Tipología y ciclo de vida de los datos

PRA1: Web Scraping

Creación de un bot para analizar pisos de alquiler en Fotocasa

---

Autores:

Javier Guimerans Alonso

y

Gerson Villalba Arana

Abril 2022

Índice

1. Contexto

Los datos que se pretenden recolectar contienen información sobre los alquileres de viviendas en distintas ciudades españolas. Por lo tanto, se plantean diferentes alternativas de webs que puedan proporcionar dicha información. Hay dos portales principales en el mercado español de viviendas que pueden proporcionar los datos:

- **Idealista**
- **Fotocasa**

Para la recoleción de datos, se elige el segundo de ellos porque dispone de unas restricciones menores a la hora de hacer *web scraping*. Existen más portales, pero el resto tienen una cantidad de viviendas sustancialmente menor que los dos mencionados, y por ello se descartan.

En el momento de realización de este informe, Fotocasa da como resultado un total de 44.341 viviendas en alquiler en toda España, por lo que el potencial de recolección de datos a través de esta web es muy alto.

El objetivo del proyecto que se propone es el de poder obtener los datos deseados de una ciudad determinada o un conjunto de ellas, y poder generar un dataset con ellos, que pueda servir como punto a partida para multitud de aplicaciones. Para ejemplificar una de estas aplicaciones, se realiza un análisis visual de los datos mediante diferentes representaciones gráficas.

2. Título

El título elegido para el dataset es: "Datos de alquileres de viviendas en España".

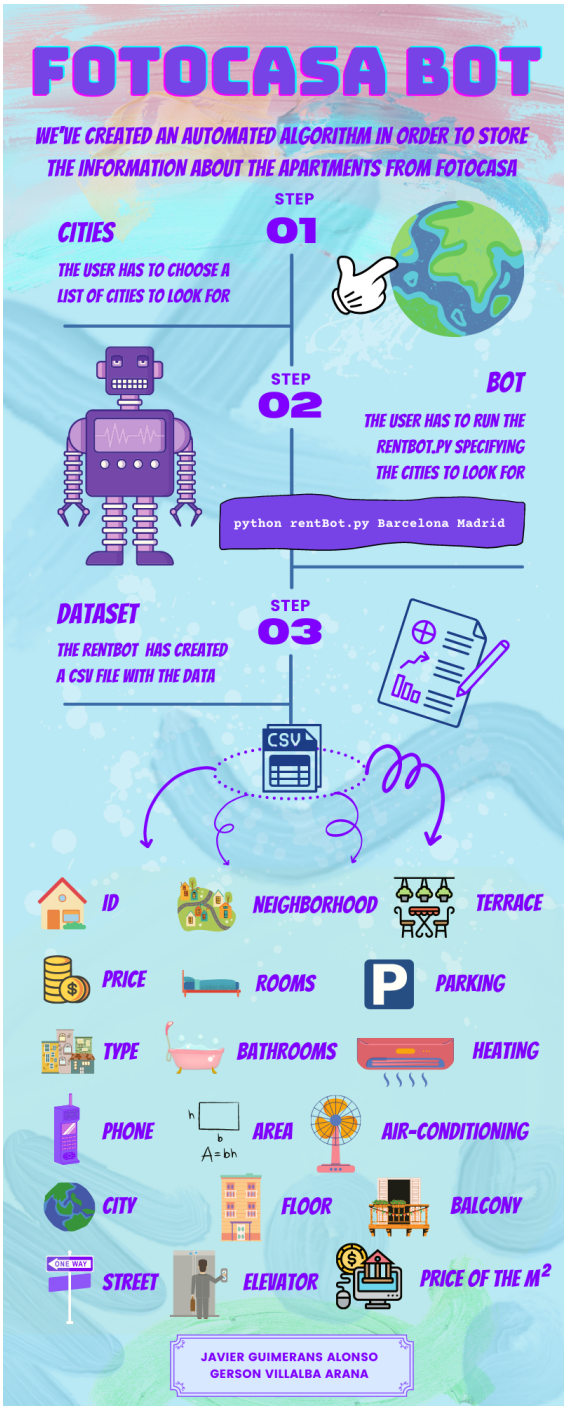
3. Descripción del dataset

El dataset recolectado a partir del *web scraping* sobre la web de Fotocasa contiene las características disponibles en el anuncio sobre la vivienda en cuestión ofertada, además del precio mensual del alquiler y número de teléfono del anunciante.

