

FIFTY

OUTLET STORES

SINCE 1997

# WEB SCRAPING PROJECT

**AUTORS:** Miquel Tomé i Guillem Fernández

**CURS:** Màster en Data Science 2020-2021

**ASSIGNATURA:** Tipologia i Cicle de Vida de les Dades

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

El nostre objectiu consistia en poder crear un script que de manera setmanal ens proporcionés dades sobre banyadors per home, un dels productes més venuts durant la primavera per la proximitat a l'estiu. Així doncs, teníem la necessitat de trobar una pàgina web que recollís informació de diferents banyadors i marques.

Aquesta pràctica, doncs, es centra en l'extracció de dades a partir d'una pàgina web de roba. Específicament, ens centrem en els articles de bany masculins que es troben a FiftyOutlet, pàgina web que recull peces de roba de temporades passades amb un descompte important (tipus outlet) de les marques Cortefiel, Pedro del Hierro, Springfield, Jack & Jones i Milano. Ens hem decantat per FiftyOutlet (Grupo Cortefiel) per fiabilitat de marca i perquè és una marca que ja coneixem i ens agraden els seus productes.

Després de revisar el fitxer *robots.txt* hem pogut comprovar que els propietaris de la pàgina ens permeten fer el web scraping sense masses restriccions.

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

Seguint allò descrit en el context, un títol descriptiu i adequat seria "Articles de bany masculins", ja que les dades que hi trobarem seran sobre banyadors per home (breu descripció, preu, link de compra, imatge, etc).

### **3- Descripció del dataset. Desenvolupar una descripció breu del conjunt de dades que s'ha extret.**

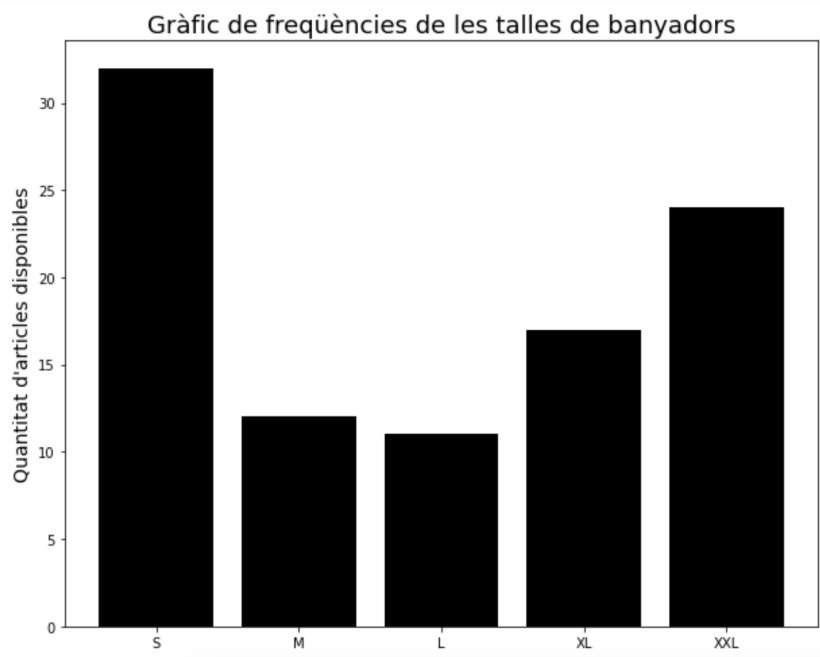
D'una banda, s'ha decidit extreure les dades descriptives del producte, com podrien ser la marca, el color, les talles disponibles i el seu nom o descripció, i d'altra banda s'han extret les dades relatives al seu preu actual i la rebaixa que suposa, és a dir, el seu preu original. Addicionalment, també s'ha inclòs el link a la pàgina web del producte per tal de poder donar accés a aquest en cas de decidir comprar-lo.

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

FiftyOutlet en definitiva és un outlet online de productes de marques concretes. Així doncs, les dues característiques que defineixen un outlet són:

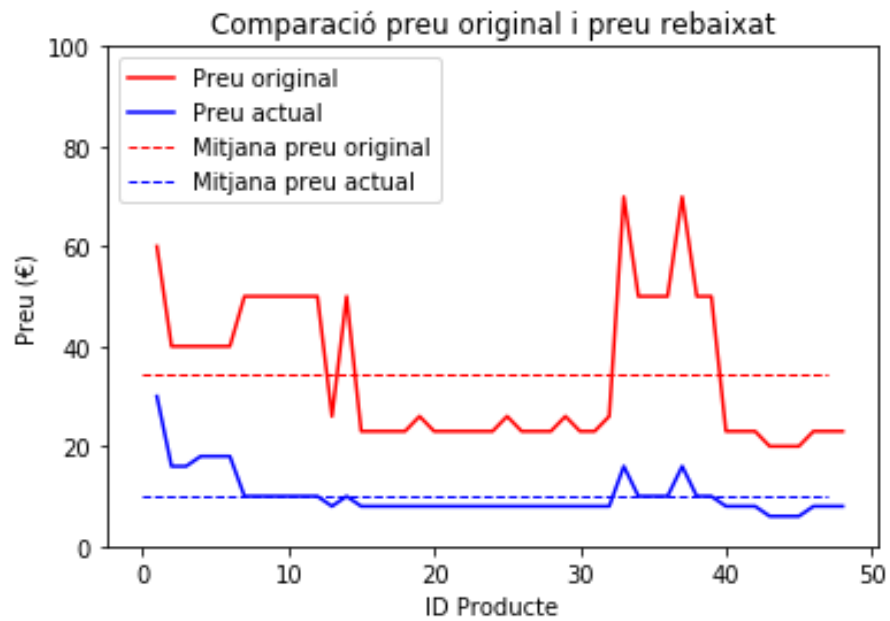
- Els grans descomptes sobre els articles.
- La poca disponibilitat de talles al tractar-se de productes de temporades anteriors.

