HSLU Hochschule Luzern

Master of Applied Information and Data Science
Module Data Collection, Integration and Preprocessing

# Web Scraping of Grocery Pricing: A Comparative Study of Pasta, Rice, and Sauces Across Three Swiss Supermarkets


**Group 03**

**Group Members:**
Fatima Barcina (fatima.barcina@stud.hslu.ch)
Martina Diaz (martina.diaz@stud.hslu.ch)
Catalina Roth (catalina.barriosroth@stud.hslu.ch)

Hand in Date: 15 November 2024

# Table of Contents

# 1. Introduction, Motivation, Purpose and Scope

Recognizing the importance of this Data Collection, Integration and Preprocessing project as future data scientists, we began brainstorming ideas, seeking topics and websites that were of our interest with a practical and real-world focus.

After group evaluation of various proposals, we found a common motivation with clear purpose: analyze market strategies by examining price competitiveness and product diversity in the Swiss retail sector, with the aim of establishing a market comparison. To achieve this, we selected three of the main and popular Swiss supermarkets.

Migros, Switzerland's largest retail company, stands as the country's leading supermarket chain. Additionally, it ranks among the forty largest retailers worldwide. Lidl Switzerland AG is a Swiss retail company that operates a nationwide discount store network. As part of the German Lidl Stiftung & Co. KG, Lidl Switzerland is ranked among the 100 largest companies in the country.

This analysis and comparison are based on three basic and essential product categories: rice, pasta, and sauces (tomato and pesto) of the chosen supermarkets.

# 2. Research Questions

The research question that are addressed through the web scraping are:

*1. Which supermarket has the most competitive prices?*
*2. Which competitor offers more brands across distinct categories?*
*3. How much more expensive are own brands compared to traditional brands for each competitor*?

The approach to answering these three main questions and the analysis conducted are detailed in Chapter 3.3 of this report.

# 3. Methods

## 3.1 Data Collection

The data collection aimed at scraping all the products of the three categories pasta, rice and sauces through the two competitors' websites.

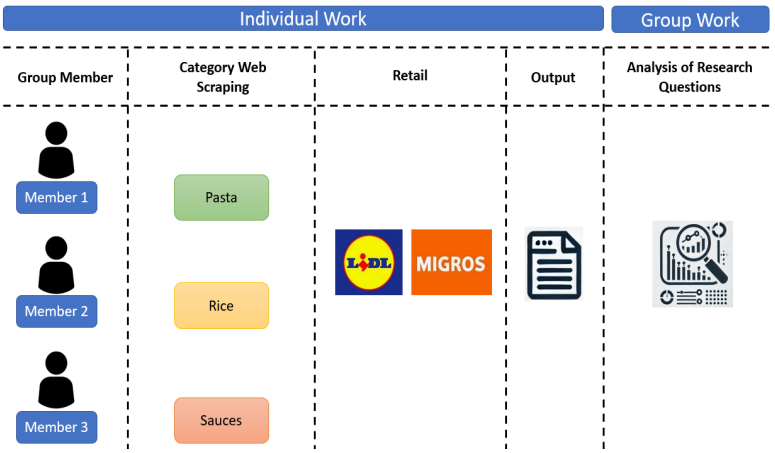


*Figure 1: Work Process for the web scraping*

The main tools used in the scripts are *Selenium*, *BeautifulSoup* and *Pandas*. The main functions used, and their application are listed below.

- Webdriver from Selenium allows to access webpages and interact with those through searching methods according to id, name, css_selector, or xpath.
- ActionChains from Selenium allows to perform actions on the web page, for example clicking on bottoms or interacting with sliders. In the Lidl website, for example, enable to scroll down the menu and access the following web pages with products. In the rice category for both competitors, the process starts from the initial webpage link and navigates through multiple links until reaching the rice section. For Lidl, however, this navigation becomes especially challenging (details provided later).
- BeautifulSoup is a library to extrapolate and parse the html of webpages, searching for elements according to tags, class names, or other attributes.
- Pandas is a tool to create and manage data frames. Its use in this project includes formatting the scraped data at the end of the data collection, the data preprocessing, and the data analysis.

Other tools used are, for example, the exceptions handling and the wait time pose offered by *Selenium*, and the datetime to register the data of the web scraping.

