BAX 442 Data Design & Representation

Unlocking Business Insights: Web Scraping for E-commerce Health Products (HealthWarehouse.com)

Data Design & Representation Team

Sakshi Arya, Stuti Shekhar, Vivienne Xiang

Word count: 1935

Table of Content

1. Executive Summary

The project aims to delve into *'How price information of medicines varies across OTC categories and impacts individuals with no insurance coverage'*. The growing market of e-commerce companies in the health sector has helped change the way of purchasing medicines and its availability. The study utilizes **Selenium** to extract product information from healthwarehouse.com, revealing potential disparities that impact affordability. Additionally, to ensure minimal data loss and maintain data consistency, the extracted result was stored in a **MongoDB** database.

Important insights into the dataset were offered by the extracted characteristics, which included title, price, description, variations, availability, and category in a products table. The analysis yielded information for over the counter medicines such as ibuprofen being a **cheap** non prescription medicine, which is also considered as the most commonly bought drug in the market. The 'availability' data for products showed a notable out-of-stock ratio for popular over-the-counter medications such as in category Cold. It also demonstrated the wide range of average prices for drugs across different categories, spanning from$4 to $250. This variation could pose a challenge for individuals facing financial constraints.

2. Background Context

The diverse E-commerce market offers a wide variety of products delivered right to your footstep with just a few clicks. Being in the healthcare sector, the company operates in a highly competitive domain that is characterized by pricing, product availability, insurance and user experience. This all converges to a common goal of customer acquisition and retention. However, the policies of a company significantly impact the individuals categorized by their insurance coverages.

HealthWarehouse is a platform that facilitates the online purchase of healthcare products and services. They span a diverse range including prescribed Rx medications, Diabetes supplies, OTC products and

Home Medicals. The website features tabs showcasing products with important header details, sorted based on bestseller medications. To enhance customer convenience, the team also offers customer support services, ensuring a seamless user experience. Diving deep into the data of Over the Counter medicines,which are accessible without a prescription, this project aims to analyze price discrepancies for everyday-use medicines.

3. Methodology

3.1 Data Source

The primary data source for this project is *https://www.healthwarehouse.com/,* a leading provider of affordable and digital pharmacy services in America's online health market. The company's strategic use of technology and sourcing aims to simplify the supply chain, reducing intermediary costs.. The website features an organized list of products categorized by their nature and type of medication.

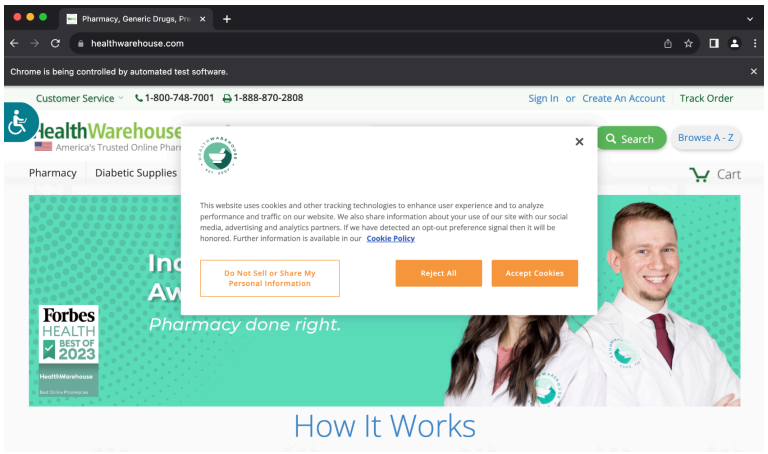
Iterating over multiple pages for data collection, the project gathers information about medicines. It stores from various **categories** such as: Cold - 10 pages, Pain & Fever - 20 pages, Gas Relief - 3 pages, Antacid - 5 pages, Antidiarrheal - 3 pages, Allergy - 10 pages, Haircare - 8 pages, Ointments 14 pages. In total, the project iterates over **67 pages**, with each page containing 20 items. With this comprehensive method, a wide variety of healthcare goods are included in the dataset for analysis.

3.2 Web Scraping Routine

The project utilizes Selenium to initialize the web driver for automated browsing and interaction with web elements, helping in capturing the page content. The team uses the request.url method to access specific URLs and retrieve information dynamically. Controlled delays and intervals are implemented using time.sleep() to ensure smooth navigation and mimic human behavior.

