habitaciones
- *bathrooms*: Números de baños
- *floor*: Planta donde se localiza el inmueble
- *longitude*: Coordenada longitud.
- *latitude*: Coordenada latitud.
- *url*: Enlace al anuncio.
- *exact_position*: Si las coordenadas corresponden a una localización exacta.
- *recently_date*: Fecha de publicación
- *is_promo*: Si es una promoción.
- *image_url*: Foto principal del inmueble.
- *owner_name*: Nombre del publicador del anuncio.
- *old_price*: Si ha habido una rebaja, correspondería al precio anterior.

Por otro lado, se ha obtenido la siguiente información relativa al paro por municipio de Mallorca.

- **Código Municipio:** Código postal del municipio.
- **Municipio:** Nombre del municipio.
- **total Paro Registrado:** Número total de parados.
- **Paro hombre edad < 25:** Total de hombres parados menores de 25 años.
- **Paro hombre edad 25 - 45 :** Total de hombres parados entre 25 y 45 años.
- **Paro hombre edad >=45 :** Total de hombres parados mayores de 45 años.
- **Paro mujer edad < 25:** Total de mujeres paradas menores de 25 años.
- **Paro mujer edad 25-45 :** Total de mujeres paradas entre 25 y 45 años.
- **Paro mujer edad <=45 :** Total de mujeres paradas mayores de 45 años.
- **Paro Agricultura :** Total de parados en el sector de la agricultura.
- **Paro Industria :** Total de parados en el sector de la industria.
- **Paro Construcción:** Total de parados en el sector de la construcción.
- **Paro Servicios:** Total de parados en el sector de servicios.
- **Paro Sin empleo Anterior:** Total de parados que no han tenido empleo con anterioridad.

Cómo se ya se ha comentado con anterioridad, los datos no han pasado por ningún proceso de limpieza, será en una segunda parte cuando se realizará la limpieza y se enlazará la información obtenida del portal inmobiliario, con el paro por municipio.

Los datos fueron recogidos a través de *web scraping* en lenguaje *python* sobre la página web: pisos.com. Como se ha indicado previamente, los datos extraídos se han guardado en un fichero CSV. Por lo tanto, para realizar este proyecto, se han seguido los siguientes pasos:

1. Al tratarse de un equipo, aunque solo de dos personas, hemos empleado trello.com para gestionar las tareas que se tenían que realizar.

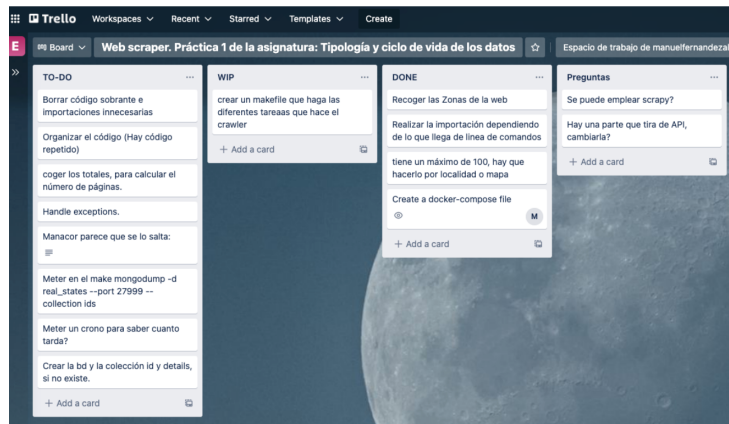


Figura 2: Imagen del panel *trello*

2. Para proceder con el *scraping*, hemos entrado en la web de pisos.com, seleccionando "Islas Baleares", ya que nuestro análisis será exclusivamente de la isla de Mallorca.

