

# TIPOLOGIA I CICLE DE VIDA DE LES DADES

## Pràctica 1

### Índex

0. Components .....	2
1. Context. Explicar en quin context s'ha recol·lectat la informació. Explicar per què el lloc web triat proporciona aquesta informació. ....	2
2. Definir un títol pel dataset. Triar un títol que sigui descriptiu. ....	3
3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat). ....	4
4. Representació gràfica. Presentar una imatge o esquema que identifiqui el dataset visualment .....	4
5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit. ....	5
6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha). ....	6
7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre. ....	6
8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció.....	6
9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.....	7
10. Dataset. Presentar el dataset en format CSV .....	7

## 0. Components

La totalitat d'aquesta pràctica ha estat realitzada pel dos firmants que consten al document:

- Joan Ribera Monfort (JRM)

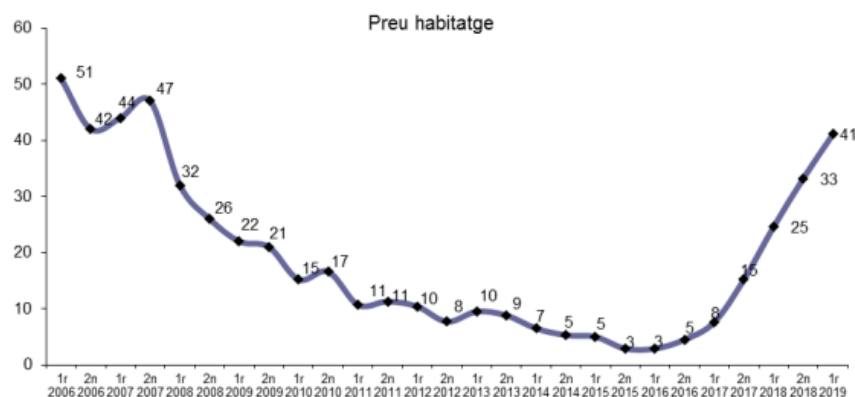
- Jordi Fuster Arion (JFA)

JFA

## 1. Context. Explicar en quin context s'ha recollert la informació. Explicar per què el lloc web triat proporciona aquesta informació.

En els darrers anys a Andorra s'ha experimentat un creixement molt important dels preus dels pisos de lloguer degut a una disminució de l'oferta i un increment de la demanda.

Segons reflecteix l'enquesta de l'Observatori<sup>1</sup> del Centre de Recerca i Estudis Sociològics de l'Institut d'Estudis Andorrans del 11 de juliol de 2019, des del primer semestre de 2018 el preu de l'habitatge es situa com a primera preocupació pel 41,1% de la població andorrana. Com s'aprecia al gràfic que hi ha a continuació, aquesta preocupació ha crescut molt des del 2017.

