

## Web scraping

```
In [4]: import bs4 as bs
import urllib
import urllib.request
import re
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import io

In [5]: HTML = "https://www.trovimap.com/precio-vivienda/barcelona"

In [6]: # Amb .read_html() podem obtenir fàcilment la taula que ens mostra la web.

df = pd.read_html(HTML)
df[0].head()

Out[6]: Unnamed: 0 Variación Mensual Variación 3 meses Variación anual €/m2
0 Alella -0.08% +0.19% +2.23% 2.867 €
1 Arenys de Mar -1.93% +1.03% +7.64% 2.287 €
2 Badalona -0.90% -1.43% -2.23% 2.408 €
3 Barberà del Vallès +2.77% -1.11% -2.72% 2.096 €
4 Barcelona +0.10% -0.86% -1.51% 4.059 €

In [7]: # Amb BeautifulSoup, podrem obtenir tot el String del codi HTML per si volem obtenir més informació.

historypage = urllib.request.urlopen(HTML)
soup = bs.BeautifulSoup(historypage,'html.parser')
makeitastring = ''.join(map(str, soup))

#soup
#makeitastring

In [8]: #retorna un llistat de ciutats i la seva variació de preu segons el html ja llegit per BeautifulSoup que li hem passat. Aquest html està format segons
# La província. Per exemple: HTML = "https://www.trovimap.com/precio-vivienda/barcelona" retornarà un llistat amb les
# ciutats que es troben en aquesta pàgina que son les ciutats de Barcelona i les seves variacions de preus mensual, els últims 3
mesos
# anual i preu euro/ metre quadrat.

def obtenerCiutatsVariacio(htmlSoup):
    tableProv = soup.find('table', {'class':'table table-condensed precio-medio-table'})
    tbody = tableProv.find('tbody')

    llistaVarCiutats = []
    for row in tbody.findAll('tr'):
        cells = row.findAll('td')
        link = cells[0].find('a')
        ciutat = link.find(text=True)
        print(ciutat)
        varMensual = cells[1].find(text=True)
        var3mesos = cells[2].find(text=True)
        varAnual = cells[3].find(text=True)
        eumetre = cells[4].find(text=True)
        element=[ciutat,varMensual,var3mesos,varAnual,eumetre]
        llistaVarCiutats.append(element)

    return llistaVarCiutats

HTML = "https://www.trovimap.com/precio-vivienda/barcelona"

VarCiutat = obtenerCiutatsVariacio(soup)
print(VarCiutat[0])
['Alella', '-0.08%', '+0.19%', '+2.23%', '2.867 €']

In [ ]:

In [ ]:

In [9]: HTML = "https://www.trovimap.com/precio-vivienda/"

In [10]: HTMLEspaña = HTML + "espana"

In [11]: df = pd.read_html(HTMLEspaña)
dfProvinciesEspana = df[0]
dfProvinciesEspana = dfProvinciesEspana.rename(columns={"Unnamed: 0": "Provincies"})
dfProvinciesEspana.head()

Out[10]: Provincies Variación Mensual Variación 3 meses Variación anual €/m2
0 A Coruña +1.48% -0.16% -6.86% 1.489 €
1 Albacete +3.29% +2.60% -1.62% 1.262 €
2 Alicante/Alicant -0.71% -0.05% -2.16% 1.722 €
3 Almería +3.52% +4.26% +3.28% 1.193 €
4 Araba/Álava +0.18% -0.27% -4.28% 2.092 €

Obtenim una taula amb la variació mensual de cada província de l'estat.

Si volem obtenir una taula, tenint per files les ciutats més importants de cada província tal com hem fet anteriorment amb la de Barcelona:
```

```
In [11]: dfProvinciesEspana.iloc[0]['Provincies']

Out[11]: 'A Coruña'

Si volem obtenir una taula, tenint per files les ciutats més importants de Catalunya:
```

```
In [12]: # Agafar la taula per cada una de les quatre províncies de Catalunya.
dfBarcelona = pd.read_html(HTML,"barcelona")
dfTarragona = pd.read_html(HTML,"tarragona")
dfGirona = pd.read_html(HTML,"girona")
dfLleida = pd.read_html(HTML,"lleida")
dfBarcelona = dfBarcelona[0]
dfTarragona = dfTarragona[0]
dfGirona = dfGirona[0]
dfLleida = dfLleida[0]

# Concatenar Taulas.
dfCatalunya = pd.concat([dfBarcelona, dfTarragona], ignore_index=True)
dfCatalunya = pd.concat([dfCatalunya, dfGirona], ignore_index=True)
dfCatalunya = pd.concat([dfCatalunya, dfLleida], ignore_index=True)

# Renombrar columns.
dfCatalunya = dfCatalunya.rename(columns={"Unnamed: 0": "Ciutat", "Variación Mensual": "Variació mensual",
                                            "Variación 3 meses": "Variació tres mesos", "Variación anual": "Variació anual"})

dfCatalunya.tail()

Out[12]: Provincies Variació mensual Variació tres mesos Variació anual €/m2
93 Tossa de Mar -2.64% -3.05% -2.97% 2.682 €
94 Vídures +0.08% +0.22% -3.04% 1.423 €
95 Lleida +1.59% +1.30% +6.53% 1.199 €
96 Naut Aran +1.95% - - 3.948 €
97 Vielha e Mijaran - - -0.94% 2.381 €

Ara volem treure totes les ciutats importants de Espanya. Per fer això necesitem totes les url de cada província a partir de l'html.