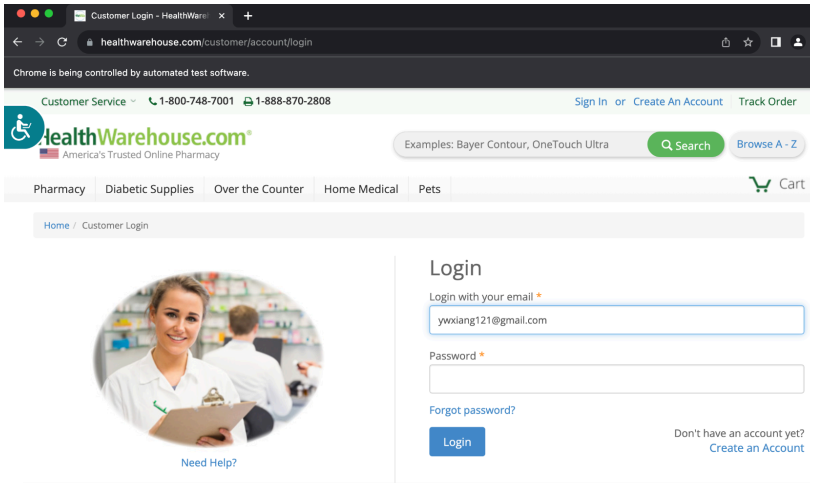
3.2.1 Calling the Chrome Driver and Rejecting Cookies:

The web scrape project starts with importing all the necessary libraries and calling the **chrome driver** with the help of *webdriver.Chrome()*. The cookies dialog box always pops up whenever a user enters the website, which is rejected by automating the program and clicking on the 'Reject All' Button. This is done by specifying the element with .XPATH and text associated with it ‘'Reject All'. The function .**click()** is used to click on the element and facilitates an interaction.



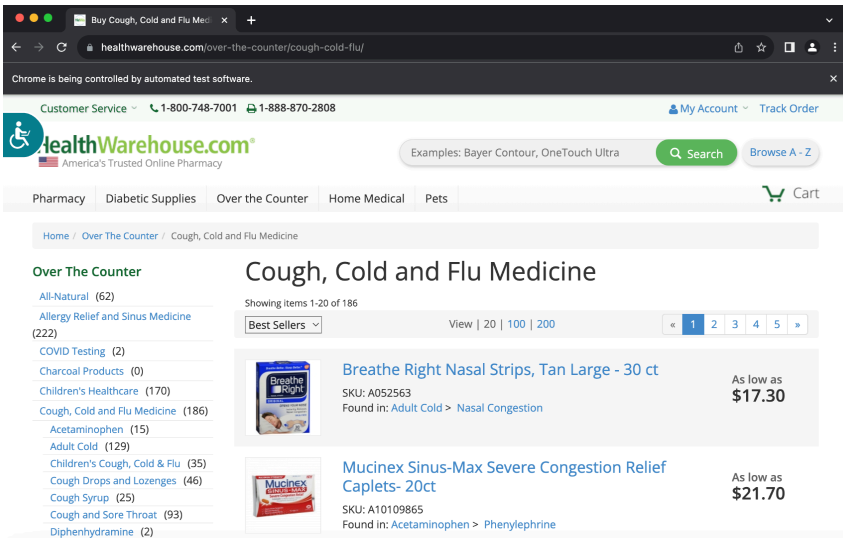
3.2.2 Logging in ‘healthwarehouse.com’ as a registered user

The user registers via the User Interface, and Python is subsequently programmed to log in to the page. The sign-in button is clicked by specifying text through the .XPATH Selector, followed by sending the email and password keys, identified by the elements with ids 'pemail' and 'ppasswd', respectively.



3.2.3 Navigating to the OTC tab with the help of driver

Using the driver, the user navigates to the OTC tab for 'Cough & Cold' medicines by utilizing .click(). Interaction on the page is facilitated through the identification of elements using static texts. The resulting page presents a list of medicine products, including their prices, reviews, SKUs, and other information. This valuable data can be stored in a product database, offering insights into product availability.



3.2.4 Observations on url changes:

The URL can trigger page changes by modifying the parameter **{page}** in the browser's address bar. Additionally, the URL is set to limit the quantity to 20 per page by default. The variable associated with sorting is "**sort**," which changes its value to "bestseller," "Newest," or "Alphabetical" when the filter is adjusted.

URL - f"https://www.healthwarehouse.com/over-the-counter/cough-cold-flu?limit=20&page={page}&sort=bestseller"

3.2.5 Saving HTML files for each page:

The GET request with the specified parameters, modifies the response to display the data and reloads the page. Upon receiving the modified response, the page source is parsed using BeautifulSoup and is stored in a HTML file. The filename for each file is dynamic, incorporating the category name and page numbers to ensure organized storage of the scraped data. Steps mentioned in 3.2.4 and 3.2.5 are iterated

over for other categories, ensuring homogeneity of data collection, along with a diverse line of products to do analysis on.

