

**Gustavo Salgado Ocampo
Carlos Andrés Rivera Rivera**

Practica 1:

Comparador de Precios de Supermercados utilizando Web Scraping

DOI: 10.5281/zenodo.7859421

Tipología y ciclo de vida de los datos

Tabla de Contenido

Contexto.....	3
Título.....	3
Descripción del conjunto de datos.....	3
Representación gráfica.....	4
Contenido.....	4
Propietario.....	5
Inspiración.....	5
Licencia.....	6
Código.....	6
DataSet.....	7
Video.....	8
Contribuciones.....	8

Contexto

Este proyecto se desarrolló para comparar y analizar los precios de productos en dos supermercados diferentes que operan en Colombia. La información se recolectó a través de web scraping de los sitios web de estos supermercados, ya que proporcionan precios actualizados y detallados de los productos.

Los sitios web de los supermercados son las siguientes:

- <https://www.exito.com/>
- <https://www.tiendasjumbo.co>

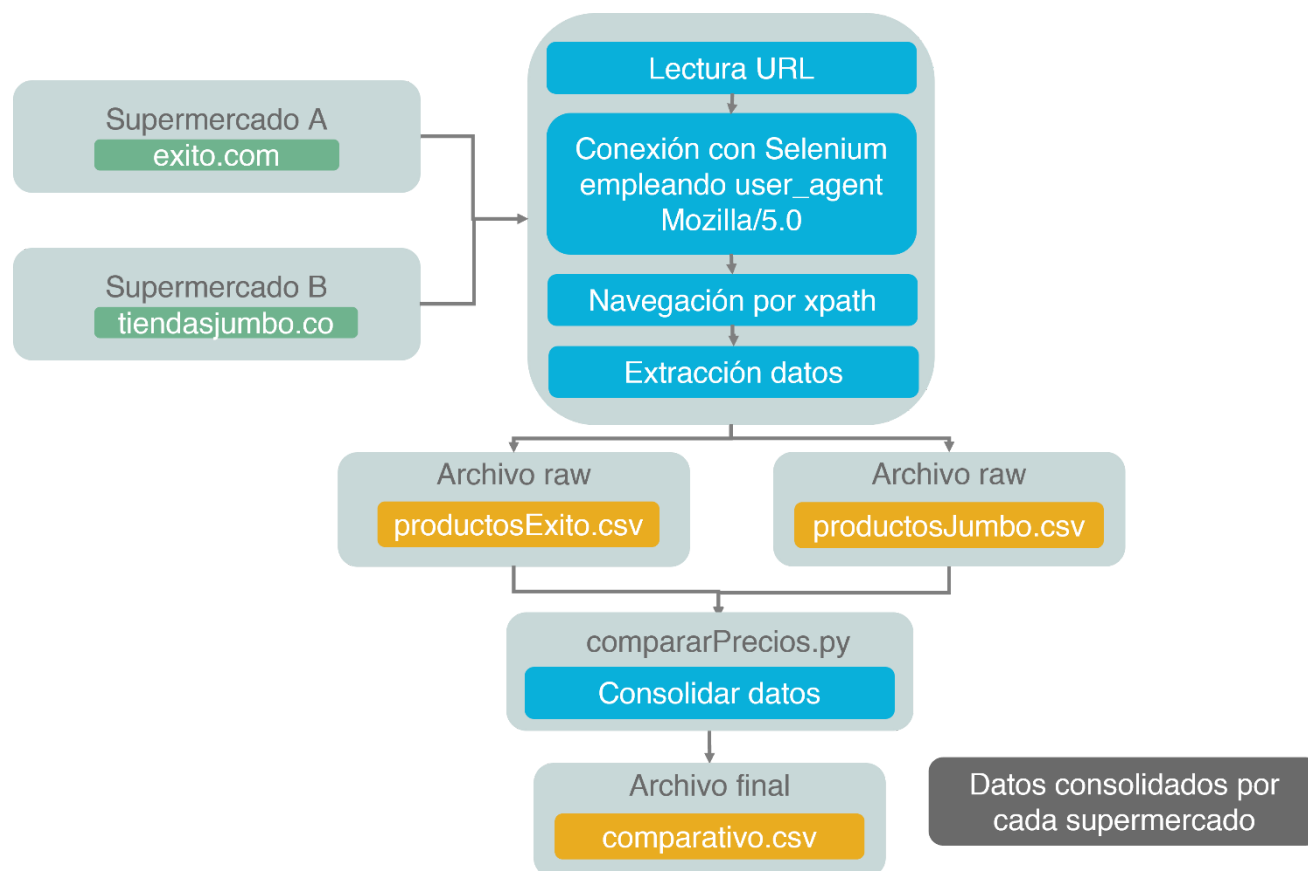
Título

Comparador de precios de supermercados.