The final output obtained from the individual web scraping phase for each product category consists of a data frame with the following fields: ID, Competitor, Category, Product_Description, Brand, Regular_Price (CHF), Grammage, Unit, Link, Scraping_Date, Discount, Actual_Price (CHF), Regular_Price/Unit, Actual_price/Unit and a custom field according to the individual analysis.

## 3.2 Transformation Techniques

### 3.2.1 Cleaning, formatting and structure of the scraped data for the analysis

In the case of "Pasta Sauces" category, the data collection from Lidl website led to scraping all items under the categories labeled 'Konserven' on the website, which include also products different from pasta sauces, requiring a cleaning phase to remove the different items through keywords in the product descriptions.

To merge the data from the individual scraping activity, some conventions have been established, which required a phase of cleaning and formatting before merging the data frames.

In some cases, a cleaning stage of the **Product_Description** aimed at correcting character formats.

As regards **Grammage** and **Unit**, from the Lidl website it is possible to download strings from which extrapolate both values. For example, from a string like "pro 265g | 100g = 0.67 CHF". Furthermore, the quantities and units have been converted in Kg and L.

The **Actual_Price (CHF)** is calculated considering the regular price and the discount scraped from the webpage (if applicable). For the Rice category, the Regular Price is calculated using the Actual_Price (CHF) and the discount information scraped from the websites.

Again, in the case of "Pasta Sauces" category, a more specific formatting regarded the **Brand** as some of those are composed of more than one world that could be extracted from the product's description.

In case of no applicable discount in the scraping date, the **Discount** record has been registered as 'no discount'.

The final output obtained merging the different data frames consists of a data frame with the following fields:
ID, Competitor, Category, Product_Description, Brand, Regular_Price (CHF), Grammage, Unit, Link, Scraping_Date, Discount, Actual_Price (CHF), Regular_Price/Unit, Actual_price/Unit.

### 3.2.2 Missing Values

Following the methodology to handle missing values after check them, very low percentage of missing data across the dataset by each product category. Most missing values were easily filled by retrieving information directly from the websites.

As a result of this detailed data cleaning process, the final merged data frame from the three sources was completed with no missing values.

However, there were some special cases where specific formatting on the webpage affected the scraping process.
For example, in most cases, product weights or quantities were published as a number followed by a unit (e.g., "250g" or "1L"). However, in a few instances, quantities were presented as a multiplication (e.g., "2 x 500g"), which the general scraping code did not recognize, resulting in missing values for these items. In the rice web scraping, this specific case was eventually addressed to avoid filling NaN values.

Also mentioning that Brand field on the Lidl website: the brand name only appeared after clicking into each individual product page. Without doing so, the field would return as NaN. Given the volume of products, we addressed this by scraping brand information across all categories.

Additionally, in the "Rice" category, as example, three items had some essential information missing. However, by reviewing the product images with the product link, the missing fields were manually filled.

In the "Rice" category, an organized method has been implemented to fill missing fields column by column, using the input function.

## 3.3 Analysis

### 3.3.1 Address of Research Question

*1. Price competitiveness*

The analysis of price competitiveness focuses on determining which supermarket, between Migros and Lidl, offers the most competitive prices in selected categories (Pasta, Pasta Sauces, and Rice). To achieve this, both regular prices and current prices (including discounts) are examined, and a percentage comparison of the average prices of both competitors is performed by category. Additionally, prices have been categorized into defined ranges for each competitor and category, allowing for a more granular view of pricing structures. This methodology allows for an evaluation of prices not only in absolute terms but also in relative terms, providing a clear view of each competitor's pricing strategy.

Furthermore, to complement the price competitiveness analysis, the study includes which supermarket applies a higher percentage of discounts during the web scraping dates, offering further insights into the frequency and aggressiveness of each competitor's discount policies.

*2. Number of brands offered by category and competitor*

The question required, firstly, to create two subsets of the data frame according to the Competitor, including only the Category and the Brand fields. A second step calculate the number of different unique Brand for each competitor according to the category. Finally, the results are plot for each competitor and as comparison between the two competitors. Moreover, the number of products with unknown brand are counted for each category in the offers by the competitors.

*3. Price comparison between private label products and traditional brands*

To address this question, a new flag is created in the dataset to indicate, for each product, whether the brand corresponds to a private label or a traditional one. The objective is to calculate the average of both prices (Regular and Actual, including any discount) per category and competitor and to determine the price advantage and savings for a client when purchasing a private label product.