```

```
In [43]: historypage = urllib.request.urlopen(HTMLEspaña)
soup = bs.BeautifulSoup(historypage,'html.parser')
makeitastring = ''.join(map(str, soup))
#soup
#makeitastring

# Buscar el patró dins del html on es guarda els noms de les províncies per després buscar-les en url.
capitals = '/'.join(map(str, re.findall("precio-vivienda.(+?)>",makeitastring)))
capitals = capitals[46:570]
print(capitals)

# Ja tenim un array amb tots els noms de les províncies que haurem de posar a la url.
capitalsArray = capitals.split("//")
capitalsArray

Out[43]: ['a-coruña',
'albacete',
'alicante-alacant',
'almeria',
'araba-alava',
'asturias',
'avila',
'badajoz',
'barcelona',
'bikaiia',
'burgos',
'caceres',
'cadiz',
'cantabria',
'castellon-castello',
'ceuta',
'ciudad-real',
'cordoba',
'cueta',
'gipuzkoa',
'girona',
'granada',
'guadalajara',
'huelva',
'huesca',
'illes-baleares',
'jaen',
'la-rioja',
'las-palmas',
'leon',
'lleida',
'lugo',
'madrid',
'malaga',
'melilla',
'murcia',
'navarra',
'ourense',
'palencia',
'pontevedra',
'salamanca',
'santa-cruz-de-tenerife',
'segovia',
'sevilla',
'soria',
'tarragona',
'teruel',
'toledo',
'velencia-valencia',
'veladoloid',
'zamora',
'zaragoza']

In [48]: HTML = "https://www.trovimap.com/precio-vivienda/"
HTML = https://www.trovimap.com/precio-vivienda/[25]

df = pd.read_html(HTML + capitalsArray[25])
df[0].head()

Out[48]: Unnamed: 0 Variación Mensual Variación 3 meses Variación anual €/m2
0 Alaior -0.67% - - +10.07% 2.130 €
1 Alcudia +0.22% -5.53% -1.42% 2.784 €
2 Andrabi +3.61% +0.85% +5.89% 2.944 €
3 Calvià +0.29% +0.05% +5.34% 3.769 €
4 Campos -2.92% -2.34% +17.20% 2.449 €

In [64]: # Veiem que serveix, ara només hem de posar-lo en un loop per obtenir totes les ciutats importants d'Espanya.
dfEspana = pd.DataFrame()
for capital in capitalsArray:
    if capital != "ceuta" and capital != "soria":
        # Per a cada url agafar la taula
        dfAux = pd.read_html(HTML+capital)
        dfAux = dfAux[0]
        #Anar concatenant ciutats
        dfEspana = pd.concat([dfEspana, dfAux], ignore_index=True)
        print(capital)

# Renombrar columns.
dfEspana = dfEspana.rename(columns={"Unnamed: 0": "Ciutat", "Variación Mensual": "Variació mensual",
                                            "Variación 3 meses": "Variació tres mesos", "Variación anual": "Variació anual"})

dfEspana.tail()
```

```
a-coruna
albacete
alicante-alacant
almeria
araba-alava
asturias
avila
badajoz
barcelona
bikaiia
burgos
caceres
cadiz
cantabria
castellon-castello
ceuta
ciudad-real
cordoba
cueta
gipuzkoa
girona
granada
guadalajara
huelva
huesca
illes-baleares
jaen
la-rioja
las-palmas
leon
lleida
lugo
madrid
malaga
melilla
murcia
navarra
ourense
palencia
pontevedra
salamanca
santa-cruz-de-tenerife
segovia
sevilla
soria
tarragona
teruel
toledo
velencia-valencia
veladoloid
zamora
zaragoza

Out[64]: Provincies Variació mensual Variació tres mesos Variació anual €/m2
617 Zamora +0.65% +2.24% +1.27% 1.284 €
618 Calatayud +2.31% - +6.26% 882 €
619 Cuarte de Huerva +2.42% +3.17% -6.14% 1.352 €
620 Utebo - -6.41% - 1.363 €
621 Zaragoza -0.47% -0.85% +1.00% 1.684 €

Ja tenim taules amb les variacions econòmiques de les ciutats més importants tant de Catalunya com d'Espanya.
```