Activitat grupal.
Pràctica 1: Tipologia i cicle de vida de les dades

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.

Segons l'Institut nacional d'estadística (2022) el preu de l'habitatge s'ha incrementat en més d'un 20% respecte 2018. Aquest fet ha estat propiciat en base a diverses variables que conjuntament han incrementat els preus generals. Alguns factors clau han estat la manca d'oferta de productes, l'acumulació d'estalvis en temps de pandèmia, la recent guerra a Ucraïna entre d'altres. En tot cas, el sector immobiliari no n'ha estat exempt. Per altra banda, el preu del lloguer també presenta una situació excepcional. "La llei de l'oferta i la demanda en la seva màxima expressió: l'oferta de pisos de lloguer a Barcelona està en mínims històrics i, per tant, els preus han tornat a màxims." (Cerro, G, Xavier. 2022)

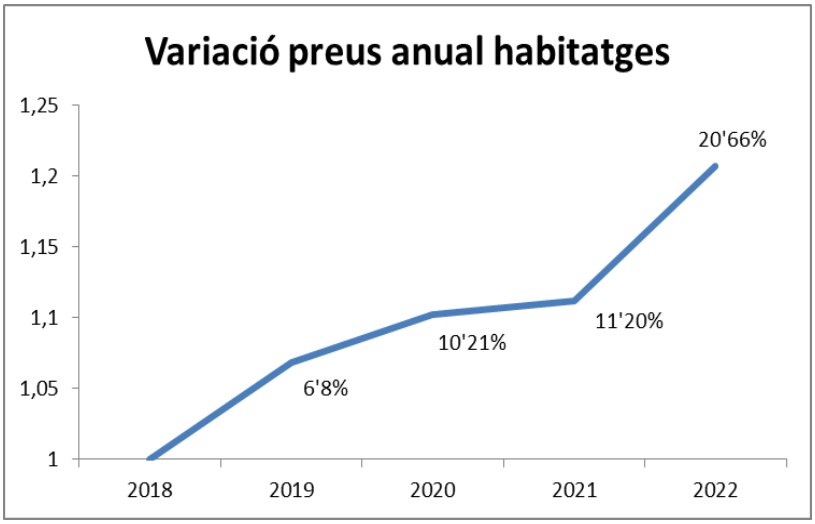

Figura 1. Variació en percentatge de preus de l'habitatge a España. **Font:** Adaptació INE 2022.

El portal web Habitaclia es situa entre els tres buscadors immobiliaris més consultats a l'estat espanyol. Aquest fet, sumat al coneixement del mateix per part dels autors del projecte, fan que es determini el portal web Habitaclia.com com el més adequat per utilitzar en el treball proposat.

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

Habitaclia. Quina població desitges cercar?

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.

El projecte de web scrapping realitzat presenta un anàlisis del portal web per finalment generar un document csv amb un llistat d'immobles i les seves característiques. De manera més detallada, el procés del web scrapping utilitzat s'executarà, demana la població desitjada i un cop acabi el procés de codi generarà un document amb el llistat d'habitatges en venda de la població.
El codi presentat realitza la recerca dels urls necessaris per a cobrir tots els habitatges publicats al portal web fins que no troba més immobles.

Column1	type	size	price	rooms	baths
0	Piso	85	135000	3	2
1	Casa	350	845000	5	3
2	Piso	105	115500	4	2
3	Piso	50	85000	1	1
4	Chalet	350	368000	5	2
5	Piso	60	105000	0	0
6	Masía	700	2500000	6	5
7	Dúplex	50	95000	2	1
8	Piso	79	138000	4	2
9	Casa	219	169900	4	2
10	Masía	350	430000	5	1