3.2.6 Extracting Information from html file:

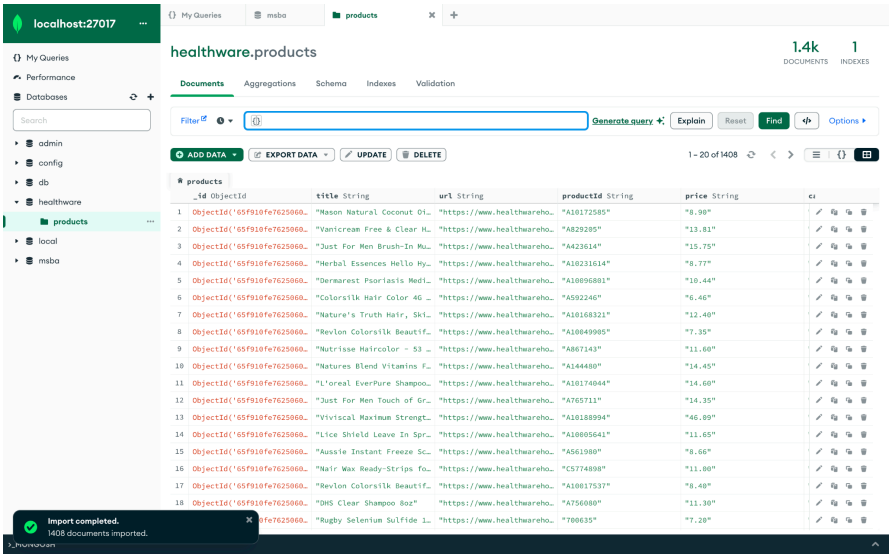
The html file is opened by checking if the directory exists and if the file ends with .html extension. All the files are iterated over and parsed using beautiful soup. Attributes such as title and link are extracted by specifying the **'h3'** element. If the item exists, the URL for the detailed product information is called, and other attributes such as price, SKU, description, variants, and availability are fetched by specifying elements and attributes. Conditions are applied using if statements to handle exceptions and null values.

```
Count: 0

Title: Mason Natural Coconut Oil Beauty Cream 2 oz.
Link: https://www.healthwarehouse.com/-2782755.html
ProductId: A10172585
Price: 8.90
Category: haircare
Availability: In Stock
Image: https://www.healthwarehouse.com/skin/frontend/hewa/newdesign/images/logo-top.png
Count: 1

Title: Vanicream Free & Clear Hair Styling Gel For Sensitive Skin 7oz
Link: https://www.healthwarehouse.com/-2740454.html
ProductId: A829205
Price: 13.81
Category: haircare
Availability: In Stock
```

Additionally, categories for each data is fetched using **regex** (regular expressions), which matches the pattern 'healthcare-(\w+)-page\d+.html' and assigns the word in between to the category. Here, it matches healthcare followed by word, page and /d indicating digits followed by .html.

3.2.7 Storing in Database

During the iteration process, the Python code stores all values as key-value pairs in a dictionary. This dictionary is then appended to a list called productDataList. The code initializes a connection to **MongoDB**, creating a database named 'healthware' with a collection named 'products'. The tag element are converted into url or string values to match compatibility with BSON before inserting the data.



3.3 Dataset Description

3.3.1 Database Choice

While implementing both MongoDB and mySQL were used, but MongoDB was selected for the following reasons:

1. **Speed in inserting and retrieval:** With 1408 items in the scraped item, mySQL requires much more processing time than MongoDB. Additionally, MongoDB proves to be faster for tasks like aggregation and filtering compared to MySQL, making it a more efficient choice for your project.
2. **Flexibility:** MongoDB offers more flexibility for data analysis, particularly when using Python, allowing a seamless workflow.

3.3.2 Dataset Dictionary

The dataset created consists of 9 attributes as described:

Columns	Description
Title	Name or title of the product
Link	Includes the URL link to the product's detailed page on the website
ProductId	Unique identifier for each product
Price	Price of each product,

Category	Category to which the product belongs to (Eg. Cough, Cold and Flu)
Availability	Availability Status (In stock/ Out of Stock)
Description	Detailed product Information
Image	URL link to the product's image
variant	Substitute for the product

4. Business Insights

4.2.1 How the dataset will help answer business questions

The dataset collected encompasses important features like d*rug name, id, price, category, description, availability, image, and variants* of each product. Through data analysis, insights can be uncovered such as the distribution of price among categories, and medicine demand trends. In addition, we can help pharmacies by offering strategic inventory management recommendations based on availability data. Furthermore, we can combine different features like price and variant, to develop alternative medicine plans for consumers from various income groups.