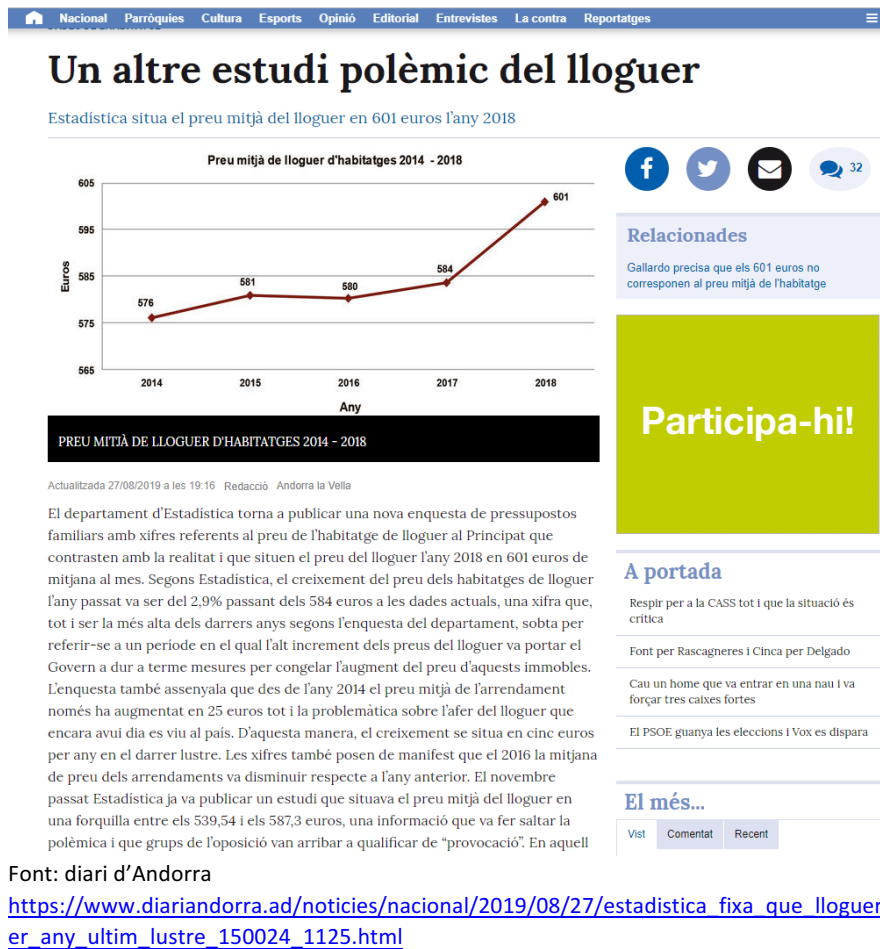


Font: Centre de Recerca i Estudis Sociològics (IEA)

Aquest creixement de la preocupació no es veu reflectit a les dades que publica el Departament d'Estadística d'Andorra amb les dades de l'Enquesta de pressupostos familiars (EPF). Aquesta enquesta recull les dades que paguen de mitjana la població, tenint en compte que gran part de la població disposa de contractes antics que disminueixen aquesta mitjana. Els problemes apareixen a la població que ha de realitzar un nou contracte o una renovació de l'antic. Per

<sup>1</sup> [https://www.iea.ad/images/cres/Observatori/2019/2019\\_Observatori\\_1erSem.pdf](https://www.iea.ad/images/cres/Observatori/2019/2019_Observatori_1erSem.pdf)

aquests motius, quan es publiquen les dades de l'EPF hi ha fortes crítiques per part de la població i de partits polítics, ja que segons ells es impossible disposar d'un pis de lloguer als preus que marca l'estudi.



Per a donar resposta a la inquietud i contrastar les dades del Departament d'Estadística, s'han extret les dades del portal buscocasa.ad (major portal immobiliària d'Andorra) amb Python.

Aquest portal immobiliària recull les dades d'habitatges, despatxos, naus industrials, aparcaments, etc., però en aquest estudi només s'analitzaran els habitatges en lloguer. Per fer-ho, degut que s'ha vist que hi ha anuncis mal catalogats (immobles que no són habitatges catalogats a les categories de vivendes, s'ha filtrat en la pròpia URL, forçant que disposi d'un bany com a mínim.

