Ein Bild, das Schwarz, Dunkelheit enthält.

Automatisch generierte Beschreibung

**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**](mailto: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 2](#_Toc182059379)

[2. Research Questions 2](#_Toc182059380)

[3. Methods 2](#_Toc182059381)

[3.1 Data Collection (Martina) 2](#_Toc182059382)

[3.2 Transformation Techniques 3](#_Toc182059383)

[3.3 Analysis 3](#_Toc182059384)

[4. Results 4](#_Toc182059385)

[4.1 Findings 4](#_Toc182059386)

[4.2 Summary of key Data 4](#_Toc182059387)

[4.3 Outcomes 4](#_Toc182059388)

[4.3.1 Competitive Advantage 4](#_Toc182059389)

[4.3.2 Consumer Benefits 4](#_Toc182059390)

[4.4 Observations 5](#_Toc182059391)

[5. Conclusion 5](#_Toc182059392)

[5.1 Summary About Learning 5](#_Toc182059393)

[5.2 Limitations 5](#_Toc182059394)

[5.3 Potential Future Steps to Improve (Cata) 5](#_Toc182059395)

# 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

The methods and techniques for data collection, transformation, and analysis are explained below.

## **3.1 Data Collection (Martina)**

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

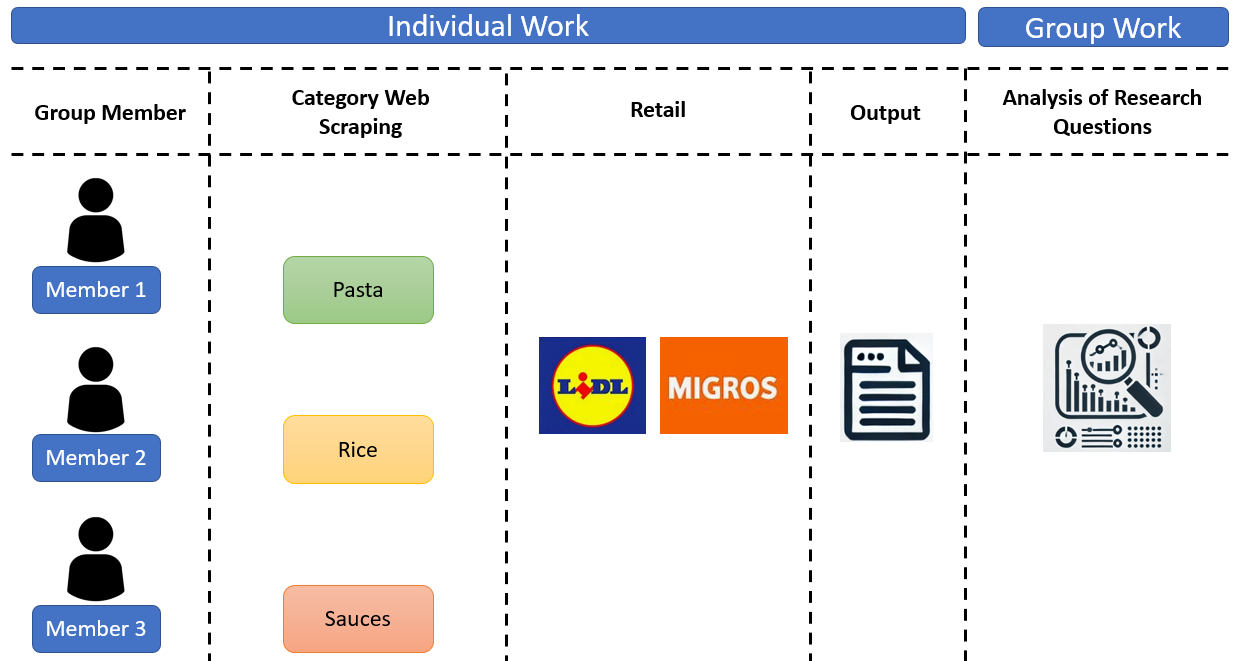


Figure 1: Work Process

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.
* *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**

1. **Cleaning, formatting and structure of the scraped data for the analysis** (Martina – Cata - Fatima)

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).