Descripción del conjunto de datos

El conjunto de datos contiene información sobre los precios de productos en los supermercados Éxito y Tiendas Jumbo para una categoría específica de productos Licores y Vinos. Los datos incluyen el código del supermercado, nombre del producto, descripción del producto, información sobre su cantidad (si la tiene), su precio, descuento (si lo tiene) y la fecha de extracción de la información.

Representación gráfica



Contenido

El dataset incluye los siguientes campos:

- Supermercado: A para el supermercado Éxito y B para el supermercado Tiendas Jumbo.
- Nombre: Nombre del producto
- Descripción: Detalle del producto extraído de la información publicada por cada producto.
- Mililitros: Unidad de Medida del producto; este campo puede no estar presente en todos los casos y depende mucho si la descripción del producto contiene algún estándar de nomenclatura.
- Precio: Precio del producto
- Descuento: Porcentaje de descuento que tiene el producto, si no tiene puede venir vacío o en cero.

- PrecioConDescuento: Precio que incluye la aplicación del descuento.
- FechaHoraScraping: Marca de tiempo en formato Año-Mes-Día-Hora-Minutos-Segundos del momento en que se hizo la captura de los datos.

Propietario

Los datos pertenecen a los supermercados Exito y Tiendas Jumbo en Colombia. Se han seguido las políticas y directrices de estos supermercados en términos de uso de datos, así como las leyes locales aplicables al web scraping. Además, se ha prestado atención a los principios éticos y de privacidad, evitando la recolección de información personal o sensible.

Inspiración

Este conjunto de datos es interesante porque puede ayudar a los consumidores a encontrar los mejores precios en los supermercados y también puede revelar patrones y tendencias en los precios de los productos. Las preguntas que se pretenden responder con este conjunto de datos incluyen:

- ¿En qué supermercado se pueden encontrar precios más bajos para una categoría específica de productos?
- ¿Cómo fluctúan los precios de los productos en ambos supermercados a lo largo del tiempo?
- ¿Existen patrones estacionales o de otro tipo en los precios de los productos en estos supermercados?
- Se pueden comparar los resultados de este análisis con estudios anteriores o similares para obtener una visión más completa de las tendencias y fluctuaciones en el mercado de productos de consumo.

Licencia

Se considera adecuado la licencia Creative Commons Attribution 4.0 International (CC BY 4.0); esta licencia permite a otros compartir, copiar y redistribuir el material en cualquier medio o formato, y adaptar, remezclar, transformar y construir sobre el material para cualquier propósito, siempre que se otorgue el crédito apropiado y se indique si se realizaron cambios.

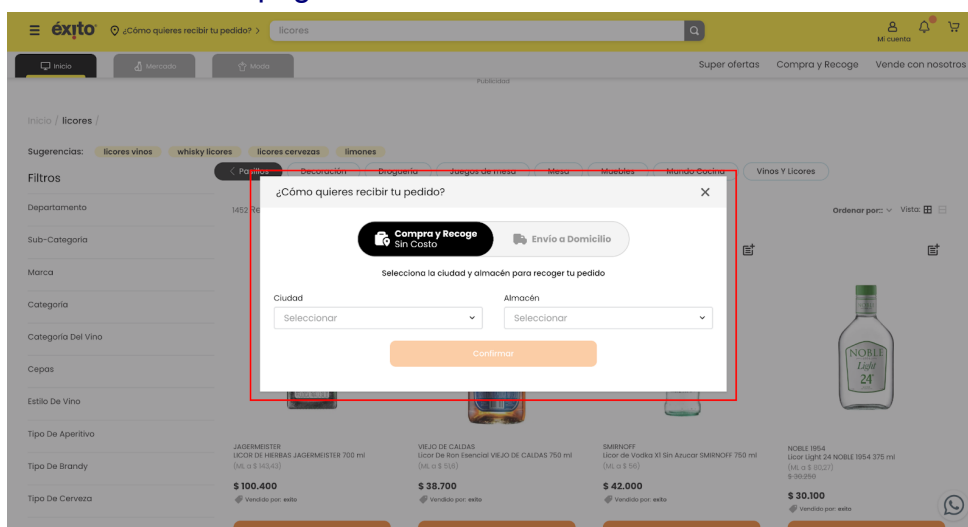
Código

El código construido para la ejecución de web scraping, se realizó en lenguaje python usando librerías de webdriver y selenium.