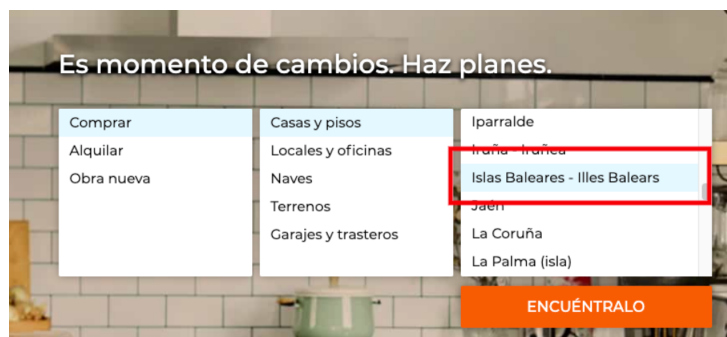


Figura 3: Página de selección de provincia

3. A continuación, seleccionamos Mallorca:

Casas y pisos en venta en Islas Baleares - Illes Balears

[Ver 9.867 resultados](#)



Figura 4: Selección de Mallorca en el mapa

4. Como la web solo permite llegar hasta 100 páginas en las búsquedas, decidimos parsear por municipio en lugar de la isla entera.

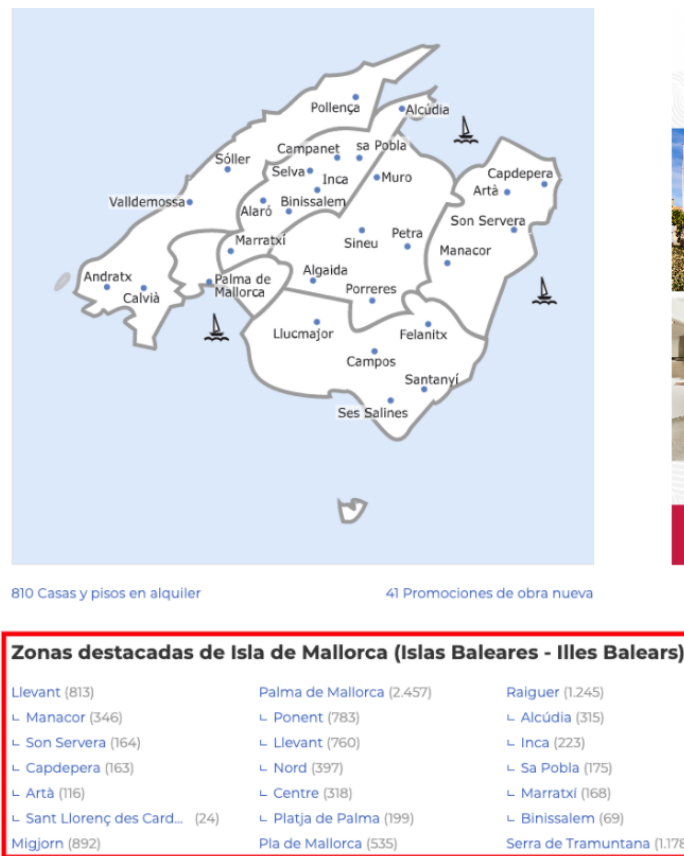


Figura 5: Apartado con el listado de municipios

Por lo tanto, parseamos los enlaces y los totales (en cada página de búsqueda hay 30 anuncios). Lo descargamos en un csv, para que el siguiente *spider* pueda hacer uso de él. (zones.py)