Primerament, analitzarem la disponibilitat de talles a partir de les dades descarregades mitjançant la representació gràfica d'un histograma que ens permeti veure quina és la distribució que segueix aquesta variable del dataset.



Com veiem, els valors extrems (talles S i XXL) són les que més disponibilitat tenen, ja que són talles que tenen menys sortida al mercat. Per altra banda, les talles M, L i XL, talles que sol comprar molta més gent, són les que tenen menys disponibilitat.

D'altra banda, s'ha estudiat el descompte aplicat als articles disponibles per tal de tenir una visió global sobre la diferència entre el preu original i actual de tots els articles. La següent figura mostra la comparació entre els preus originals i els preus rebaixats, amb les seves respectives mitjanes.



Com es pot observar, existeix una diferència considerable entre ambdues variables i les seves mitjanes. Concretament, la diferència entre les dues mitjanes és d'aproximadament uns 24 €, sent les dues mitjanes de 34,36 € i 10,12 €. Aquest gran descompte és un dels reclams de la pàgina FiftyOutlet.

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

Les dades que es troben a la pàgina web de Fifty Outlet han sigut extretes mitjançant un script de Python que, a grans trets, recollia la informació del codi HTML de la pàgina en un diccionari en el qual cada article exposat tenia associat un identificador únic que feia el paper de clau primària. D'aquesta manera s'han recollit els atributs que s'han considerat rellevants i, després de ser convertits a una variable de tipus DataFrame mitjançant la llibreria Pandas, s'han exportat a un fitxer en format csv.

Els camps o atributs que s'han recollit al *dataset* són els següents:

1. **ID:** Variable que identifica de manera única cada registre del dataset.
2. **Descripció:** Descripció o nom del producte tal i com està recollit a la pàgina web.
3. **Marca:** Al tractar-se d'una web on hi ha més d'una marca, aquesta variable especifica a quina marca correspon el producte.
4. **Preu inicial:** Preu inicial del producte abans de la rebaixa.
5. **Preu final:** Preu final del producte després de la rebaixa.
6. **Disponibilitat de colors:** Colors en els quals està disponible el producte.
7. **Nombre de colors disponibles:** Nombre de colors en els quals està disponible el producte.
8. **Disponibilitat de talles:** Talles en les quals està disponible el producte.
9. **Link de compra:** Link que ens dirigeix al producte en cas que estiguem interessats en comprar-ho.
10. **Link de la imatge:** Link que mostra la imatge del producte corresponent.

L'extracció de dades s'ha dut a terme a data de 10/04/2021.

Cal esmentar que les dades del dataset s'actualitzaran automàticament quan un dels links deixi d'estar actiu. Hem pres aquesta decisió perquè considerem que quan un link d'un producte determinat deixi d'existir aleshores voldrà dir que aquell producte ja no existeix o ja no està disponible per comprar i, per tant, s'esborrarà del dataset.

Durant el desenvolupament del projecte de web scraping s'ha comprovat que el nombre de registres del set de dades ha anat variant entre 36 i 55 ítems, tot i que és possible que depenent del període de l'any en el qual s'extreguin les dades, aquest nombre pugui ser tant inferior com superior a l'interval esmentat.

**6- Agraïments. Presentar el propietari del conjunt de dades. És necessari incloure cites d'anàlisis anteriors o, en cas de no haver-les, justificar aquesta cerca amb anàlisis similars.**

