1. **Context.** Explicar en quin context s’ha recollit la informació. Explicar per què el

lloc web triat proporciona aquesta informació. Indicar l’adreça del lloc web.

Donada la creixen inestabilitat econòmica, increment de la inflació i la conseqüent pujada dels tipus d’interès, el sector immobiliari sembla estar a punt d’entrar en un canvi de cicle. Això genera dubtes i preocupacions a les persones que estan buscant comprar o vendre un habitatge.

És per això que sembla interessant poder obtenir informació sobre habitatges en venda per tal de poder seguir l’evolució d’aquests en el temps i per tan poder obtenir algunes respostes o ser capaços de poder preveure com evolucionar el sector.

Per fer-ho hem decidit fer web scrapping en un dels portals de venda de cases més important del país, Habitaclia.

Més en concret ens hem centrat en Barcelona ciutat per tal de poder simplificar el data set, amb la qual cosa l’enllaç de referència a sigut:

<https://www.habitaclia.com/viviendas-barcelona.htm>

El motiu d’aquesta selecció és que d’entre els portals estudiats d’on podríem obtenir la informació requerida, HABITACLIA ens ha semblat el que permetia fer el scrapping d’una manera més senzilla i ràpida.

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

El títol del data set és: Houses to Sell

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 escollit.

Les dades extretes i guardades en Houses to Sell són un conjunt de dades que ens permeten entendre alguns dels paràmetres més rellevants dels habitatges a la venda en la zona de Barcelona.

Paràmetres com preu de venda, superfície de l’habitatge o ubicació, entre d’altres, ens permetran poder fer un anàlisis sobre l’evolució de l’oferta d’habitatges en l’àrea de Barcelona en funció del temps.

Comentar també la intenció de que aquest data set sigui dinàmic. És a dir la idea és que de manera periòdica (cada dia, setmana, etc) es faci un run del main.py i la informació s’afegeixi al data set existent (especificant la data i hora). D’aquesta manera podrem estudiar com evolucionen els preus dels habitatge en funció del temps i podrem treure paràmetres com per exemple temps en que es tarda a vendre un habitatge en una zona determinada.

4. **Representació gràfica.** Dibuixar un esquema o diagrama que identifiqui el

dataset visualment i el projecte escollit.

5. **Contingut.** Explicar els camps que inclou el dataset i el període de temps de

les dades.

**Product\_id**: ens indica el codi de l’habitatge. És la primary key de cada habitatge en el data set

**Advert\_id**: és el codi de l’anunci en particular. La diferència entre aquest i l’anterior és que per exemple un habitatge pot estar anunciat per diferents agències, de manera que en aquest cas el product\_id seria el mateix però el advert\_id seria diferent.

**Price:** indica el preu de venda de l’habitatge.

**City\_name:** és el nom de la ciutat on es troba l’habitatge. En el nostre cas sempre serà Barcelona ja que ens hem centrat en aquesta ciutat.

**Zone:** és el nom de la zona o el barri on es troba l’habitatge.

**City\_code:** és el codi de la ciutat on es troba l’habitatge, en el nostre cas sempre serà 1930008. Aquest valor és redundant a City\_name, però el considerem útil ja que en certes ocasions pot ser útil tenir el codi en format numèric.

**Zone\_code:** és el codi de la zona o barri. És més útil que el Zone ja que aquest al ser una string pot tenir moltes diferències entre anuncis, mentre que el Zone\_code és únic a cada zona el que el fa molt útil per filtrar o operar per zones.

**Sqaure\_meters:** proporciona la superfície de l’habitatge en metres quadrats.

**Bed\_rooms:** indica el número d’habitacions de l’habitatge.

**Toilets:** són el número de lavabos de l’habitatge.

**Advert\_url:** és l’enllaç de l’anunci per si es vol consultar. Ens ha sigut molt útil per a poder comprovar que el data set era correcte.

**Today’s\_date:** en aquest paràmetre es fixa l’aspecte temporal. Ens indica la data i hora del moment de descarregar dels valors de l’habitatge.

Inicialment aquests són els paràmetres que guardarem en el data set, però destacant que es poden afegir i treure paràmetres en funció de les necessitats específiques.

6. **Propietari.** Presentar el propietari del conjunt de dades. És necessari incloure

cites d'anàlisis anteriors o, en cas de no haver-n’hi, justificar aquesta cerca amb

anàlisis similars. Justificar quins passos s’han seguit per actuar d’acord amb els

principis ètics i legals en el context del projecte.

7. **Inspiració.** Explicar per què és interessant aquest conjunt de dades i quines

preguntes es pretenen respondre. És necessari comparar amb les anàlisis

anteriors presentades a l’apartat 6.

8. **Llicència.** Seleccionar una d'aquestes llicències pel dataset resultant i justificar

el motiu de la seva selecció. Exemples de llicències que poden considerar-se:

● 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.

● Altres (especificar quina).

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

alternativament, en R.

● El codi haurà de situar-se a la carpeta **/source** del repositori.

● S'han d'indicar les llibreries i versions utilitzades. P. ex., en Python

poden obtenir-se mitjançant la comanda

pip3 freeze > requirements.txt

● Al document PDF s'han de comentar els aspectes més rellevants sobre

com el codi realitza el procés de recol·lecció de dades, quines dificultats

presenta el lloc web triat, i com les heu resolt.

10. **Dataset.** Publicar el dataset obtingut en format CSV a Zenodo, incloent-hi una

breu descripció. Obtenir i adjuntar l'enllaç del DOI del dataset

(https://doi.org/…). El dataset també haurà d’incloure’s a la carpeta **/dataset** del

repositori.

Si existeix qualsevol circumstància que impedeixi publicar obertament el

dataset real a Zenodo, s’haurà de: (1) comentar aquesta circumstància i

justificar el motiu en aquest apartat; (2) generar un dataset simulat i publicar-lo

a Zenodo, obtenint l'enllaç del DOI; i (3) comunicar al professor el dataset real

de manera privada (p. ex., utilitzant un repositori privat).

11. **Vídeo**. Realitzar un breu vídeo explicatiu de la pràctica **(màxim 10 minuts)**,

que haurà de comptar amb la participació dels dos integrants del grup. Al vídeo

s'haurà de realitzar una presentació del projecte, destacant els punts més

rellevants, tant de les respostes als apartats com del codi utilitzat per a extreure

les dades. Indicar l'enllaç del vídeo (https://drive.google.com/…), que haurà

d’estar al Google Drive de la UOC.

|  |  |
| --- | --- |
| **Contribució** | **Signatura** |
| Investigació prèvia | XBF, RAC |
| Redacció de respostes | XBF, RAC |
| Desenvolupament del codi | XBF, RAC |
| Participació al vídeo | XBF, RAC |