### 3.3.2 Exploratory Data Analysis

3.3.2.1. Statistics Summary

| Variable | Mean | Median | Mode | Min | Max |
|---|---|---|---|---|---|
| Regular_Price (CHF) | 3.14 | 2.99 | 2.95 | 0.39 | 10.90 |
| Actual_Price (CHF) | 3.10 | 2.95 | 2.95 | 0.39 | 10.90 |

*Table 1: Statistics Summary*

3.3.2.2. Boxplots for Prices per Category

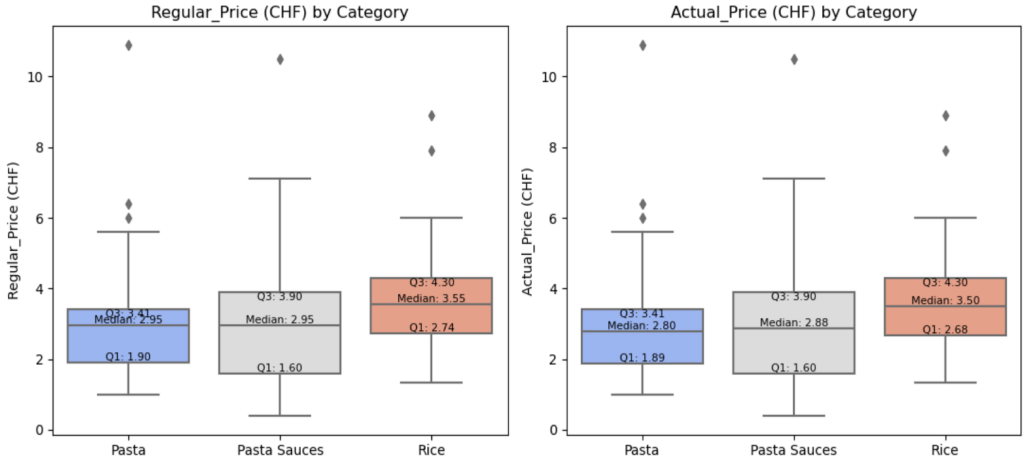


*Figure 2: Boxplot for regular and actual price according to the product category*

As it can be observed from the boxplot above, there are some outliers, however we don't have statistically evidence they were caused by measurement errors, so they will remain in the analysis to provide a better understanding of the data.

## 4. Results

### 4.1 Findings

#### 4.1.1 Results of Regular Average Price per category and Competitor

In general, there is no significant difference in the pattern of Regular Prices and Actual Prices for both supermarkets across the three categories.

The difference between the two supermarkets is slightly smaller when looking at Actual Prices compared to Regular Prices. The average regular price per category is higher for products offered by Migros.

Figure 3 (below) shows the price differences between the two competitors across the selected categories. In general, Lidl's prices appear more concentrated at lower values, indicating a competitive pricing strategy relative to Migros.

Migros consistently has higher average regular prices across all categories compared to Lidl, particularly for the "Pasta Sauces" category. In this category, Lidl has a notably high concentration of products in the lowest price range (0 to 5 CHF), focusing on offering affordable prices.

Therefore, Migros might offer a broader range of products, including higher-priced options, whereas Lidl's offerings are more concentrated in the lower price range, indicating competitive pricing strategy.