Les dades extretes de la pàgina web <https://fiftyoutlet.com/es/es/hombre/bano> són propietat de l'empresa Tendam Retail S.A. S'ha comprovat el propietari de les dades amb la funció whois del mòdul amb el mateix nom (veure figura següent).

```
>>> import whois
>>> print(whois.whois("https://fiftyoutlet.com/es/es/hombre/bano"))
{
  "domain_name": [
    "FIFTYOUTLET.COM",
    "fiftyoutlet.com"
  ],
  "registrar": "Entorno Digital, S.A.",
  "whois_server": "whois.entorno.com",
  "referral_url": null,
  "updated_date": [
    "2021-01-18 23:49:30",
    "2021-03-20 11:47:02"
  ],
  "creation_date": "2018-04-19 11:12:08",
  "expiration_date": "2021-04-19 11:12:08",
  "name_servers": [
    "NS10.HOSTING-MEYTEL.NET",
    "NS11.HOSTING-MEYTEL.NET",
    "NS12.HOSTING-MEYTEL.NET",
    "NS13.HOSTING-MEYTEL.NET",
    "NS14.HOSTING-MEYTEL.NET",
    "NS15.HOSTING-MEYTEL.NET",
    "ns10.hosting-meytel.net",
    "ns11.hosting-meytel.net",
    "ns12.hosting-meytel.net",
    "ns13.hosting-meytel.net"
  ],
  "status": [
    "ok https://icann.org/epp#ok",
    "ok https://www.icann.org/epp#ok"
  ],
  "emails": "abuse@entorno.es",
  "dnssec": "unsigned",
  "name": null,
  "org": "Tendam Retail S.A.",
  "address": null,
  "city": null,
  "state": "MADRID",
  "zipcode": null,
  "country": "ES"
}
>>>
```

Anàlogament, s'han cercat anàlisis similars de la pàgina web escollida tant a GitHub com al buscador de Google i no s'han trobat resultats similars a l'anàlisi similar, per la qual cosa es considera que no és necessari fer esment a estudis realitzats amb anterioritat.



**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 conjunt de dades extretes i les posteriors versions que es poden obtenir amb el codi elaborat, que correspondran al conjunt de productes disponibles al lloc web en el moment de l'execució, poden ser d'utilitat per les marques que tenen els seus productes en venda a la web de Fifty Outlet i que volen estudiar-ne l'evolució temporal dels preus, la disponibilitat d'articles o la preferència de colors dels usuaris.

D'altra banda, poden ser útils per altres marques que busquin constituir un lloc web per vendre els seus productes rebaixats a mode d'outlet. D'aquesta manera podrien estudiar els preus de la competència i veure quina és la seva oferta.

**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

S'ha considerat oportú escollir la llicència CC BY-NC-SA 4.0 per la publicació del *dataset* generat pels següents motius:

- S'ha d'aportar el nom dels creadors del conjunt de dades generat, indicant els canvis que s'han realitzat respecte la versió original. Això implica que es segueixi reconeixent el treball realitzat pels autors originals i, alhora, les aportacions posteriors dels usuaris que hagin modificat el projecte original.
- No es permet l'ús comercial d'aquest treball ja que s'ha enfocat estrictament per a funcions d'aprenentatge i difusió de coneixement. Així doncs, l'objectiu és que pugui ser reaprofitat per un objectiu acadèmic i en cap cas per un ús comercial.
- Les contribucions realitzades a posteriori sobre el treball publicat sota aquesta llicència hauran de publicar-se sota la mateixa llicència. D'aquesta manera, els autors originals veurà versions posteriors del seu treball publicades amb les condicions i termes que ell mateix ha escollit.

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

**10- Dataset. Publicar el dataset en format CSV a Zenodo (obtenció del DOI) amb una breu descripció.**

El dataset resultant del nostre codi ha estat penjat correctament a Zenodo.

El DOI obtingut ha estat: [10.5281/zenodo.4679508](https://doi.org/10.5281/zenodo.4679508)

The screenshot shows the Zenodo interface for a dataset named 'men\_swimwear\_fiftyOutlet'. The page includes a search bar, user profile 'guillem.fdez@gmail.com', and buttons for 'Dataset' and 'Open Access'. The dataset is dated April 11, 2021, and is attributed to Miquel Tomé Carreño and Guillem Fernández Pallarès. It contains 4 files. A table provides details for each file, including its name, description, URL, price, and availability. The files are: 1. PedrodellHiero (Bañador vichy PdH), 2. milano (Bañador básico secado rápido), 3. milano (Bañador básico secado rápido), and 4. milano (Bañador estampado). The dataset is indexed in OpenAIRE and has a Creative Commons Attribution 4.0 International license.

File ID	File Name	Description	URL	Price (€)	Availability	Count
1	PedrodellHiero	Bañador vichy PdH	/es/es/hombre/bano/banador-vichy-pdh/9149112.html	59.99	29.99	0
2	milano	Bañador básico secado rápido	/es/es/hombre/bano/banador-basico-secado-rapido/9149120.html	39.99	15.99	1
3	milano	Bañador básico secado rápido	/es/es/hombre/bano/banador-basico-secado-rapido/9149120.html	39.99	15.99	1
4	milano	Bañador estampado	/es/es/hombre/bano/banador-estampado-secado-	39.99	17.99	0

Contribucions	Signa
Recerca inicial	Guillem Fernández Pallarès Miquel Tomé Carreño
Redacció de les respostes	Guillem Fernández Pallarès Miquel Tomé Carreño
Desenvolupament del codi	Guillem Fernández Pallarès Miquel Tomé Carreño