

Anàlisi comparativa de Preus de Supermercat: Un Conjunt de Dades Extret a Partir de Web Scraping

Pràctica 2

Enric Sintes Arguimbau i Carlos Romero Matarin

Tipologia i cicle de vida de les dades -UOC

# Memòria del Projecte: Anàlisi Comparativa de Preus de Supermercat

**Integrants del Grup:**

* Carlos Romero Matarin
* Enric Sintes Arguimbau

**Enllaç al Lloc Web Elegit:**

* Dia: <https://www.dia.es/>
* Consum: <https://www.consum.es/>
* Mas: <https://www.supermercadosmas.com/>

**Enllaç al repositori amb el codi de la pràctica:**

<https://github.com/cromeroUOC/Analysis-Web-Scraping-Supermarket>

Arxiu comprimit del repositori:

**Enllaç al dataset publicat a Zenodo:**

**Enllaç al vídeo de presentació de la pràctica:**

* Carpeta de la pràctica:
* Vídeo de la pràctica:
* Vídeo comprimit de la pràctica:

# Taula de continguts

[**Memòria del Projecte: Anàlisi Comparativa de Preus de Supermercat 2**](#_fi82ghuuqeeh)

[**Taula de continguts 3**](#_w5z33jfu6zwl)

[**1. Context: 4**](#_d7qhul98k14h)

[**2. Títol 4**](#_j09ow7tria0n)

[**3. Descripció del dataset. 5**](#_s588beu31a7r)

[**4. Representació gràfica. 7**](#_wdx487kvh2m3)

[**5. Contingut: 8**](#_8vae62je1m8i)

[**6. Propietari 9**](#_nurgdo6hkep0)

[**7. Inspiració 11**](#_vhtxgbr56g60)

[**8. Llicència 12**](#_uacwxzb2k3f6)

[**9. Codi 12**](#_z3lmv9p5gqei)

[**10. Dataset 13**](#_8ez752jarrgl)

[**11. Referències 14**](#_3q4miz4k00mp)

[**Taula de contribucions 14**](#_1u8j2lf66btk)

# Descripció del dataset:

**Context:**

El conjunt de dades que hem utilitzat conté dades exhaustives sobre els productes que estan disponibles a les botigues en línia de Consum, Dia i Mas. Podem veure que cada entrada conté informació important per un anàlisis de les tendències del productes y la seva classificació. A més, en alguns enllaços s'hi inclouen les categories i subcategories dels productes, però aquestes dades no són completes en tots els casos. El conjunt de dades conté 25.625 entrades que cobreixen una àmplia gamma de productes.

Dir que aquest conjunt de dades és essencial per proporcionar una comprensió completa de les estratègies de preus dels supermercats en línia. La comparació directa entre els tres supermercats permet fer diferents anàlisis per avaluar les diferències entre les cadenes en termes de preus, disponibilitat de productes i preferències.

**Variables del dataset:**

* Nombre: Nom del producte.
* Marca: Marca del producte.
* Precio: Preu total del producte.
* Supermercado: Nom del supermercat on es pot comprar el producte.
* URL: Enllaç a la pàgina del producte en la web del supermercat.
* Fecha: Data quan es va recollir la dada.
* Hora: Hora quan es va recollir la dada.
* unidad: Unitat en la qual es ven el producte.
* precio\_unidad: Preu per unitat del producte.
* Categoria: Categoria del producte segons el supermercat
* Subcategoria: Subcategoria del producte segons el supermercat.
* Estado: Estat del producte en la web, es pot comprar o no.

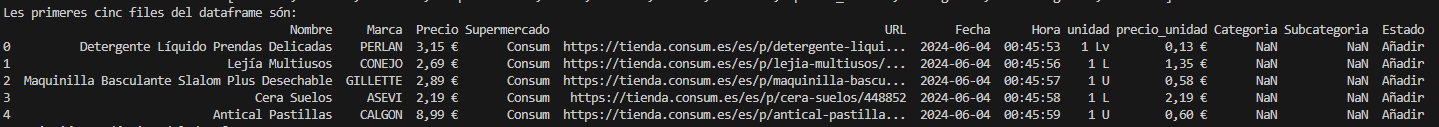
**Descripció del dataset:**

El conjunt de dades de productos.csv té 25.625 entrades, cadascuna de les quals representa un sol producte disponible per a la compra. Aquestes dades ofereixen una visió general del mercat actual en incloure cinc característiques clau: nom, marca, preu, supermercat, URL, data de obtenció, hora d’obtenció , unitat de mesura, preu unitari de mesura, categoria, subcategoria y estat.

Texto

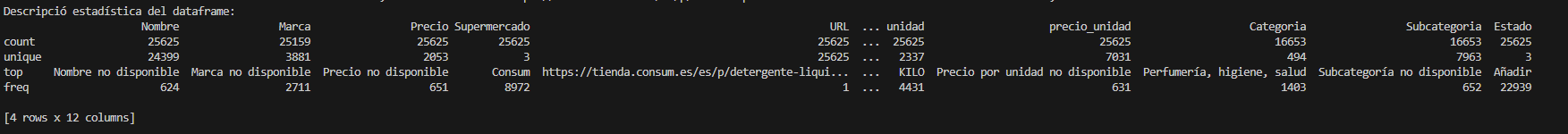
Descripción generada automáticamente

Primeres Cinc Files:



El quadre de dades comença amb una varietat d'articles, com detergents, lleixiu, maquinetes d'un sol ús, cera per a terra i pastilles anticalç. Aquesta primera mostra la varietat de productes disponibles al supermercat Consum, amb informació detallada sobre la marca, el preu i enllaços directes als productes, també podem veure com hi ha registres sense categoria.