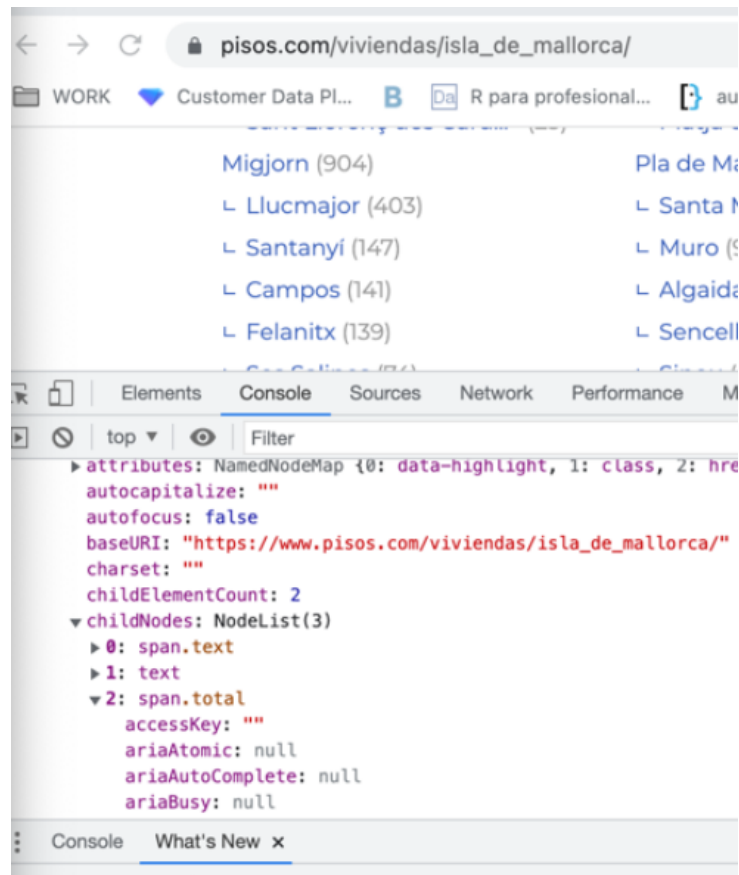


Figura 6: Extracción de enlaces usando la consola

5. Vistamos los enlaces obtenidos en el paso anterior y vamos parseando por página:

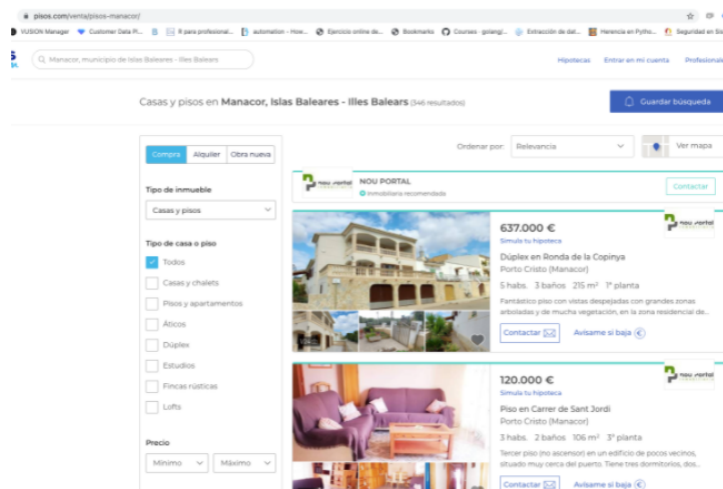


Figura 7: Listado de anuncios, paginado, de un municipio.



Figura 8: Información extraída de las fichas de los inmuebles.

6. De estos listados (Figura 8) nos descargamos la información previamente comentada:

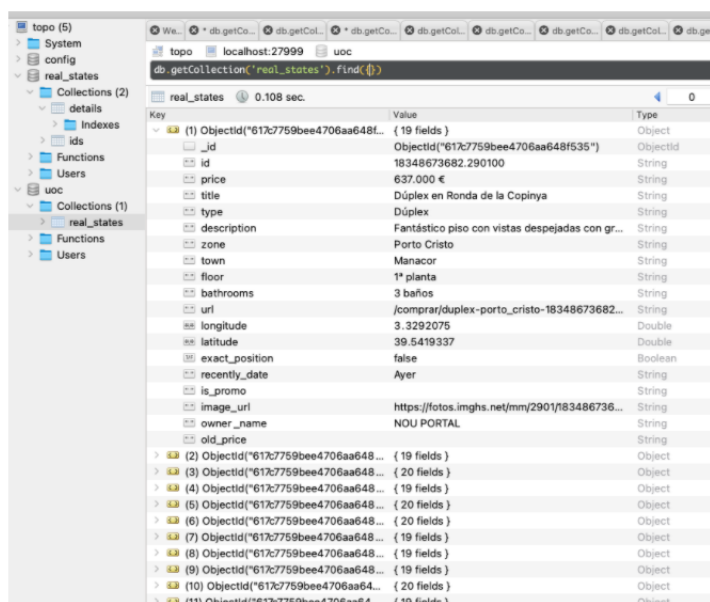
1. Imagen principal del anuncio.
2. Precio de la vivienda.
3. Tipo de vivienda.
4. Título
5. Municipio/Localidad.
6. Número de habitaciones.
7. Número de baños.
8. Superficie de la vivienda.