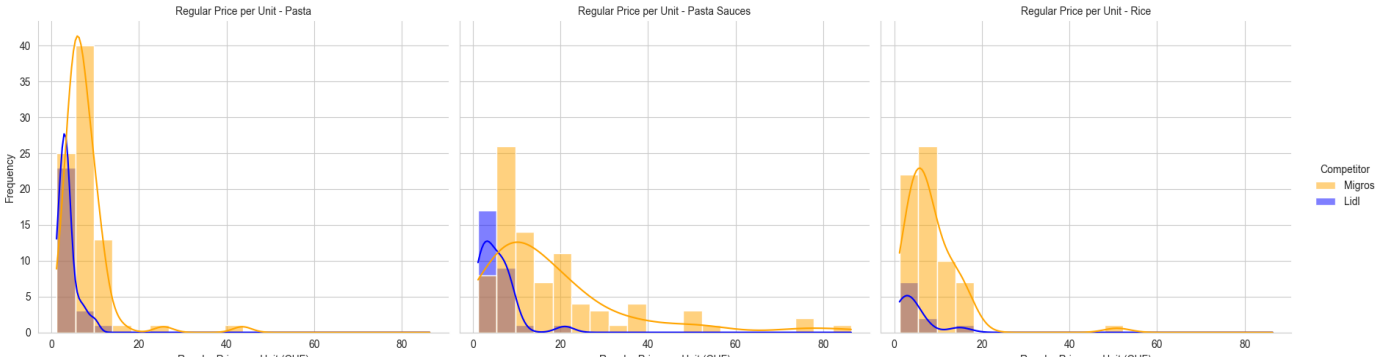


*Figure 3: Histogram - Distribution of Regular Price per Unit by competitor and Category*

#### 4.1.*2 Brand Diversity per category and Competitor*

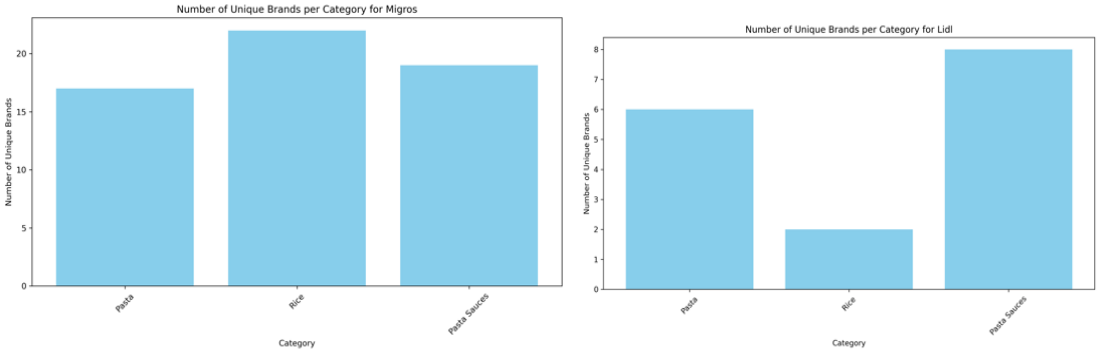


*Figure 4: Brand diversity per category and competitor*

Migros offers: 17 different brands of pasta, 22 different brands of rice, and 19 different brands of pasta sauces.
Lidl offers: 6 different brands of pasta (including 1 product with unknown brand), 2 different brands of rice, and 8 different brands of pasta sauces.

#### 4.1.3 Gap Price between Private Label and Third Brands

In Figure 5 below, a comparison is shown between the average Regular Price Per Unit and category for each competitor. For Lidl (left plot), it can be observed that, in all three categories, third brands are more expensive than private labels (own brands). The largest price gap is in the "Pasta Sauces" category (-6.61 CHF on average than third brands).

On the other hand, what stands out is that Migros (right plot) offers more competitive prices for third brands in "Pasta Sauces" (an average of +6.75 CHF than private labels), while for the "Rice" and "Pasta" categories, own brands lead in terms of price convenience.