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.

1. **Missing Values** (strange things) (Fatima)

## **3.3 Analysis**

This section explains how the group addressed the research questions and set up an exploratory data analysis to gain a more comprehensive understanding of the dataset.

**3.3.1 Address of Research Question**

1. *Average Prices Comparison (Fatima)*
2. How we compared the prices (Average Regular prices – Actual prices)
3. *Number of brands offered by category and competitor (Martina)*

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.

1. *Price comparison between private label products and traditional brands (Cata)*

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** (Cata)

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

1. Boxplots for Prices per Category

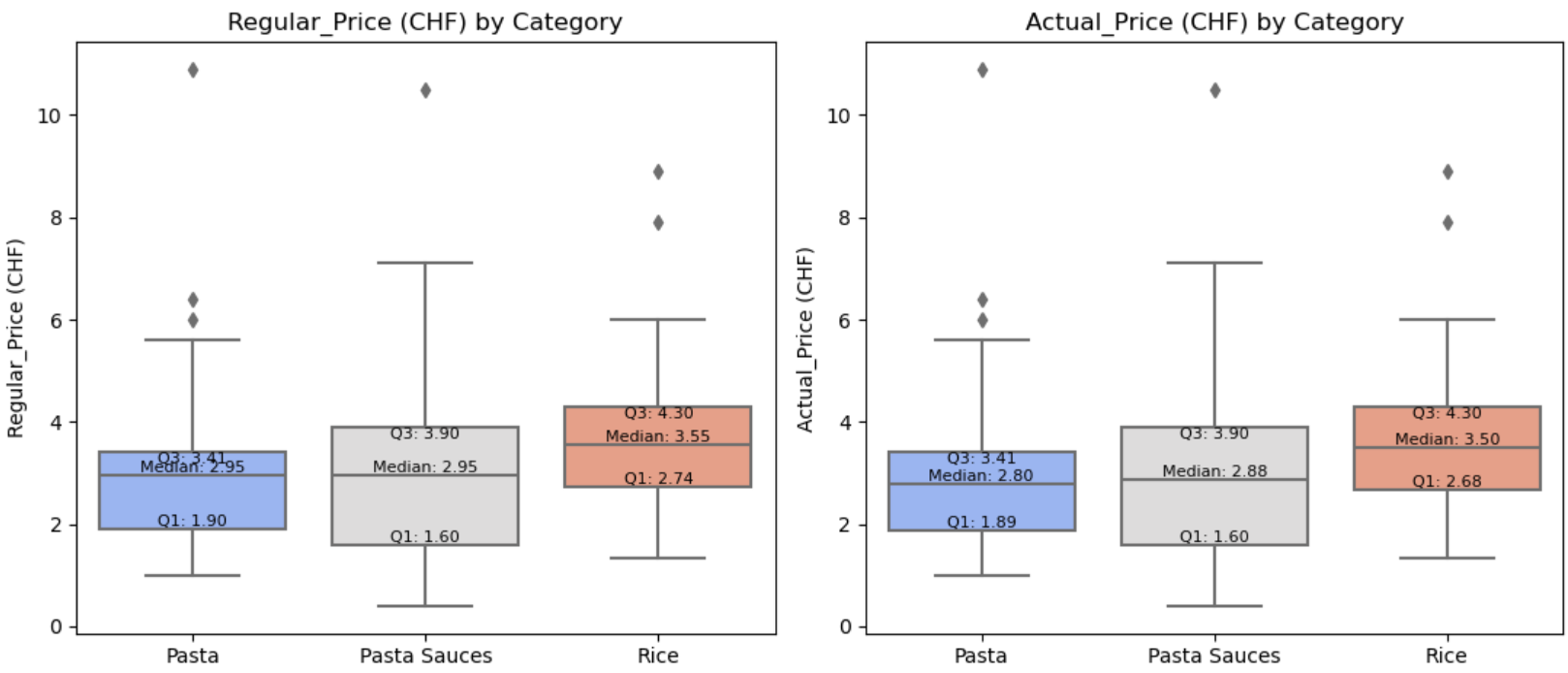


Figure 2: Boxplot for prices

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

# 4. Results

A summary and the main insights from the analysis is presented below.