4. Representació gràfica. Dibuixar un esquema o diagrama que identifiqui el dataset visualment i el projecte escollit.

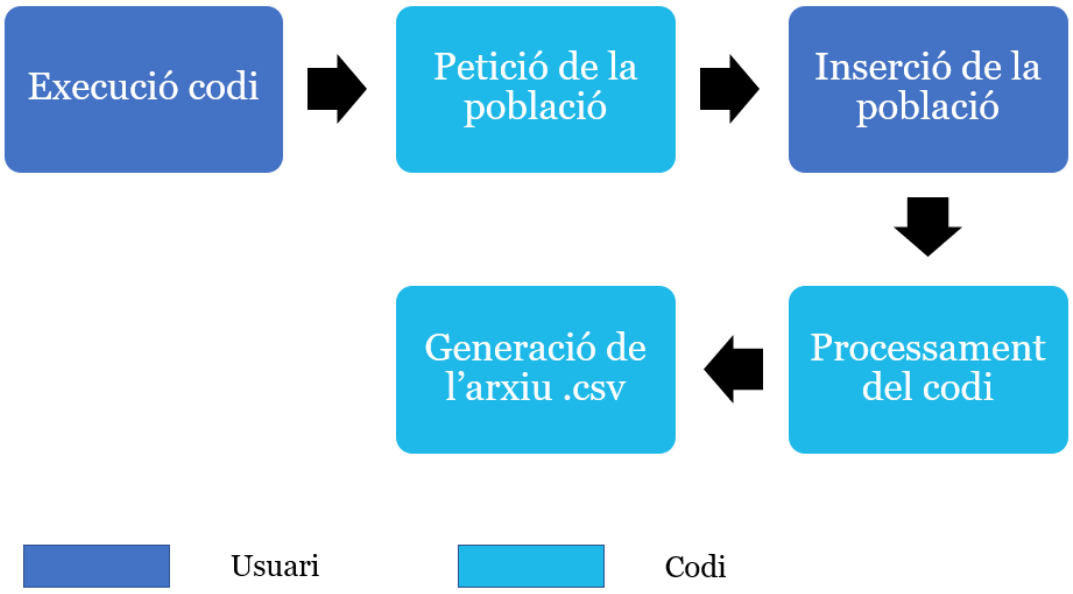

Figura 2. Representació gràfica del procés.
Font: Elaboració pròpia.

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

Número de registre: Enumera cada registre
Type: Tipus d'habitatge (per exemple: piso, casa, apartamento, chalet, ático)
Size: Mida, metres quadrats de l'habitatge
Price: Preu
Rooms: Nombre d'habitacions
Baths: Nombre de banys

Data: cada extracció guarda la data en el nom del fitxer, per tant és a cada execució en diferents dates de la mateixa població obtindrem una sèrie de fitxers contenint tots els anuncis de la data que es fa la cerca.

Tal com podem veure a robot.txt del web (veure document robots.txt del repositori Github). Hi ha informació que no permet d'extreure informació sobre els seus usuaris, o captar la validació Captcha per emular navegació humana, o que no permet ordenar per valoracions positives o les millors puntuacions, limita també la captura de fotografies o de localitzacions i per això s'ha respectat al màxim els desitjos de Habitaclia.

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.

El propietari es troba amb whois (veure whois.txt del repositori Github)
En aquest cas no es pot identificar a una persona física, però deixa clar que no vol transferir Clients (clientTransferProhibited)

Anàlisis anteriors:

Títol	Detall	Url de consulta
Habitaclia Scraper	Extreu tots els anuncis d'habitatges de venda a Catalunya i els posa en un dataset	https://github.com/MarcGuerreroM/habitaclia_scraper
Habitaclia Airbnb Anàlisis	Afecta airbnb al preu dels pisos de lloguer?	https://github.com/thealgorithmlabs/habitaclia_airbnb_analysis
Habitaclia-Scrapping	Crea un dataset amb els pisos de lloguer	https://github.com/zechao/habitaclia-scraping
Pràctica 1	Lloguer a Barcelona en un moment determinat	https://github.com/earinos/practica1

Taula 1. Relació d'anàlisis anteriors del projecte. Font: Elaboració pròpia.

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.

El que diferencia el codi que estem presentant és que permet fer una cerca dinàmica triant una sèrie de valors per identificar els pisos de venda en una població Espanyola.
S'ha mirat de fer-lo operatiu en cas de noms compostos afegint guió baix tal com es fa a la pròpia web.

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.

Després d'analitzar i cercar quin tipus de copyright podria tenir el portal web habitaclia concloure que és de domini públic, és a dir, habitaclia publica de manera pública el contingut de la seva web per a que els usuaris puguin reutilitzar-lo amb finalitats pròpies.
En aquest sentit, la llicència a considerar seria Released Under Cco: Public Domain License.

Segons Creative Commons (2017) CC0 permet que contingut protegit per drets d'autor o bases de dades renunciar a aquests interessos en les seves, de manera que altres puguin basar-se lliurement sobre, millorar i reutilitzar les obres per a qualsevol propòsit sense restriccions segons la llei de drets d'autor o bases de dades.