9. Planta

10. Breve descripción del anuncio.

Para procesar los campos de la ficha del anuncio, no solo hemos empleado selectores CSS3, también hemos hecho uso de expresiones regulares, como por ejemplo, para extraer el tipo de vivienda.

7. Esta información la hemos guardado en una base de datos NoSQL MongoDB, para después completarla con llamadas a una API en un tercer paso.



Key	Value	Type
(1) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object
_id	ObjectId("617c7759bee4706aa648f535")	ObjectId
id	18348673682.290100	String
price	637.000 €	String
title	Dúplex en Ronda de la Copinya	String
type	Dúplex	String
description	Fantástico piso con vistas despejadas con gr...	String
zone	Porto Cristo	String
town	Manacor	String
floor	1ª planta	String
bathrooms	3 baños	String
url	/comprar/duplex-porto_cristo-18348673682...	String
longitude	3.3292075	Double
latitude	39.5419337	Double
exact_position	false	Boolean
recently_date	Ayer	String
is_promo		String
image_url	https://fotos.imghs.net/mm/2901/183486736...	String
owner_name	NOU PORTAL	String
old_price		String
(2) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object
(3) ObjectId("617c7759bee4706aa648f535")	{ 20 fields }	Object
(4) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object
(5) ObjectId("617c7759bee4706aa648f535")	{ 20 fields }	Object
(6) ObjectId("617c7759bee4706aa648f535")	{ 20 fields }	Object
(7) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object
(8) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object
(9) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object
(10) ObjectId("617c7759bee4706aa648f535")	{ 20 fields }	Object
(11) ObjectId("617c7759bee4706aa648f535")	{ 19 fields }	Object

Figura 9: Colección Mongo con la información básica

8. Por otro lado, nos dimos cuenta que en el mapa que contiene los anuncios, se construía a través de llamadas a una API.

