**Definir un títol pel data set. Triar un títol que sigui descriptiu.**

Preu i característiques dels habitatges de lloguer a les principals ciutats catalanes.

**Descripció del data set. Desenvolupar una descripció breu del conjunt de dades que s’ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat)**

El data set recull informació sobre el preu i les característiques dels habitatges de lloguer anunciats al portal immobiliari [www.tucasa.com](http://www.tucasa.com) a les ciutats de Girona, Tarragona, Lleida i Barcelona.

**Inspiració. Explicar perquè és interessant aquest conjunt de dades i quines preguntes es pretenen respondre**

En els darrers anys els preus de l’habitatge de lloguer a Catalunya han augmentat a un ritme molt més accelerat que els salaris de la població. Aquest fet augmenta la vulnerabilitat de moltes famílies, que veuen com els preus de lloguer queden lluny o fora de l’abast del seu poder adquisitiu, i posen en risc el seu dret fonamental d’accés a l’habitatge [1]. En aquest context, el preu del lloguer pren una rellevància especial, ja que pot representar un indicador de risc i exclusió social.

El conjunt de dades que presentem conté, a més del preu de lloguer, informació sobre la localització geogràfica de cada pis (Girona, Tarragona, Lleida, Barcelona; i barri dins d’aquestes ciutats) i sobre les seves característiques bàsiques (nombre d’habitacions, nombre de banys, m2). Així doncs, aquestes dades permeten estudiar com afecten la zona geogràfica i les característiques estructurals dels pisos al seu preu de lloguer.

En resum, les principals preguntes que pot respondre aquest conjunt de dades són les següents:

* Quin és l’estat del preu de l’habitatge?
* Com varia el preu de lloguer segons zona geogràfica?
* Quines característiques estructurals dels pisos determinen el preu del lloguer?

Més enllà d’aquest data set, el web scraper desenvolupat és una eina pot servir per actualitzar les dades en el temps i fer un seguiment de l’evolució dels preus del lloguer.

**Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació.**

El nostre objectiu és crear un dataset que contingui informació dels pisos de lloguer de les 4 capitals de província de Catalunya. El lloc a cercar aquesta informació doncs, seran els diferents portals de compra/lloguers de pisos de Catalunya i Espanya.

Hem fet una recerca sobre 8 portals de lloguers de pisos per escollir el que més fàcilment ens proporcionarà les dades que volem. Aquests portals són: fotocasa, idealista, habitaclia, API.cat, buscocasa, pisos.com, yaencontre i tucasa.

A continuació mostrem les conclusions d’aquesta recerca i el portat escollit per recol·lectar la informació:

- **Fotocasa** i **Idealista:**  són els primers en descartar ja que retornen un codi 405 (mètode no permès) i 403 (prohibit) respectivament. Ja que tenim alternatives decidim descartar aquests portals. Els següents tots retornen un codi 200 (Ok).

**- Buscocasa:** s’ha descartat perquè hi ha poques ofertes i totes de fa més d’un mes. Tenim alternatives amb un volum de dades més gran i més actualitzat.

**- API.cat:** s’ha descartat perquè aporta poca informació dels pisos de cada anunci. No aporta informació de número d’habitacions o número de lavabos cosa que altres portals sí que fan.

- **Habitaclia:** ara ja comencem a trobar portals molt complets en informació i en volum de pisos. Segurament a partir d’aquí tots ens servirien per crear un bon dataset però, ja que podem escollir, triarem el portal que presenti la informació de la manera més còmode de recol·lectar. En aquest cas hem descartat aquest portal perquè la informació del pis (metres quadrats, habitacions, lavabos...) no està diferenciada per etiquetes sinó que és una sola cadena de caràcters en la mateixa etiqueta. Com hem dit no impossibilita la recol·lecció de informació però tenim alternatives que ens deixaran un codi més clar i net.

**- Pisos.com:** també té la informació molt completa i en aquest cas sí que la diferent informació està separada en diferents etiquetes. S’ha descartat, però, perquè totes les classes d’aquestes etiquetes es diuen igual (class=item) i tenim alternatives que ho tenen més ben classificat.

**- Yaencontre i Tucasa:** finalment arribem als dos últims portals. Ens presenten informació completa i un número de pisos molt elevat. A més la informació que volem està molt ben classificada per etiquetes úniques. Així doncs tenim dues opcions molt vàlides. Decidm quedar-nos amb **Tucasa** ja que en aquest portal els títols dels anuncis són els noms dels barris d’una forma més estandarditzada que Yaencontre i permet fer cerques a nivell de província.

**Contingut. Explicar els camps que inclou el data set, el període de temps de les dades i com s’ha recollit**

Els camps que inclou el data set són:

**- ciutat:** Ciutat on es troba l’habitatge de lloguer. Com que hem decidit fer les capitals de província de Catalunya pot prendre els valors Barcelona, Girona, Lleida o Tarragona.

**- pàgina:** La pàgina on es trobava l’anunci en el moment de fer l’scraping. Com que hi ha 20 anuncis per pàgina, cada 20 anuncis canvia de nom.

**- zona:** Zona de la ciutat on es troba l’habitatge, generalment nom del barri.

**- carrer:** Nom del carrer on es troba l’habitatge.

**- preu:** Preu mensual a pagar pel lloguer de l’habitatge.

**- m2:** Metres quadrats de l’habitatge

- **preu\_m2:** Preu per metre quadrat de l’habitatge.

**- nHab:** Nombre d’habitacions de l’habitatge

**- nBanys:** Nombre de banys de l’habitatge.

Per recollir aquestes dades hem plantejat un bucle sobre cada capital de província, dins aquest, un bucle sobre cada pàgina d’anuncis i, dins d’aquest, un bucle sobre cada anunci d’habitatge de lloguer. A partir d’aquí recollim tota la informació que ens interessa de cada anunci (per cada pàgina i cada capital de província).

**Llicència. Seleccionar una d’aquestes llicències pel data set resultant i explicar el motiu de la seva selecció**

Per seleccionar una llicència hem tingut en compte que les dades que hi ha al nostre data set no son nostres sinó de l’empresa que les facilita. Per això ens hem decantat cap a la llicència **CC BY-NC-SA 4.0 License** ja que implica donar crèdit a l’origen de dades (en aquest cas [www.tucasa.com](http://www.tucasa.com)) i a més en limita l’ús comercial.

**Enllaç del data set a Zenodo**

**Rental housing market in Catalonia in April 2020**

This data set includes a snapshot of the state of the rental housing market in Catalonia (Spain) on the XXth April 2020. The data was obtained using a web scraping tool on the real estate website [www.tucasa.com](http://www.tucasa.com). The web scraper collected data on price, square metres, price per square metre, number of rooms, number of bathrooms, neighbourhood and city; in the four main cities of Catalonia (Barcelona, Girona, Lleida and Tarragona). The web scraping code can be found in the following git repository: [www.github.com/scraper-preus-lloguer](http://www.github.com/scraper-preus-lloguer)

**DOI github**

**Codi. Adjuntar el codi amb el qual s’ha generat el data set, preferiblement en python o, alternativament, en R.**

**Agraïments. Presentar el propietari del conjunt de dades.**

Agraïm a l’empresa propietària de les dades tucasa.com, pàgina d’anuncis de compra i lloguers d’habitatges (i d’espais i terrenys en general) de tota Espanya.

**Es necessari incloure cites de recerca o anàlisi anteriors (si n’hi ha).**

[1] <https://www.ccma.cat/324/expulsats-de-barcelona/>

**Representació gràfica. Presentar una imatge o esquema que identifiqui el  dataset visualment**