Descripció Estadística:



Com hem comentat el conjunt de dades conté 25.625 entrades registrades, on cada una representa un producte únic. Podem veure que hi ha 24.399 noms de producte diferents, destacant la diversitat i amplada d'assortiment disponible.

Dir que encara que la majoria de marques són identificables, un total de 2.711 productes són etiquetats com a "Marca no disponible", reflectint possibles deficiències o inconsistències en la recopilació de dades. A més, 651 preus estan etiquetats com "no disponibles", indicant que aquestes entrades poden necessitar una verificació o actualització de dades.

Amb més cal dir que hi ha 7.031 preus per unitat únics i preus que varien àmpliament, per això podem dir que el conjunt de dades ofereix dades per a l'anàlisi de tendències de preus i comportaments de compra dins del context dels tres supermercats inclosos, destacant especialment Consum, on s'han registrat la majoria de les entrades.

**El nombre de productes per supermercat:**

Gráfico, Gráfico de barras

Descripción generada automáticamente

vUna captura de pantalla de un celular

Descripción generada automáticamente con confianza media

El gràfic mostra que el supermercat amb més productes registrats és "Consum" (8.972), seguit de prop per "Dia" (8.840) i "Supermercados Mas" (7.813). Segons aquesta distribució, "consum" i "dia" són els principals proveïdors de productes de la base de dades.

**El Top 10 dels productes:**

Gráfico, Gráfico de barras

Descripción generada automáticamente

El gràfic mostra els deu articles més comuns a la col·lecció, cosa que dona una idea de quins articles són més comuns o populars als supermercats.

# Integració i selecció:

S'ha optat per modificar i ampliar el dataset original utilitzat a la Pràctica 1 per a l'activitat d'integració. El desig principal d’aquesta ampliació es proporcionar una base de dades més completa i detallada és allò que justifica aquesta decisió. Podem veure que aquest conjunt de dades es capaç de suportar anàlisis més profundes i variades sobre el comportament dels preus entre diferents cadenes de supermercats.

Dir que la limitació de la varietat i la profunditat de les dades inicials fa que calgui ampliar la les dades original. Com un exemple ràpid dir que hem millorat significativament la qualitat de l'anàlisi en afegir característiques com "Categoria", "Subcategoria" i "Estat" als productes, cosa que permet segmentacions i comparacions més precises entre els productes. Podem dir que és especialment útil als estudis de mercat on la categorització dels productes pot afectar la percepció del preu i les decisions de compra dels consumidors.

Per fer l'ampliació del conjunt de dades i incloure noves variables com "Categoria", "Subcategoria" i "Estado", s'ha modificat el script original de Python utilitzat per a l'extracció de dades via web scraping. El que hem desenvolupat és que les modificacions recullin informació addicional de les pàgines de productes. El script actualitzat està disponible en el repositori GitHub del projecte.

# Neteja de les dades:

# Anàlisi de les dades

## 4.1. Aplica un model supervisat i un no supervisat a les dades i comenta els resultats obtinguts.

## 4.2. Aplica una prova per contrast d'hipòtesi. Tingues en compte que algunes d'aquestes proves requereixen verificar prèviament la normalitat i homocedasticitat de les dades.

# Representació dels resultats:

# Resolució del problema:

# Codi

# Referències

* Selenium IDE · Open source record and playback test automation for the web. (n.d.). <https://www.selenium.dev/selenium-ide/>
* Beautiful Soup Documentation — Beautiful Soup 4.12.0 documentation. (n.d.). <https://www.crummy.com/software/BeautifulSoup/bs4/doc/>
* Introduction to HTML - Web Scraping Overview. (n.d.). DataCamp. <https://campus.datacamp.com/courses/web-scraping-with-python/introduction-to-html?ex=1>
* A. (n.d.). *GitHub - alexandersotoc/supermarket-web-scrapping: Scrapping data from https://www.plazavea.com.pe website*. GitHub. <https://github.com/AlexanderSotoC/supermarket-web-scrapping>
* I. (2023, September 4). How To Make The Right Choice Of Web Scrapers To Collect Grocery Store Prices Data. Medium. <https://iwebdatascrapingservices.medium.com/how-to-make-the-right-choice-of-web-scrapers-to-collect-grocery-store-prices-data-6f2117fd328>
* Automate the Boring Stuff with Python. (n.d.). <https://automatetheboringstuff.com/chapter11/>

# Taula de contribucions

|  |  |
| --- | --- |
| **Contribucions** | **Signatura** |
| Investigació prèvia | CRM, ESA |
| Redacció de les respostes | CRM, ESA |
| Desenvolupament del codi | CRM, ESA |
| Participació al vídeo | CRM, ESA |

CRM: Carlos Romero Matarin

ESA: Enric Sintes Arguimbau