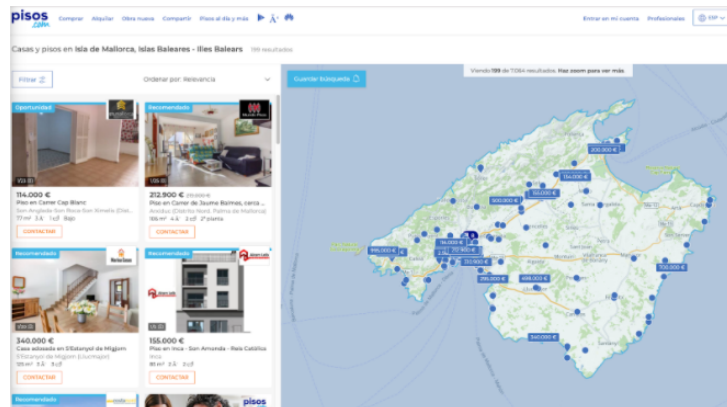


Figura 10: Mapa interactivo de Mallorca

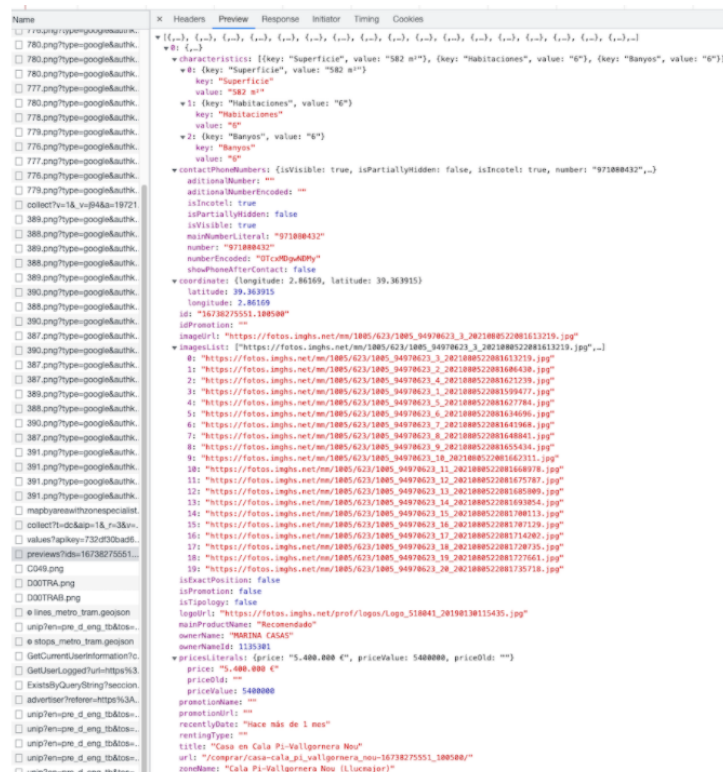


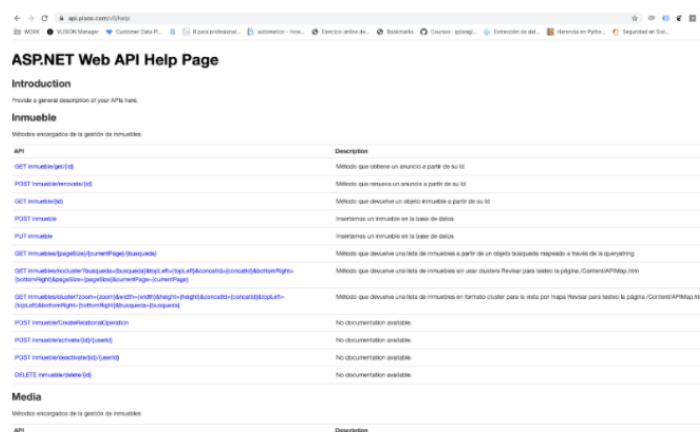
Figura 11: Inspección de una llamada a la API

Con el fin de recolectar la información, hemos completado la información obtenida de las fichas de los anuncios con el resultado de llamar a esta API.

Para poder llamar a la API, tenemos que emplear una cookie y una api-key asociada a ella, por lo que nos vimos en la obligación de añadir esta información a la cabeceras de las solicitudes.

Completamos la información obtenida de la ficha del producto, con los siguientes campos: *url*, *longitude*, *latitude*, *exact_position*, *recently_date*, *is_promo*, *image_url*, *owner_name*, *old_price*.

Por nuestra experiencia, sabíamos que, normalmente, las APIs, en la raíz, disponen de ayuda, por lo que descubrimos que, visitando la raíz, podíamos ver los distintos *endpoints* de los que disponen en la web, tal y como podemos ver en la siguiente imagen:



API	Description
GET /api/properties/{id}	Método que obtiene un anuncio a partir de su id
POST /api/properties/{id}	Método que inserta un anuncio a partir de su id
GET /api/properties	Método que devuelve un objeto inmueble a partir de su id
POST /api/properties	Insertamos un inmueble en la base de datos
PUT /api/properties	Insertamos un inmueble en la base de datos
GET /api/properties/{id}/search/{searchTerm}	Método que devuelve una lista de inmuebles a partir de un objeto búsqueda pasado a través de la querystring
GET /api/properties/{id}/search/{searchTerm}/page/{pageNumber}	Método que devuelve una lista de inmuebles en una página específica para todos los países. ContentAPIMap.htm
GET /api/properties/{id}/search/{searchTerm}/page/{pageNumber}/page/{pageNumber}	Método que devuelve una lista de inmuebles en formato cluster para la vista por mapa. Ver más en la página ContentAPIMap.htm
POST /api/properties/{id}/search/{searchTerm}	No documentation available.
POST /api/properties/{id}/search/{searchTerm}	No documentation available.
POST /api/properties/{id}/search/{searchTerm}	No documentation available.
DELETE /api/properties/{id}	No documentation available.

Figura 12: Documentación de la API

Después de probar varios *endpoints*, nos dimos cuenta de que, a parte de carecer de documentación, muchos de ellos no funcionaban.

9. Por último, con el fin de realizar un análisis de posible correlación del paro de un municipio con el precio de las viviendas, realizamos un procesado de la [web del paro](#), extrayendo información de paro por municipio del país.

[illegible]

Figura 14: Previsualización excel de gob.es

15

PARO REGISTRADO POR MUNICIPIOS DESGLOSADO POR SEXO, TRAMOS DE EDAD Y SECTOR DE LA ACTIVIDAD ECON

mi	mes	Código de	Comunidad	Código Pr	Provincia	Código M	Municipio	total Paro Reg	Paro hombre	Paro hombre
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		Provincia	43	57	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			40	74	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			142	155	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			32	64	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			60	94	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			55	72	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			3	4	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			54	92	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			9	8	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			35	44	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			267	386	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			13	29	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			68	90	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			68	83	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			141	185	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			19	36	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			10	24	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			4	2	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			0	1	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			22	34	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			0	3	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			154	193	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			12	13	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			41	61	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			3	2	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			405	513	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			350	445	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			7	20	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			60	92	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			11	27	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes			247	373	
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		07032 Maó	1255	52	185
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		07033 Manacor	2290	141	376
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		07034 Mancor de la	44	2	6
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		07035 Maria de la Sa	93	5	20
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		07036 Marratxí	1471	99	219
2107	Julio de 2021	4	Balears, Illes	7	Balears, Illes		07037 Mercadal, Es	127	5	15

Figura 15: Filtrado de excel por Baleares

Como estamos más acostumbrados a leer CSVs desde un *jupyter notebook*, abrimos uno para extraer la información que necesitamos del excel y después pasar el código resultante al *script* con los *spiders*.

```

In [3]: import pandas as pd

In [11]: df = pd.read_excel("data/Paro_por_municipios_segundo_semestre_2021.xls", header=1)

In [17]: df_balears = df[df["Provincia"]=="Balears, Illes"]

In [25]: last_date = df_balears["mes"].unique()[0]
last_date

Out[25]: 'Septiembre de 2021'

In [31]: df_balears_last_date = df_balears[df_balears["mes"]==last_date]
df_balears_last_date.to_dict("records")

Out[31]: [{"Código mes": 202109,
'mes': 'Septiembre de 2021',
'Código de CA': 4,
'Comunidad Autónoma': 'Balears, Illes',
'Código Provincia': 7,
'Provincia': 'Balears, Illes',
'Código Municipio': 7001,
'Municipio': 'Maó',
'total Paro Registrado': 247,
'Paro hombre edad < 25': 16,
'Paro hombre edad 25 -45': 40,
'Paro hombre edad >=45': 63,
'Paro mujer edad < 25': 5,
'Paro mujer edad 25 -45': 56,
'Paro mujer edad >=45': 67,
'Paro Agricultura': 4,
'Paro Industria': 12,
'Paro Construcción': 31,
'Paro Servicios': 182,
}

```

Figura 16: Hoja *jupyter* para hacer las pruebas

Con el script, almacenamos la información en la base de datos NoSQL, en la colección "*unemployment*".