## 2. Definir un títol pel dataset. Triar un títol que sigui descriptiu.

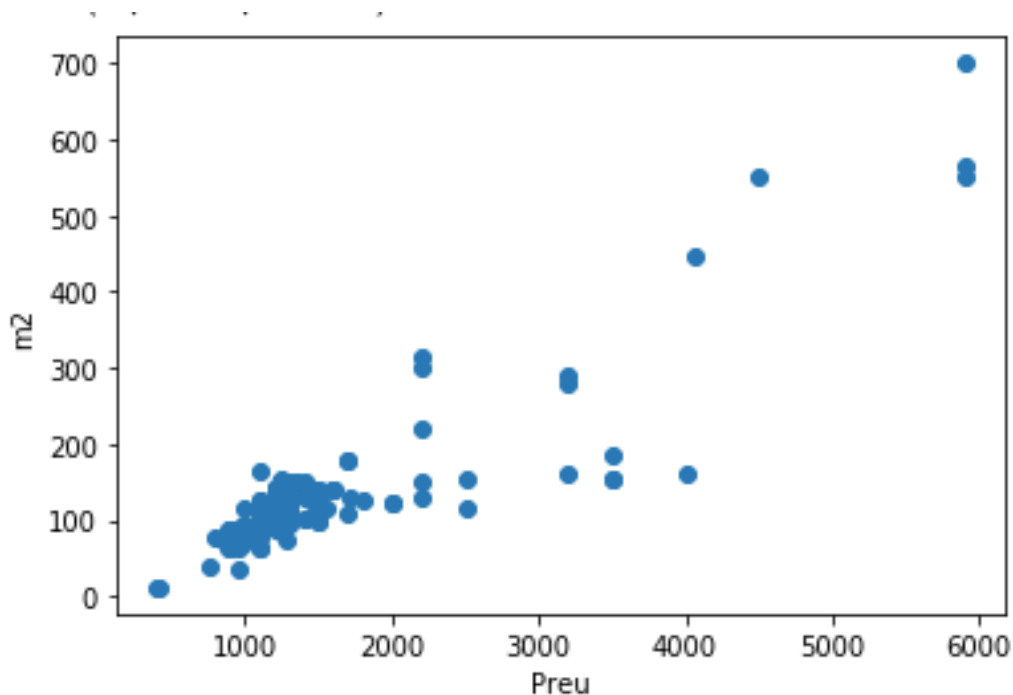
anuncis\_habitatges\_andorra.csv

3. Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret (és necessari que aquesta descripció tingui sentit amb el títol triat).

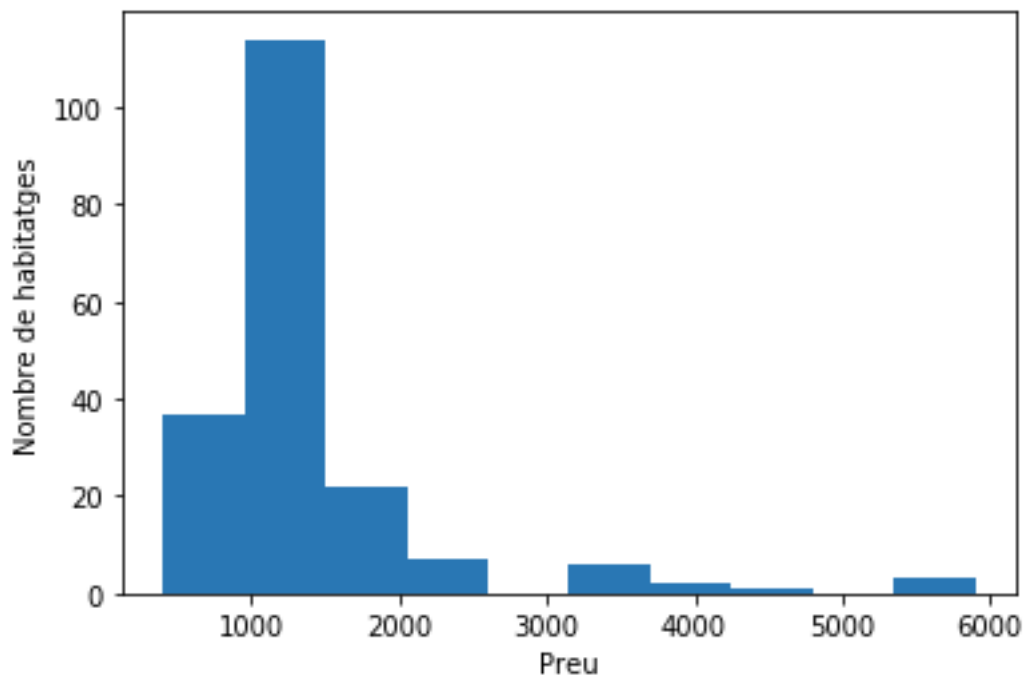
El dataset inclou tots els anuncis publicats al portal immobiliari, a dia 10/11/2019 (darrera execució). Inclou també les principals variables d'interès (metres quadrats, preu, tipus d'habitatge, parròquia, etc.). S'ha contemplat la idea d'anar emmagatzemant les dades històricament, però finalment només es guarden les últimes dades extretes.

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

En el següent gràfic s'observa un scatterplot entre el preu i els metres quadrats.



A continuació s'observa que la majoria de pisos en lloguer es troben en la franja d'uns 1.000 euros.



5. Contingut. Explicar els camps que inclou el dataset, el període de temps de les dades i com s'ha recollit.

Les dades contingudes fan referència a tots els habitatges anunciats a la web a data 10/11/2019.

Les dades s'han recollit utilitzant tècniques de web scraping amb diferents llibreries especialitzades amb el programari Python. S'han recollit totes aquelles dades disponibles a la web que són considerades d'interès:

Variable	Tipus	Descripció
descripcio	Caràcter	Descripció completa de l'anunci.
tipusHabitatge	Caràcter	Tipologia d'habitatge (pis, casa, xalet, etc.)
poble	Caràcter	Poble on es troba l'habitatge
parroquia	Caràcter	Parròquia on es troba l'habitatge
dataAnunci	Data	Data de publicació de l'anunci
m2	Numèric	Metres quadrats que té el pis
numHabitacions	Numèric	Nombre d'habitacions de l'habitatge
tipusOferta	Caràcter	Tipus d'oferta de l'anunci: lloguer, propietat
numVisites	Numèric	Nombre de visites que ha rebut l'anunci
preu	Numèric	Preu al que s'anuncia l'habitatge.