4.2.2 Design Choices and Business Value

In terms of design choices and business value, MongoDB was selected as the database system because of its superior performance in retrieval, filtering, and insertion operations. Its compatibility with Python also contributed to its flexibility.

The scraped data was stored in both **HTML format** and dictionaries. During web scraping, apart from capturing essential details such as product names and prices, comprehensive information like descriptions and variants was extracted, enabling deeper levels of data analysis

4.2.3 Recommendations

1. **Analysis for Pharmaceutical companies:**

   To sustain competitiveness in the market and enable strategy adaptation, pharma companies can have a constant monitoring plan for pricing changes gained from the **analysis of average prices**

grouped by categories. It is also possible to comprehend the impact that insurance coverage has on individuals, emphasizing how prices may be changed and outliers can be identified.

```
Category: antacid, Average Price: 14.49
Category: antidiarrheal, Average Price: 12.17
Category: digestive, Average Price: 11.59
Category: fever, Average Price: 13.27
Category: haircare, Average Price: 13.5
Category: cold, Average Price: 14.87
Category: allergy, Average Price: 17.71
Category: ointment, Average Price: 13.53
```

2. **Inventory Management Optimization for pharmacies:**

   Examining product availability in several categories, we provide pharmacists with suggestions for improving their inventory control methods. This includes giving pharmacists advice on how to effectively satisfy customer demand by keeping a careful eye on stock levels and quickly refilling drugs that sell quickly. Providing in **stock/ out of stock availability**, could help the stores in inventory management of most used medicines such as cold.

```
CATEGORY: antacid, In Stock: 89, Out of Stock: 11
CATEGORY: haircare, In Stock: 145, Out of Stock: 15
CATEGORY: fever, In Stock: 347, Out of Stock: 31
CATEGORY: ointment, In Stock: 255, Out of Stock: 25
CATEGORY: digestive, In Stock: 60, Out of Stock: 0
CATEGORY: antidiarrheal, In Stock: 40, Out of Stock: 0
CATEGORY: cold, In Stock: 160, Out of Stock: 27
CATEGORY: allergy, In Stock: 199, Out of Stock: 1
```

3. **Price Level insights for consumers:**

   For consumers, the focus could be to educate on the **misuse of ibuprofen drugs** for various purposes. Additionally promotion of variants with lower price could also be done to make the customer educated and not use the cheap OTC as a substitute. The main goal should be to have the same salt in medicine which could help in **actually treating the illness** rather than just subsiding the pain with ibuprofen. This could additionally help insurance companies also, since they would have to shell out less money for lesser variants.

5. Summary and conclusions

The project aims to help uninsured people in getting the appropriate medications that meet their healthcare needs. Insurance in the states plays a vital role in people having the access to care and even medicines. The escalating prices have made it challenging for uninsured individuals to afford comprehensive healthcare.

Our data analysis indicates significant price variations both across and within different categories, supporting these observations. The price analysis was conducted to assess its impact on individuals lacking insurance coverage. Findings reveal that while the **average price** stands at **$14.18**, prices can surge to **$258.14** for allergy medications and drop as low as **$4.54** for an Advil tablet. The affordability of ibuprofen, being the least expensive, also contributes to the concerning trend of Americans misusing over-the-counter medications, particularly as substitutes for Advil.

These trends may stem from several factors, including the affordability and non-prescription nature of certain medications, making them popular alternatives for pain relief across multiple conditions. Additionally, it also suggests that Advil has 52 numbers of products attached to it as variants, highlighting the broad range of painkillers.. Even though tylenol has more variants, the low price influences consumer behavior. Additionally, the data indicates that cold medicines exhibit the highest out-of-stock ratios, supporting their widespread usage.

References

Corky Siemaszko., (2018). Americans are abusing over-the-counter drugs as well as opioids, study shows https://www.nbcnews.com/storyline/americas-heroin-epidemic/americans-are-abusing-over-counter-drugs-well-opioids-study-shows-n846401

Ibuprofen misuse and overdose. *Reactions Weekly* **1837**, 326 (2021). https://doi.org/10.1007/s40278-021-88925-y

Selenium. https://www.selenium.dev/documentation/

MongoDB. https://www.mongodb.com/