4. Representación gráfica

La Figura 1 muestra una representación gráfica del funcionamiento del script, así como sus intrucciones de uso y el contenido del dataset que genera.

Figura 1: Representación gráfica del funcionamiento del script



Las Figuras 2 - 6 muestran varias representaciones gráficas de diferentes atributos del dataset obtenido en una ejecución del script. En concreto, se ha ejecutado el script para que recolecte información sobre los alquileres de viviendas en Barcelona, Madrid, Valencia, Mallorca y Granada.

Figura 2: Tipo de vivienda

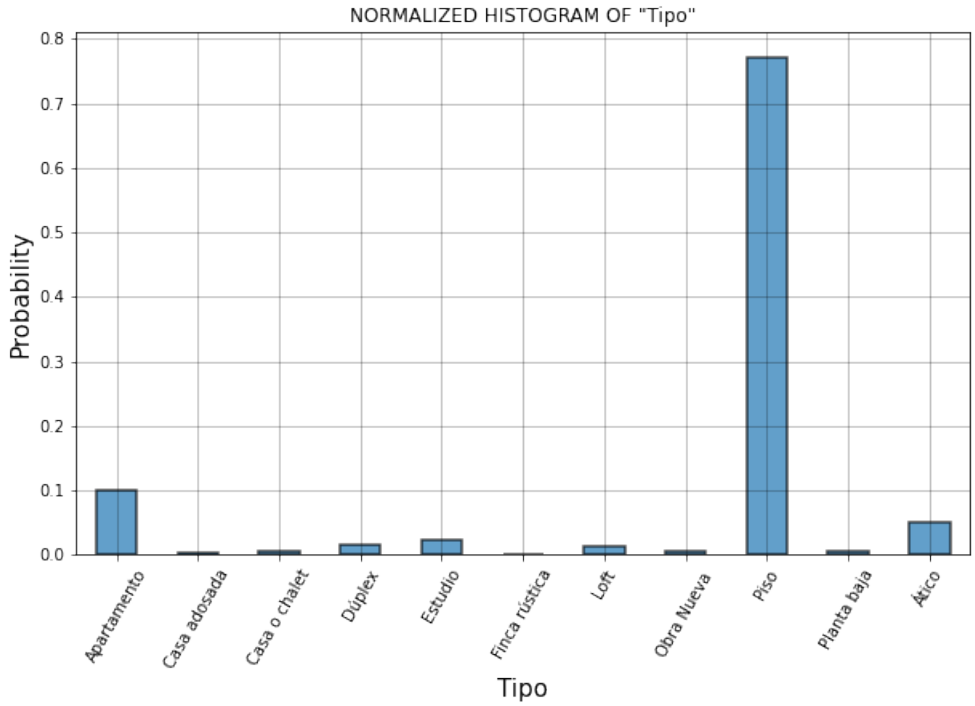


Figura 3: Precio de alquiler por ciudad

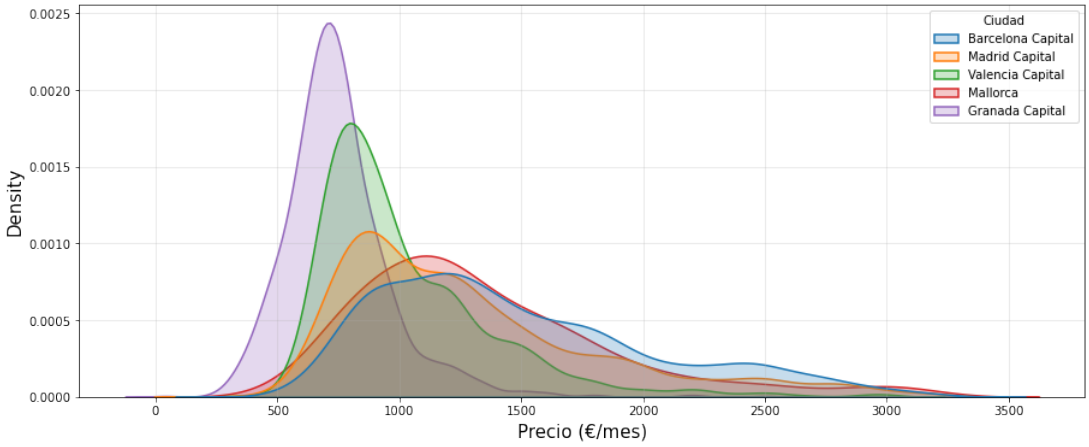


Figura 4: Superficie de la vivienda en diferentes ciudades

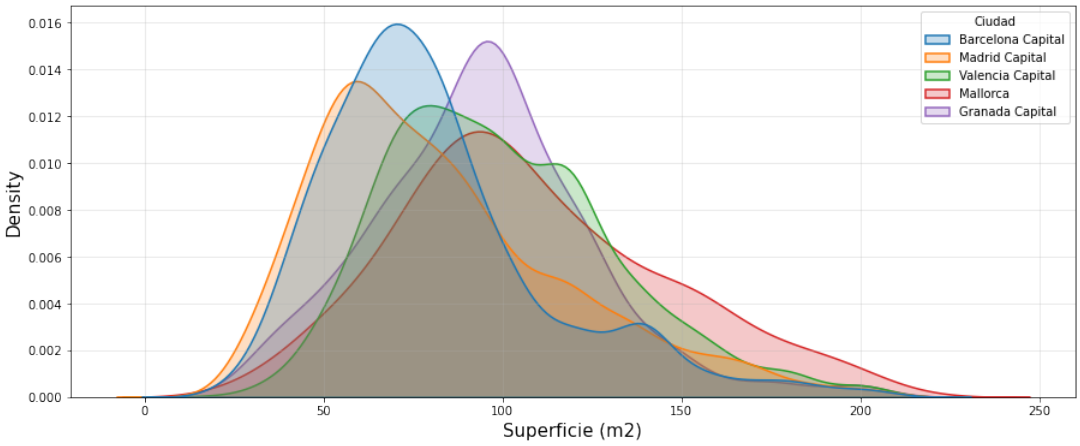


Figura 5: Precio del alquiler por metro cuadrado en distintas ciudades

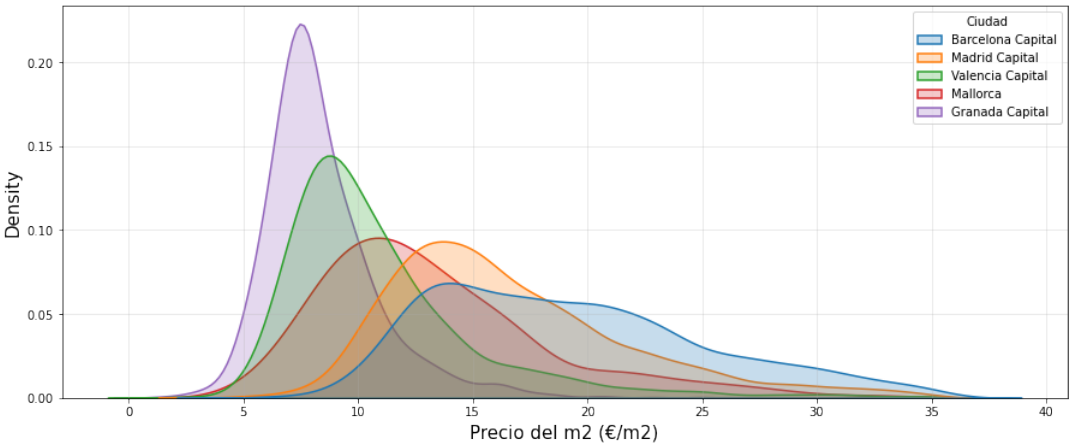
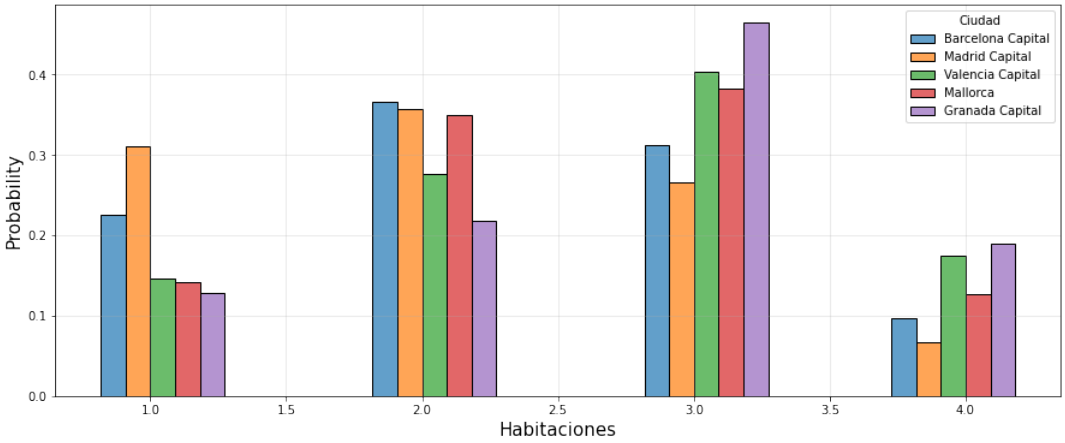