## **4.1 Findings**

The plots for the analysis of each research question are shown next.

**4.1.1 Results of Regular Average Price per category and Competitor (Fatima)**

**4.1.2 Brand Diversity per category and Competitor (Martina)**

**A graph of blue rectangular shapes

Description automatically generated with medium confidence A graph with blue rectangles

Description automatically generated**

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 (Cata)**

In Figure 3 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.

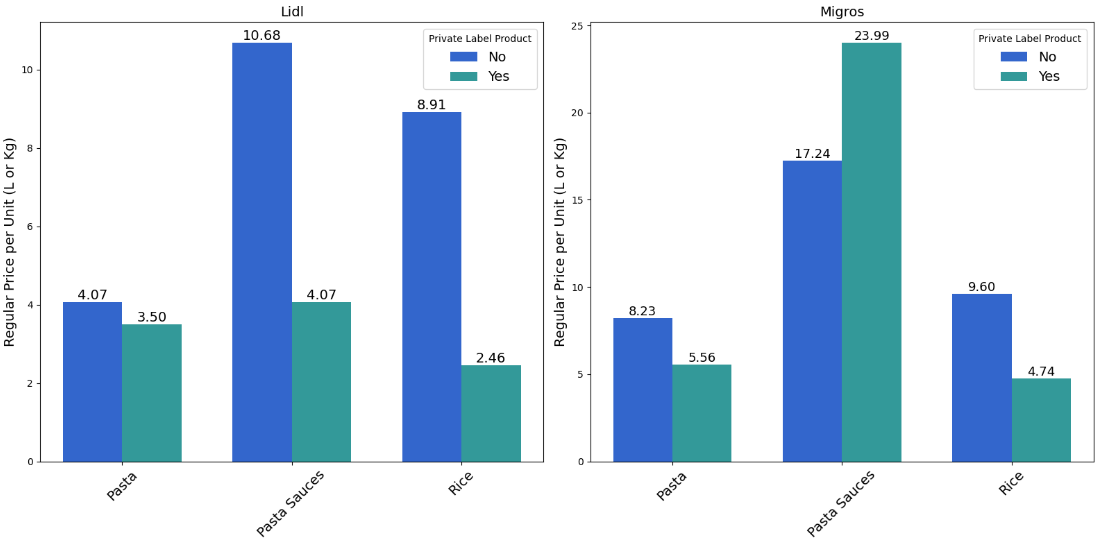
****

Figure 3: Comparison Average Price Between Private Labels and Third Brands

## **4.2 Summary of key Data**

* Summary of the main key insights of the plots above (Fatima, Martina, Cata )

## **4.3 Outcomes**

### **4.3.1 Competitive Advantage**

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

*Which supermarket offers more brands and how might this attract certain costumers who value more brand options?* (Martina)

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

* for “Pasta” categrory, Migros offers approximately 283% of the number of unique brands offered in Lidl,
* for “Rice” categrory, Migros offers approximately 1100% of the number of unique brands offered in Lidl,
* for “Pasta Sauces” categrory, Migros offers approximately 237.5% of the number of unique brands offered in Lidl.

A graph with blue and pink bars

Description automatically generated

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

### **4.3.2 Consumer Benefits**



Table 2: Consumer Benefits

In the plot above, 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.

## **4.4 Observations**

* Comment Unexpected Results or notable differences in pricing and variety of brands. (Cata)

# 5. Conclusion

## **5.1 Summary About Learning**

* Skills gained for web scraping and market strategies (Fatima)

## **5.2 Challenges** (Martina)

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.

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** (Cata)

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.