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

- **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.**

```
# Importem llibreries
from datetime import datetime
import pandas as pd
import requests
import re
from bs4 import BeautifulSoup

# Apliquem opció de dos decimals al data
pd.set_option('display.float_format', lambda x: '%.2f' % x)

# Generem la data que posteriorment aplicarem al nom del dataset
fecha=datetime.now()

# Modifica user agent
headers = {'User-Agent': 'Mozilla/5.0'}

# Generem input per a que l'usuari apliqui la població desitjada
URL_town = input('Quina població desitjes cercar? ')

# Generem dataframe
df = pd.DataFrame()

# Generem llistes buides amb les variables que ens generarà el codi
type_list = []
price_list = []
size_list = []
bath_list = []
room_list = []

# Muntem loop que anirà buscant en les pàgines web les variables i els afegirà a la llista
for i in range(5000):
    URL_habitaclia = 'https://www.habitaclia.com/viviendas-'
    URL_town = re.sub(" ","_", URL_town)
    URL_final = '-' + str(i) + '.html'
    URL_total = URL_habitaclia + URL_town + URL_final
    response = requests.get(URL_total)
    soup = BeautifulSoup(response.text, "html.parser")
    check = soup.findAll('div', attrs={"class":"list-no-result-title"})
    check = str(check)
    if re.search('No', check):
        break;
    else:
```

```
        item_content_first = soup.findAll('section', attrs={"class":"list-item-content"})
        item_content_second = soup.find_all('section', attrs={"class":"list-item-content-second"})
        for element in item_content_second:
            preu = element.findAll('span', attrs={"class":"font-2"})
            preu = str(preu)
            preu = re.search('>(.+?) €<', preu)
            try:
                price_list.append(preu.group(1))
            except AttributeError:
                price_list.append(preu)

        for element in item_content_first:

            title = element.findAll('a', href=True)
            title = str(title)
            title = re.search('title=["](.+?)["]>', title)
            try:
                title = re.search('(.+?)[" "]', title.group(1))
            except AttributeError:
                title = re.search('(.+?)[" "]', str(title))
            try:
                type_list.append(title.group(1))
            except AttributeError:
                type_list.append(title)
            size = element.findAll('p', attrs={"class":"list-item-feature"})
            size = str(size)
            try:
                size = re.search('[0-9]+m', size).group(0)
            except AttributeError:
                size = re.search('[0-9]+m', size)
            try:
                size = re.search('[0-9]+', str(size)).group(0)
            except AttributeError:
                size = re.search('[0-9]+', str(size))
            size_list.append(size)
            room = element.findAll('p', attrs={"class":"list-item-feature"})
            room = str(room)
            room = re.search('[0-9][" "]hab', room)
```

```
            if room:
                room = re.search('[0-9]+[" "]', room.group(0)).group(0)
            else:
                room = "0"
            room_list.append(room)
            bath = element.findAll('p', attrs={"class":"list-item-feature"})
            bath = str(bath)
            bath = re.search('[0-9][" "]baño', bath)
            if bath:
                bath = re.search('[0-9]+[" "]', bath.group(0)).group(0)
                bath_list.append(bath)
            else:
                bath = "0"
                bath_list.append(bath)

# Generem el data amb les columnes igual a les llistes anteriorment completades.
df = df.assign(type=type_list,size=size_list, price=price_list, rooms=room_list, baths=bath_list)

# Generem el document csv amb el format de població + data +.csv
df.to_csv((str(URL_town)+"_"+format(fecha.strftime("%Y%m%d"))+".csv"), index=True)
```

El codi s'ha fet amb Python mitjançant codi:
https://github.com/Alzatrak/Practica_1

Document csv de sortida:
https://github.com/Alzatrak/Practica_1/blob/main/Dataset/castellar_del_valles_20221116.csv

En la realització d'aquesta pràctica no hem estat lliures de dificultats.
No podem obviar el fet que és la pràctica s'ha de fer en grup i que tot i que havíem coincidit en una altra assignatura la possibilitat de trobar-nos presencialment ha estat escassa, per viure a diferents poblacions i tenir horaris laborals totalment diferent.
Ens hem hagut d'adaptar força un a l'altre i suplir les diferents dificultats o entrebancs trobats.
Un cop triat el tema i la web a la que fer scraping vam buscar a Github exemples similars, però en molts casos no funcionaven a l'executar-los.
Un de nosaltres s'havia canviat el portàtil recentment i la versió d' Anaconda instal·lada té un spyder amb problemes i cada cop a l'entrar havia d'executar una correcció a la lá línea de comandes d'Anaconda i actualitzar el Kernel cada vegada abans de començar. Afortunadament el portàtil antic tenia una versió sense problemes amb input() i es va poder continuar treballant sense problemes.
Es va considerar d'afegir latitud i longitud al dataset per posteriorment poder fer anàlisis amb geolocalització, però calia entrar dins de cada anunci per aprofundir en la informació i hauria incrementat el temps d'execució del codi i segurament trobaria més traves per part dels mecanismes preparats pel propietari del portal per evitar extreure dades que no es volen permetre de capturar i difondre.

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).

https://doi.org/10.5281/zenodo.7327923

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.

Contribucions	Signatura
Investigació prèvia	SGM, LTA
Redacció de respostes	SGM, LTA
Desenvolupament del codi	SGM, LTA
Participació al vídeo	SGM, LTA

Bibliografía

Cerro, X. G. del. (19 d'octubre de 2022). Es desploma l'oferta d'habitatge de lloguer a Barcelona. Ara.cat. recuperat el 27 d'octubre de 2022 de: https://www.ara.cat/economia/immobiliari/oferta-d-habitatge-lloguer-barcelona-minims_1_4522933.html

Creative Commons. (22 de juny de 2017). *CCO*. https://creativecommons.org/share-your-work/public-domain/cc0/

INE - Instituto Nacional de Estadística. (2on trimestre de 2022). *Índice de precios de vivienda. Últimos datos*. Recuperado el 27 d'octubre de 2022, de: https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C

Resolució del problema de versió Spyder a Anaconda.

https://github.com/spyder-ide/spyder/issues/17616#issuecomment-1088750490