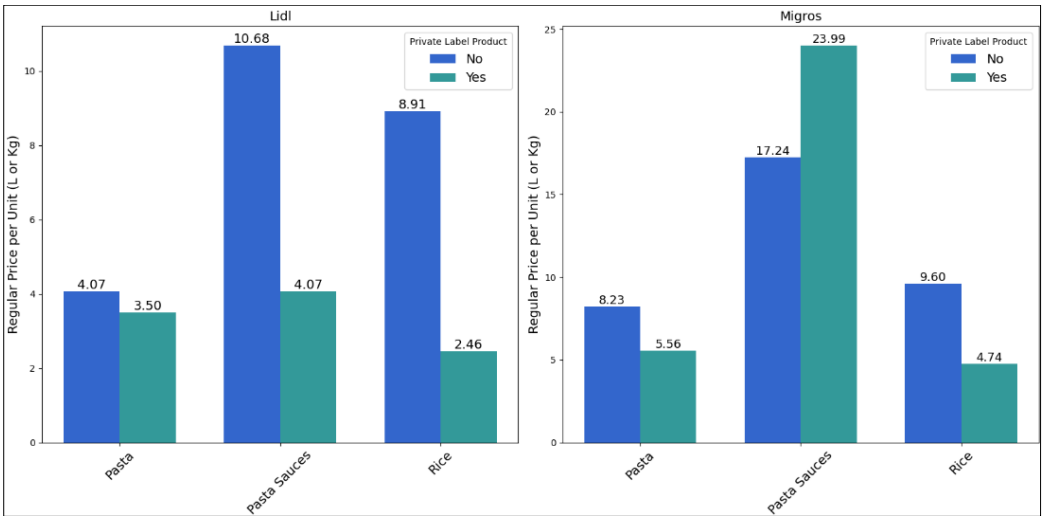


*Figure 5: Comparison Average Price Between Private Labels and Third Brands*

### 4.2 Outcomes

#### 4.2.1 Competitive Advantage

*1. Which supermarket appears to offer a more competitive pricings strategies across distinct categories (According Regular Prices and prices with discount)?*

The first part of the analysis Competitive Advantage is based on comparing Regular and Actual prices for both competitors by category, the tables and plots shows the following insights:

- **Pasta** (Regular Price and Actual Price)

Lidl maintains a consistent average price of 3.63 CHF for both regular and actual prices, indicating a stable pricing strategy in this category. Migros shows a slight reduction in average price from 7.54 CHF (regular price) to 7.45 CHF (actual price). Although a small discount is applied, the average price remains considerably higher than Lidl's.

The percentage difference between regular prices is 107.71% in favor of Lidl, which decreases slightly to 105.23% for actual prices. Lidl thus remains the more affordable option in this category.

- **Pasta Sauces** (Regular Price and Actual Price): In this category, the price difference between both supermarkets is notable.

Lidl has an average price of 5.25 CHF for both regular and actual prices, reflecting a competitive pricing strategy without significant discounts,

whereas Migros, reduces its average price from 18.04 CHF (regular) to 17.81 CHF (actual), applying a small discount in this category.
The percentage difference is significant in both cases: 243.62% for regular prices and 239.24% for actual prices, showing that Lidl remains considerably more affordable in this category.

- **Rice** (Regular Price and Actual Price):

Lidl maintains an average price of 4.40 CHF for both regular and actual prices, highlighting stability in its pricing strategy. On the other hand, Migro shows a small discount, with the average price moving from 8.42 CHF (regular) to 8.37 CHF (actual). Despite this reduction, the price remains significantly higher than Lidl's.

The percentage difference in regular prices is 91.36%, which decreases slightly to 90.23% for actual prices, making Lidl the more economical choice in this category as well.

As conclusion, Lidl leads with a more competitive pricing strategy across categories, with consistently lower prices regardless of discounts. Migros' higher prices suggest a broader or higher-quality product range, though less accessible for price-sensitive shoppers.

The second part of the competitive analysis focuses on the discount strategy by category for each supermarket. The data tables and charts shows: Migros includes 25 discounted items across its categories, representing the entirety of discounts, distributed by category Pasta: 12 items (14.81%), Pasta Sauces: 7 items (8.24%), Rice: 6 items (9.09%). while Lidl does not apply any discounts.

Consequently, Lidl's pricing advantage comes from maintaining consistently lower base prices across all categories whereas Migros uses discounts to enhance its competitiveness.

*2. Which supermarket offers more brands and how might this attract certain costumers who value more brand options?*

The comparison of the findings shows that Migros offers a larger brand selection for all the three products categories:

- for "Pasta" category, Migros offers approximately 283% of the number of unique brands offered in Lidl,
- for "Rice" category, Migros offers approximately 1100% of the number of unique brands offered in Lidl,
- for "Pasta Sauces" category, Migros offers approximately 237.5% of the number of unique brands offered in Lidl.