Figura 6: Número de habitaciones en diferentes ciudades



5. Contenido

El dataset obtenido a partir del *web scraping* tiene los siguientes campos:

- **ID:** Número identificativo de la vivienda.
- **Precio (€/mes):** Precio de la vivienda en euros mensuales.
- **Tipo:**. Tipo de vivienda.
- **Teléfono:**. Número de teléfono del anunciante.
- **Ciudad:** Ciudad donde se encuentra el inmueble.
- **Dirección:** Calle donde se encuentra el inmueble.
- **Barrio:** Barrio donde se encuentra la vivienda.
- **Habitaciones:** Número de habitaciones.
- **Baños:** Número de baños.
- **Superficie (m2):** Superficie de la vivienda en metros cuadrados.
- **Planta:** Planta donde se encuentra la vivienda.
- **Ascensor:** La vivienda dispone de ascensor.
- **Terraza:** La vivienda dispone de terraza.
- **Parking:** La vivienda dispone de parking.
- **Calefacción:** La vivienda dispone de calefacción.
- **Aire:** La vivienda dispone de aire acondicionado.
- **Balcón:** La vivienda dispone de balcón.
- **Precio del m2 (€/m2):** Relación entre el precio y la superficie de la vivienda en euros por metro cuadrado.

Todos los datos se han obtenido a través del resultado obtenido en la web al buscar viviendas en alquiler en un municipio en concreto. Los resultados aparecen en una página que utiliza Javascript, por lo que se ha utilizado la librería Selenium para poder obtener los resultados completos de una página, ya que estos se van cargando de forma dinámica. Por otro lado, los resultados, siempre que éstos superen un número determinado, aparecen paginados, por lo que se ha tenido que extrar el número total de páginas presentes en el resultado y navegar por todas para recuperar el total de resultados.

Hay que dejar claro que los datos obtenidos han sido obtenidos exclusivamente a partir de la página de resultados, sin entrar en la propia página dedicada a una vivienda en concreto. Esto quiere decir que potencialmente existiría la opción de poder una mayor cantidad de información que la recolectada en este trabajo. Hay que tener en cuenta que los datos obtenidos tienen que estar definidos como atributos del piso por parte del anunciante y éstos aparecer en la página principal de resultados. Si estas dos condiciones no se cumplen, la característica de la vivienda no será capturada correctamente. En la Figura 7 podemos ver un ejemplo de vivienda en la web de resultados, donde hemos remarcado los campos que extraemos de él.

Figura 7: Ejemplo de anuncio de alquiler

Habría dos métodos para mejorar la calidad de los datos obtenidos:

- Entrando en el anuncio propio de cada vivienda y realizando *web scraping* sobre ésta, de forma que se pudiesen capturar correctamente todas y cada una de las características definidas por el anunciante de la vivienda.
- Además del punto anterior, realizar un análisis del texto escrito de descripción de la vivienda, para obtener nuevas características, no definidas por el anunciante como atributos de la vivienda, pero que se encuentran en el texto de descripción. Para ello, haríamos uso de herramientas de *text minning*.

Ambos métodos quedan fuera del alcance de este trabajo por su complejidad añadida al tener que cargar miles de webs, siempre teniendo cuidado de que el servidor web no nos bloquee por uso excesivo de tráfico. Por ello, nos quedamos con los atributos que podemos extraer desde la página principal de resultados.

Al tratarse de datos sobre viviendas en alquiler, los datos tienen un período de vigencia relativamente corto. Dependiendo de la aplicación, éste puede ser mayor o menor, pero podría ir desde un día a unos pocos meses. Si se quieren mantener los datos actualizados, se recomienda ejecutar el script de recolección de datos una vez al día para las ciudades deseadas. Tras una nueva ejecución, se obtendrá un nuevo dataset.

Si se desea mantener un histórico de los datos, simplemente se podrá hacer una unión de ambos datasets, eliminando los registros con ID de vivienda duplicados. En este caso, también sería interesante añadir a los datos una columna de fecha, para mantener registro de la fecha de publicación del anuncio.

