Universitat Oberta de Catalunya

uoc.edu

M2.951

TIPOLOGIA I CICLE DE VIDA DE LES DADES

PRÀCTICA 1

Alumnes: Ricard Piqué, Adrià Jaraba

Pràctica 1

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.

   Degut a la conjuntura actual inflacionista, on els preus han començat a incrementar, es vol fer una comparativa entre supermercats i els preus d'una llista de productes que tenen en comú.
   També es vol aprofitar per comparar un mercat que ha guanyat rellevància passada la crisi sanitària del COVID-19, el mercat de la compra online en supermercats.
   La direcció web escollida correspon a www.soysuper.com. En aquesta pàgina es podrà omplir la cistella i oferirà els diferents preus, dels mateixos o productes similars, en un altre supermercat conegut. Classifica els productes per categories i té en compte la zona on vius (En aquest cas s'ha triat Barcelona per la ubicació de l'estudiant).

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

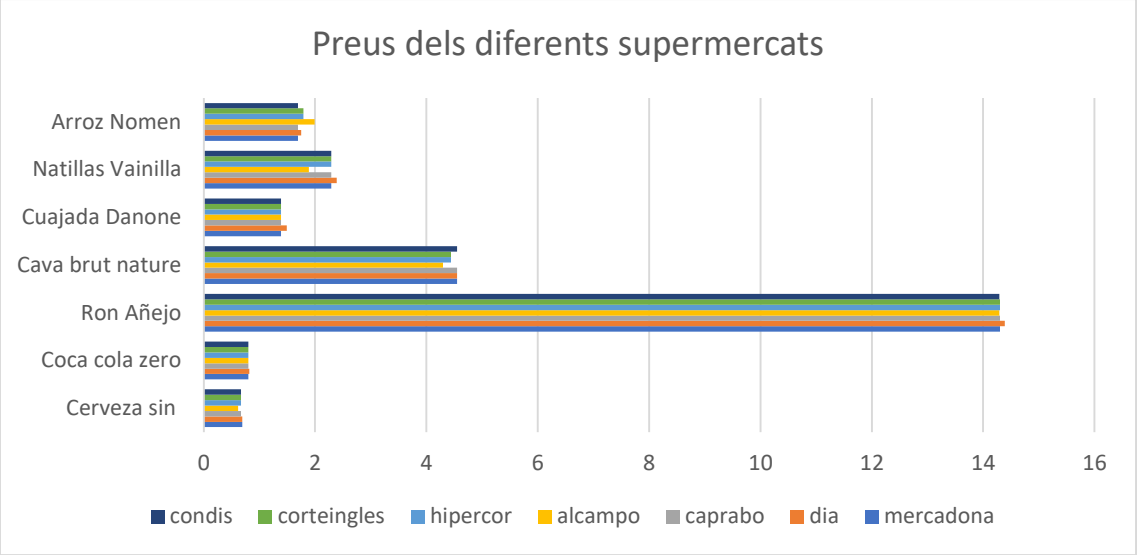
   El títol triat és: "Comparativa de preus entre diferents supermercats donada una cistella de compra comuna."

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 conjunt de dades extret al dataset, conté la comparativa d'uns productes concrets. En aquest cas, s'ha utilitzat un codi postal referent a Barcelona (08028) per trobar aquells supermercats més propers. Això és una dada a tenir en compte a l'hora de comparar preus ja que poden variar segons la zona escollida.
   Les variables triades fan referència al producte seleccionat, la data del web scraping i el preu en els diferents supermercats localitzats del voltant que porten la compra a casa.

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

   A continuació mostrem un gràfic de barres agrupades del resultat del dataset, on s'identifiquen els preus dels productes amb el seu corresponent supermercat.



Extracte de la notícia de El País que anuncia el augment de preus dels aliments bàsics de la cistella.