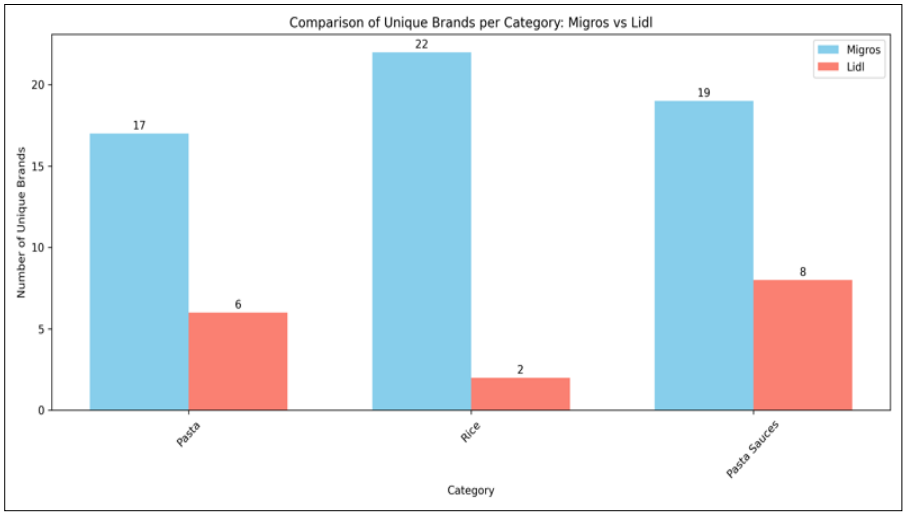


*Figure 6: Comparison between the number of unique brands per category for the two competitors*

### 4.2.2 Consumer Benefits

| | | | Count (Units) | | Avg Price_Regular/Unit (CHF/L-Kg) | |
|---|---|---|---|---|---|---|
| Variable | Category | % of Saving | Third Brand | Private Label | Third Brands | Private Label |
| Lidl | Pasta | -14.6 | 6 | 21 | 4.1 | 3.5 |
| Lidl | Pasta Sauces | -61.7 | 5 | 23 | 10.7 | 4.1 |
| Lidl | Rice | -71.9 | 3 | 7 | 8.9 | 2.5 |
| Migros | Pasta | -38 | 60 | 21 | 8.2 | 5.6 |
| Migros | Pasta Sauces | 39.5 | 75 | 10 | 17.2 | 24 |
| Migros | Rice | -51 | 50 | 16 | 9.6 | 4.7 |

*Table 2: Consumer Benefits*

In the table 2, it can be seen that Migros offers 47 private label products, while Lidl offers 51. Despite this small difference, the average savings from buying private labels instead of third brands is -16.5% for Migros and -49.4% for Lidl, respectively.

### 4.3 Observations

Unexpected pricing differences and brand variety were observed between the two competitors, where each retailer's approach appears to be different. Migros employs a marketing strategy that offers a wide range of brands for the diverse customer preferences, while Lidl appeals to budget-conscious clients through savings on private-label products.

## 5. Conclusions

### 5.1 Summary About Learning

Through this project, we gained valuable skills in web scraping and market strategies, regardless of each member's initial knowledge level. We have expanded our understanding of real-world web scraping using Beautiful Soup and Selenium, followed by data cleaning and transformation with pandas.

The goal was to apply as many tools and methods as possible through this practical case, which also allowed us to enhance and solidify our previous knowledge of Python. Facing and solving various challenges further enriched our skills and strengthened our knowledge with the Python environment.

### 5.2 Challenges

The challenges during the web scraping mostly concerned preventing block, navigating through buttons, and bypassing cookies messages.
In the first case, for example, scraping data from Migros required to split the number of products into chunks to bypass the block of the website scraping according to the IP address used by one of the group members.

As regards buttons, on the Lidl webpages, the button “Weitere Produkte laden” shows more products but at the same time access a different page, requiring a function to access the new page each time the button is clicked.

Another challenge for the “Rice” category on Lidl’s website involved navigating elements like as the "Sortiment" button, which would sometimes disappear based on the window size. This responsiveness complicated the web scraping, as key links became hidden when the window was narrow. As a solution, if the button is hidden due to the window width, the script uses an alternative approach by accessing the sitemap at the page’s footer.

Cookies settings required a function to bypass the scraping block, clicking the button to accept the cookies.

Another challenge has been bypassing errors while scraping data not available for certain products, for example as above-mentioned for the discount. In this case, a try and except function allowed to register missing values with a replacement (i.e. ‘no discount’, ‘unknown’, …).

The challenges in the data transformation and analysis included, for example, the calculation of the Actual_Price (CHF) and the analysis considering missing values.

In the case of the actual price calculation, an intermediate step required to convert the percentage of discount into a numeric value to apply it with respect to the regular price.

### 5.3 Potential Future Steps to Improve

The next improvement for this analysis would be to expand the web scraping to other markets to obtain a broader overview of price behavior in Switzerland's retail sector.

The web scraping can be complemented with machine learning models to make various price predictions. With this index, sales and demand estimates can be made, which may help improve the company's performance.