## 6. Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites de recerca o anàlisis anteriors (si n'hi ha).

Les dades han estat extretes del portal d'anuncis de pisos d'Andorra [www.buscocasa.ad](http://www.buscocasa.ad), un dels portals més utilitzats del país. Buscocasa, propietat de l'empresa DAIKKIRI SL, és el portal de referència a Andorra, amb més de 5.000 propietats anunciades i utilitzat per més de 50 immobiliàries del país. Agrair a l'empresa DAIKKIRI SL i la seva infraestructura informàtica del portal [www.buscocasa.ad](http://www.buscocasa.ad). Al ser andorra un mercat potencialment petit, no s'ha vist cap altre projecte de web scraping realitzat en aquesta web, però sí en altres del mateix sector (veure recursos).

## 7. Inspiració. Explicar per què és interessant aquest conjunt de dades i quines preguntes es pretenen respondre.

Tal i com s'ha comentat en el context actual de la conjuntura andorrana referent a l'habitatge, hem cregut convenient realitzar aquest estudi, amb l'objectiu de mostrar la realitat que es troba la població alhora de buscar un habitatge de lloguer. A part, es podria fer servir per a realitzar estudis de mercat, per a escollir el preu just del lloguer a un immoble, per a facilitar les mesures per a limitar el preu de l'habitatge (per parròquia, tipologia, etc.), o a fer prediccions a futur.

Amb aquestes dades hi ha la possibilitat de realitzar un estudi sobre el preu mitjà dels habitatges de lloguer anunciats. Aquest estudi es pot segmentar per exemple per tipologia d'habitatge, zona geogràfica o metres quadrats.

- Quin és el preu del lloguer dels habitatges a Andorra?
- A quina parròquia són més cars?
- Quina és l'evolució del preu?
- Quin és el preu per metre quadrat?
- Quin és el preu per diferent tipus d'habitatge?
- Quants habitatges hi ha de lloguer?
- Les dades s'assemblen a les publicades pel Departament d'Estadística d'Andorra?
- Realitzant un model lineal, quin és el preu d'un habitatge de dues habitacions, de 60 metres quadrats situat a la parròquia de Canillo?

## 8. Llicència. Seleccionar una d'aquestes llicències pel dataset resultant i explicar el motiu de la seva selecció.

Les dades generades a partir del script s'alliberen sota la llicència CC BY-NC-SA 4.0, amb la qual cosa queden lliures per compartir el material, i adaptar-lo i crear sense finalitats comercials. A més, les noves creacions a partir d'aquesta hauran d'anar sota la mateixa llicència.

S'ha optat per aquesta llicència perquè tothom pugui fer ús del material (sempre reconeixent l'autoria i indicant canvis que s'hagin realitzat) i compartir-lo o adaptar-lo, però sense fins

comercials, ja que les condicions d'ús de la web especifiquen que es prohibeix l'ús de les dades amb fins comercials.


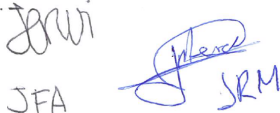
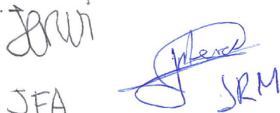
Val a dir, que per aquest projecte s'ha tingut en compte els termes i condicions de la web, així com el fitxer robots.txt de la web. També, les dades s'han extret de forma ètica, programant l'execució a hores de baix tràfic i amb intervals de temps entre peticions molt espaiat.

## 9. Codi. Adjuntar el codi amb el qual s'ha generat el dataset, preferiblement en Python o, alternativament, en R.

Veure GITHUB.

## 10. Dataset. Presentar el dataset en format CSV

Veure GITHUB.

Contribucions	Signa
Recerca prèvia	Jordi Fuster, Joan Ribera  JFA SRM
Redacció de les respostes	Jordi Fuster, Joan Ribera  JFA SRM
Desenvolupament codi	Jordi Fuster, Joan Ribera  JFA SRM

### RECURSOS:

Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd. Chapter 2. Scraping the Data.

Subirats, L., Calvo, M. (2018). Web Scraping.

<https://github.com/hmeleiro/idealisto>

<https://github.com/David-Carrasco/Scrapy-Idealista>