CincoDías

EL PAÍS ECONOMÍA

Compañías Mercados Economía Mi Dinero Fortuna / Cotizaciones

IPC

INFLACIÓN >

Algunos de los 20 alimentos básicos de la cesta de la compra suben hasta el 43% en octubre

J. GARCÍA / N. RODRIGO

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

Es mostra una taula amb la agrupació de les diferents variables que apareixen al dataset obtingut amb el seu corresponent tipus i descripció.

Cada producte conté la data d'execució del scraping i quedarà registrada per a futures comparacions, complint el propòsit del nostre estudi. En aquest cas, la data emprada ha sigut el 18/11/2022.

VARIABLE	TIPUS	DESCRIPCIÓ
X	Integer	Número identificatiu de cada registre.
Productes	Character	Nom del producte.
Data	Character	Data de la realització de scraping.
Mercadona	Numeric	Preu al supermercat Mercadona.
Dia	Numeric	Preu al supermercat Dia.
Caprabo	Numeric	Preu al supermercat Caprabo.
Alcampo	Numeric	Preu al supermercat Alcampo.
Hipercor	Numeric	Preu al supermercat Hipercor.
Corteingles	Numeric	Preu al supermercat Corteingles.
Condis	Numeric	Preu al supermercat Condis.

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 de la web www.soysuper.com correspon a la companyia Soysuper, S.L. una empresa construïda per la diversitat on els aspectes en comú són: estalviar temps i diners, facilitar-se la vida a l'hora de comprar i poder dedicar-hi aquest temps a coses més importants. Donar les gràcies per posar a la nostra disposició la seva web i tecnologies que ens han facilitat el treball.
   Un enllaç on analitza el mateix web correspon al següent:
   https://riull.ull.es/xmlui/bitstream/handle/915/25438/Sistema%20de%20informacion%20para%20la%20recopilacion%20y%20centralizacion%20de%20informacion%20sobre%20productos%20alimenticios.pdf?sequence=1&isAllowed=y

   Per tal d'actuar a favor dels principis ètics i legals s'ha comprovat el fitxer robots.txt. il'scraping s'ha realitzat de forma que no hi hagin masses sol·licituds per unitat de temps evitant així que es col·lapsi la web.

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.

   Trobem aquest projecte interessant perquè no cal recórrer tantes webs com supermercats, sinó que ho tens tot en una mateixa web on simplement hi has de navegar per treure tota la informació. Posa a l'abast els supermercats més reconeguts a nivell nacional pel que farà que hi hagi una gran varietat per escollir, i el més important de tot, saber els diners que pots estalviar. Gràcies a les dades obtingudes podrem realitzar futures comparacions en un altre moment

temporal, establir uns preus mitjans per a cada producte, descartar supermercats més cars per defecte, entre d'altres.

Podríem respondre a les següents preguntes:

- Quin supermercat té els preus més cars de mitja?
- Quant varia el preu per setmanes?
- Cada quant varia el preu del producte?
- Quins són els preus més baixos?
- Val la pena comprar sempre en el mateix supermercat?

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).

   La llicència utilitzada pel dataset resultant correspon a: GNU General Public License v3.0.
   És una llicència de software lliure que permet l'ús i redistribució de forma gratuïta. S'ha escollit aquesta llicència per tal de que permeti un ús acadèmic actual i futur en cas de ser necessària la nostra informació extreta, ja que permet modificacions. Aplicant el Copyleft, ens assegurem que en pràctiques futures, els arxius modificats preservaran les mateixes llibertats.

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

   El codi es troba detallat en el següent enllaç de Git:
   https://github.com/RicardPI/SoySuperScraper

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

    El dataset s'ha publicat a Zenodo amb DOI:

```
10.5281/zenodo.7316935
```

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.

    https://drive.google.com/file/d/1XtOrwVLSjVy36he6NajHuIdadyvccLB/view?usp=share_link

CONFIRMACIÓ DEL GRUP:

Contribucions	Signatura
Investigació prèvia	RP, AJ
Redacció de les respostes	RP, AJ
Desenvolupament del codi	RP, AJ
Participació al vídeo	RP, AJ