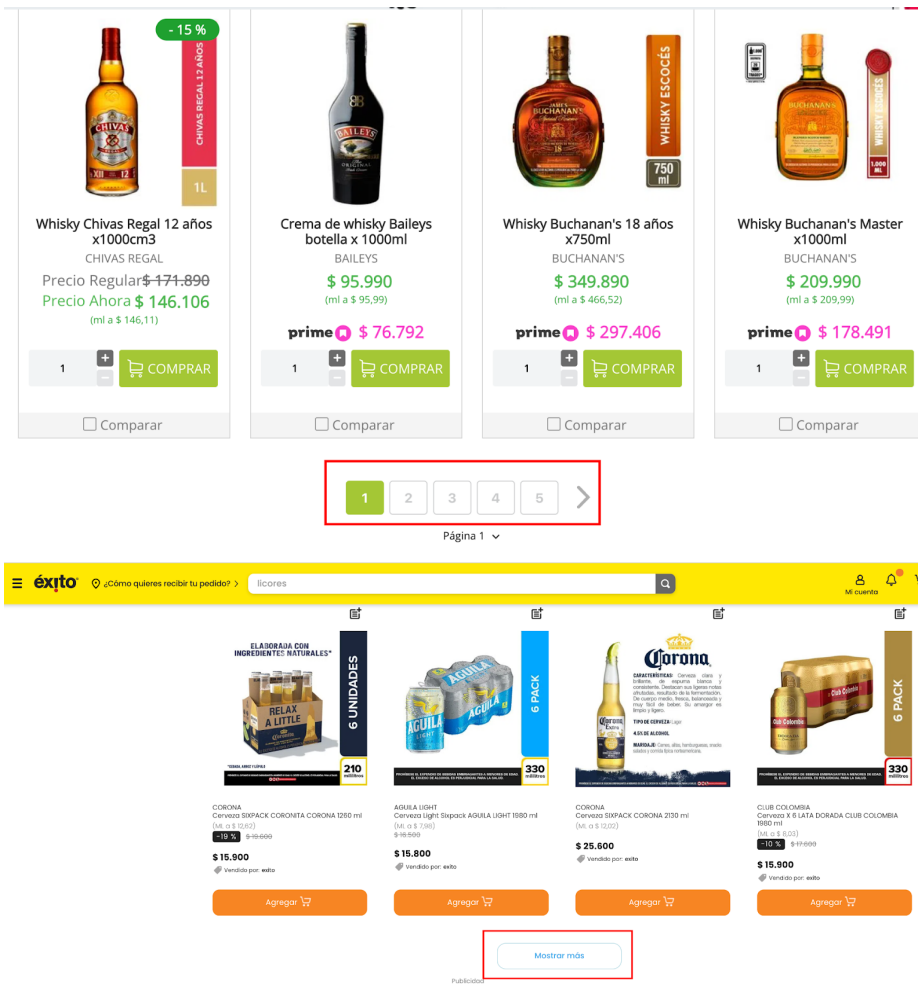
Link del repositorio: <https://github.com/Andrex2040/tiendasPriceScraper>. En el README.MD, del repositorio se puede encontrar información más detallada del proyecto.

Dificultades presentadas: Como se mencionó anteriormente, se eligieron un par de supermercados en Colombia para la exploración con web scraping. Al principio usamos la librería **scrapy** para extracción de datos, pero nos encontramos con algunos problemas:

1. En la tienda de Exito.com, de entrada nos encontramos con un modal el cual no deja interactuar con la página.



2. La información que se muestra en cada una de las páginas, se muestra distribuida en paginadores, los cuales es complejo hacer una interacción usando scrapy.



Whisky Chivas Regal 12 años x1000cm3
CHIVAS REGAL
Precio Regular \$ 171.890
Precio Ahora \$ 146.106 (ml a \$ 146,11)

Crema de whisky Baileys botella x 1000ml
BAILEYS
\$ 95.990 (ml a \$ 95,99)
prime \$ 76.792

Whisky Buchanan's 18 años x750ml
BUCHANAN'S
\$ 349.890 (ml a \$ 466,52)
prime \$ 297.406

Whisky Buchanan's Master x1000ml
BUCHANAN'S
\$ 209.990 (ml a \$ 209,99)
prime \$ 178.491

Página 1

Mostrar más

3. En el caso de las dos páginas, los resultados de productos cargados tienen implementado "lazy loading" (carga perezosa en español), por lo cual es necesario interactuar con la página haciendo scroll con el mouse para que carguen los productos por completo. Esta interacción también se nos complicaba usando scrapy unicamente.

Por las razones dadas anteriormente, decidimos usar **Selenium**, el cual nos permite interactuar con los elementos de una página simulando que es un humano el que lo realiza.

DataSet

Link: <https://zenodo.org/record/7859422#.ZEahonaZOUk>

Video

Link:

https://drive.google.com/file/d/1I4sSf_raihWw2uuz04KfPjbMW31ls5n1/view

Contribuciones

Contribuciones	Firma
Investigación Previa	Gustavo Salgado Ocampo, Carlos Andres Rivera
Redacción de las respuestas	Gustavo Salgado Ocampo, Carlos Andres Rivera
Desarrollo del código	Gustavo Salgado Ocampo, Carlos Andres Rivera
Participación en el video	Gustavo Salgado Ocampo, Carlos Andres Rivera