Key	Value	Type
(1) ObjectId("61804640b24a88110d5cf019")	{ 21 fields }	Object
_id	ObjectId("61804640b24a88110d5cf019")	ObjectId
Código mes	202109	Int32
mes	Septiembre de 2021	String
Código de CA	4	Int32
Comunidad Autónoma	Balears, Illes	String
Código Provincia	7	Int32
Provincia	Balears, Illes	String
Código Municipio	7001	Int32
Municipio	Alaró	String
total Paro Registrado	247	Int32
Paro hombre edad < 25	16	Int32
Paro hombre edad 25 -45	40	Int32
Paro hombre edad >=45	63	Int32
Paro mujer edad < 25	5	Int32
Paro mujer edad 25 -45	56	Int32
Paro mujer edad >=45	67	Int32
Paro Agricultura	4	Int32
Paro Industria	12	Int32
Paro Construcción	31	Int32
Paro Servicios	182	Int32
Paro Sin empleo Anterior	18	Int32
(2) ObjectId("61804640b24a88110d5cf01a")	{ 21 fields }	Object
_id	ObjectId("61804640b24a88110d5cf01a")	ObjectId
Código mes	202109	Int32
mes	Septiembre de 2021	String
Código de CA	4	Int32
Comunidad Autónoma	Balears, Illes	String
Código Provincia	7	Int32
Provincia	Balears, Illes	String
Código Municipio	7002	Int32
Municipio	Alaior	String
total Paro Registrado	307	Int32
Paro hombre edad < 25	19	Int32
Paro hombre edad 25 -45	49	Int32
Paro hombre edad >=45	64	Int32
Paro mujer edad < 25	13	Int32
Paro mujer edad 25 -45	72	Int32
Paro mujer edad >=45	90	Int32
Paro Agricultura	5	Int32
Paro Industria	19	Int32
Paro Construcción	31	Int32
Paro Servicios	244	Int32
Paro Sin empleo Anterior	8	Int32
(3) ObjectId("61804640b24a88110d5cf01b")	{ 21 fields }	Object

Figura 17: Colección unemployment

6. Agradecimientos

Queremos agradecer la elaboración de este proyecto a nuestras familias, por entendernos y apoyarnos en la lucha para ser mejores profesionales. Por otro lado, nos gustaría agradecer a las personas encargadas de elaborar la diferente documentación que hemos ido consultando durante este proyecto, sin ellos hubiese sido mucho más complicado. Por último, agradecer también a la web pisos.com, por dejarnos acceder a sus anuncios y conseguir, así, los datos para la realización del proyecto.

7. Inspiración

El interés en analizar este conjunto de datos es poder analizar un sector de interés público, en el cual se puedan identificar posibles oportunidades de inversión o, incluso, de estafa. Se debe destacar que el sector inmobiliario es un sector muy cambiante e influenciado, por lo que como posible mejora sería interesante tener un histórico de todos los inmuebles que se han puesto a la venta a lo largo del tiempo. Además, también sería interesante hacer el estudio a nivel estatal y no solo de Mallorca.

Por otro lado, en este análisis, se pretende responder a las siguientes cuestiones:

- ¿Cuáles son las características de las viviendas más caras? ¿Y de las más baratas?
- ¿Cuáles son las viviendas que, por sus características, están por debajo del precio medio del mercado? ¿Y las que están por encima?
- ¿Cuál es la zona más barata donde comprar una vivienda? ¿Y la más cara?
- ¿Cuáles son las variables estacionales de nuestro dataset?
- ¿Son los precios de los inmuebles más caros si el anunciante es una inmobiliaria?
- ¿Podemos estimar el precio de la vivienda con los datos obtenidos?
- ¿El paro del municipio es una variable que afecta al precio medio por municipio de los inmuebles?

