

# PRÀCTICA 1

## PRESENTACIÓ

En aquesta pràctica s'elabora un cas pràctic orientat a aprendre a identificar les dades rellevants per un projecte analític i usar les eines d'extracció de dades. Per fer aquesta pràctica haureu de treballar en grups de 2 persones. Haureu de lliurar un sol fitxer amb l'enllaç Github (<https://github.com>) on hi hagi les solucions incloent els noms dels components de l'equip. Podeu utilitzar la Wiki de Github per descriure el vostre equip i els diferents arxius del vostre lliurament. Cada membre de l'equip haurà de contribuir amb el seu usuari Github. Podeu revisar aquests exemples com a guia:

- Exemple: <https://github.com/rafoelhonrado/foodPriceScraper>
- Exemple complex: <https://github.com/tteguayco/Web-scraping>

## Competències

En aquesta PAC es desenvolupen les següents competències del Màster de Data Science:

- Capacitat d'analitzar un problema en el nivell d'abstracció adequat a cada situació i aplicar les habilitats i coneixements adquirits per resoldre-ho.
- Capacitat per aplicar les tècniques específiques de web scraping.

## Objectius

Els objectius concrets d'aquesta pràctica són:

- Aprendre a aplicar els coneixements adquirits i la seva capacitat de resolució
- de problemes en entorns nous o poc coneguts dins de contextos més amplis o multidisciplinaris.
- Saber identificar les dades rellevants que el seu tractament aporta valor a una empresa i la identificació de nous projectes analítics.
- Saber identificar les dades rellevants per dur a terme un projecte analític.
- Capturar dades de diferents fonts de dades (tals com a xarxes socials, web de dades o repositoris) i mitjançant diferents mecanismes (tals com queries, API i scraping).
- Actuar amb els principis ètics i legals relacionats amb la manipulació de dades en funció de l'àmbit d'aplicació.
- Desenvolupar la capacitat de cerca, gestió i ús d'informació i recursos en l'àmbit de la ciència de dades.

## Descripció de la Pràctica a realitzar

**L'objectiu d'aquesta activitat serà la creació d'un dataset a partir de les dades contingudes en una web. Per a la seva realització, s'han de complir els següents punts:**

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

Avui dia, cada cop és més difícil accedir a un habitatge a causa dels seus preus, la poca estabilitat laboral i la dificultat per a obtenir una hipoteca. Degut a això, cada cop és més important ajustar el preu d'aquestes a les necessitats, tenint en compte que és imprescindible i que es pot pagar per a les coses que no. En aquest *dataset* es recullen totes les cases en

venda en Catalunya en el portal d'anuncis [Habitacalia](#) podent portar a terme les tasques comentades anteriorment, entre altres.

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

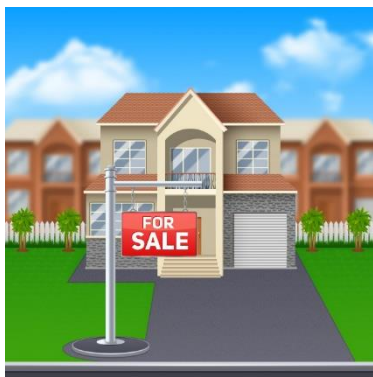
El títol triat és habitatges en venda en Catalunya, amb concordança amb el que s'ha explicat anteriorment.

**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* presenta un recull de tots els habitatges de Catalunya que hi ha en venda en el portal d'Habitacalia.

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

Una imatge que pot representar i que identifiqui el *dataset* presentat pot ser la següent:



Aquesta imatge no és pròpia, s'ha extret del lloc web freepik.es

Aquesta imatge representa una casa en venda i, per tant, té una relació directa amb el significat del *dataset*.

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

Els camps que inclou el *dataset* són els següents:

- TITLE: Títol de l'anunci.
- TOWN: Poble on està situada.
- PROVINCE: Província on està situada.
- REGION: Comarca on està situada.
- SURFACE: Metres quadrats totals.
- N\_ROOMS: Nombre d'habitacions.
- N\_BATHROOMS: Nombre de banys.
- PRICE\_METER: Preu per metre quadrat.
- TOTAL\_PRICE: Preu total.

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

Com ja s'ha comentat anteriorment, les dades han sigut obtingudes del portal web Habitacalia encarregat d'anunciar habitatges per compra i lloguer. Per la generació d'aquest *dataset*,

només s'ha considerat el cas de compra. Per la recollida, s'ha utilitzat el llenguatge *Python* per aplicar tècniques de *web scrapping*.

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

Aquest conjunt de dades és interessant perquè recull tots els habitatges de Catalunya en venda dins del portal web Habitacalia, sent aquest un dels portals més grans actualment a Espanya. A causa d'això, permet fer un anàlisi del mercat immobiliari, podent examinar si està per sobre o per sota del seu preu real. També es pot fer servir per a veure on hi ha els preus més alts i més baixos de Catalunya, podent classificar les zones.

En anàlisis més avançats, una de les avantatges que pot proporcionar aquest conjunt de dades és per a predir quin és el preu que un nou habitatge pot valer depenent de les característiques anteriors.

Algunes de les preguntes que es pretenen contestar amb aquestes dades són:

- Preu mitjà de l'habitatge depenent de província, comarca o poble.
- Preu mitjà del metre quadrat depenent de província, comarca o poble.
- Quant és el cost de viure en una zona.
- Classificació de zones més cares i barates de Catalunya.
- Cost afegit per nombre d'habitacions o banys.

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

- 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

La llicència escollida per aquest *dataset* és la *CC BY-SA 4.0 License* pels següents motius:

- S'ha de nomenar a l'autor, així com tots els canvis que hagin sigut produïts: D'aquesta manera es reconeix el treball de l'autor i s'assegura el seu reconeixement.
- Es permet l'ús comercial de l'obra i de les seves possibles obres derivades: Incrementa les possibilitats que una empresa l'utilitzi i, per tant, més reconeixement al seu autor.
- Els treballs derivats han d'utilitzar la mateixa llicència: Això assegura que es segueixi respectant els terminis que l'autor original ha volgut aplicar.

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

El codi es pot trobar en l'enllaç següent:

[https://github.com/MarcGuerreroM/habitacalia\\_scraper](https://github.com/MarcGuerreroM/habitacalia_scraper)

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

El *dataset* en format CSV es pot trobar en l'enllaç anterior juntament amb el codi que el genera.

## Taula de contribucions

Contribucions	Signa
Recerca prèvia	M.G.M
Redacció de respostes	M.G.M
Desenvolupament del codi	M.G.M

## Referències

1. Mitchel, R. (2015). *Web Scraping with Python: Collecting Data from the Modern Web*. O'Reilly Media, Inc.