6. Agradecimientos

Los datos han sido recopilados de la web https://www.fotocasa.es/es/, y, por lo tanto, debemos agradecer a esta plataforma por la información pública disponible que se ha podido recopilar para la creación de este trabajo.

Para no saturar el servidor web con peticiones, se han incluido generosos tiempos de espera en el proceso de *web scraping*. Tomando dichas precauciones, el servidor no nos ha rechazado las peticiones en ningún momento.

Además, se han respetado las reglas establecidas en el archivo *robots.txt* del sitio web, ya que se ha accedido en todo momento a las rutas */es/alquiler/**, no protegidas con ninguna regla especial frente a web scrapping.

Como no somos los propietarios de los datos, además, se han distribuido los datos con licencia que no permite el uso comercial de éstos, para evitar cualquier posible conflicto legal.

7. Inspiración

El mercado inmobiliario es uno de los que mayor peso tienen en la economía. A pesar de que, tradicionalmente, España es un país más centrado en la compra/venta de vivienda que en el alquiler, éste último ha cobrado cada vez más protagonismo. Según [1] en España hay 3.4 millones de viviendas en alquiler, por lo que el mercado es realmente grande.

La recopilación por lo tanto de datos de un mercado tan grande puede tener grandes aplicaciones. Algunas de estas pueden ser las siguientes:

- Obtención y filtrado de viviendas en alquiler según distintos parámetros.
- Realizar un análisis estadístico descriptivo sobre los precios de alquiler en distintas ciudades españolas.
- Análisis visual del mercado de alquileres en España.
- Búsqueda de correlaciones entre variables dentro de las disponibles.
- Predicción de precios de viviendas según sus características (regresión) con distintos algoritmos de *machine learning*.
- Predicción del tiempo que tardará una vivienda en ser alquilada según sus características y precio de alquiler. También sería un problema supervisado de regresión. Para esta tarea necesitaríamos, no sólamente unos datos puntuales, sino realizar una monitorización de las viviendas para contar con un histórico, de forma que podamos saber cuándo se ha puesto en alquiler la vivienda, y cuándo se ha retirado.
- Agrupamiento de los pisos según caraterísticas similares para, por ejemplo, recomendar a un cliente viviendas similares a la que está buscando. Para ello se podría utilizar un algoritmo no supervisado de *clustering*.

8. Licencia

El dataset se publica bajo una licencia CC BY-NC-SA 4.0 [2]. Esta elección se realiza porque las cláusulas de esta licencia se corresponden con las deseadas para este caso concreto:

- El usuario del dataset debe dar crédito al creador de éste, reconociendo así su trabajo.
- Las contribuciones que se puedan realizar a posteriori sobre el trabajo deben publicarse también bajo la misma licencia original.
- No se permite un uso comercial del dataset. Teniendo en cuenta que el dataset ha sido obtenido de una web con fines comerciales (https://www.fotocasa.es/es/), de esta forma evitamos problemas legales con la distribución de esta información con fines comerciales.

Por otro lado, el código fuente del script Python utilizado para obtener los datos mediante *web scraping* se publica bajo licencia "MIT License".

9. Código

El código fuente completo que se ha utilizado para la obtención del dataset se encuentra en los siguientes repositorios:

https://github.com/JavierGuimerans/rentBot
https://github.com/gvillalba86/rentBot

Además del código fuente del script para la obtención del dataset, en los mismos repositorios también se encuentra un Jupyter Notebook en el que se ha realizado un análisis visual de los datos obtenidos en forma de diferentes gráficos.

10. Dataset

El dataset puede encontrarse en el mismo repositorio que el código fuente. Por otro lado, se ha publicado también en Zenodo en el siguiente enlace:

Zenodo

11. Contribuciones

En la Cuadro 1 se muestran las contribuciones que ha realizado cada miembro del equipo al desarrollo del presentre trabajo.

Cuadro 1: Tabla de contribuciones

Contribuciones	Firma
Investigación previa	JGA, GVA
Redacción de las respuestas	JGA, GVA
Desarrollo del código	JGA, GVA

Referencias

[1] INE. *Encuesta continua de hogares. Año 2020.* URL: https://www.ine.es/prensa/ech_2020.pdf.

[2] Creative Commons. *Atribución-NoComercial-CompartirIgual 4.0 Internacional (CC BY-NC-SA 4.0).* URL: https://creativecommons.org/licenses/by-nc-sa/4.0/deed.es.