8. Licencia

Para nuestro conjunto de datos, escogeremos la licencia *CC BY-SA 4.0 License*, ya que permite proveerse del nombre del creador del *dataset* generado, indicando los posibles cambios realizados. Así, permite reconocer el trabajo de terceros.

Por otro lado, esta licencia permite su uso comercial, es decir, permite que diferentes empresas hagan uso de los datos generados, generando así nuevos proyectos a partir del original.

Por último, hay que tener en cuenta que las nuevas contribuciones deben realizarse también sobre dicha licencia. Esto permite que los términos que fueron planteados por el autor se respeten en los nuevos proyectos

9. Código

Para la implementación del *scraper* hemos empleado las siguiente tecnologías:

- Docker compose
- Docker
- MongoDB
- Python con la librería Scrapy
- Para la documentación:
 - Markdown
 - Latex

Los pasos para la ejecución del proyecto están explicados en el archivo README.md dentro de la carpeta con el código.

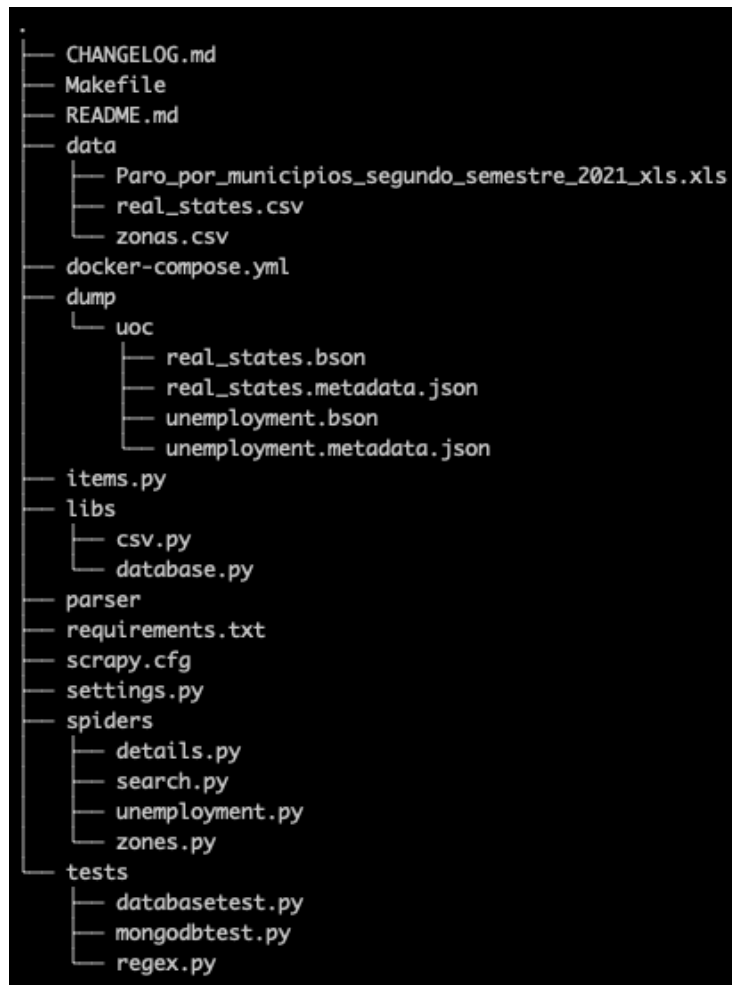


Figura 18: Estructura de directorios del proyecto

10. *Dataset*

Es una representación en csv de la información volcada en la base de datos NoSQL.

Referencias

- [1] Calvo, M., Pérez, D., Subirats, L. (2019). *Introducción al ciclo de vida de los datos*. Editorial Universitat Oberta de Catalunya.
- [2] Subirats, L., Calvo, M. (2019). *Web Scraping* Editorial Universitat Oberta de Catalunya.
- [3] Minguillón, J. (2016). *Fundamentos de data Science*. Editorial Universitat Oberta de Catalunya.