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)**

* Description of the web scraping setup and tools (Use of Selenium and Beautiful Soup and Pandas)
* Step for accessing and extracting the Data

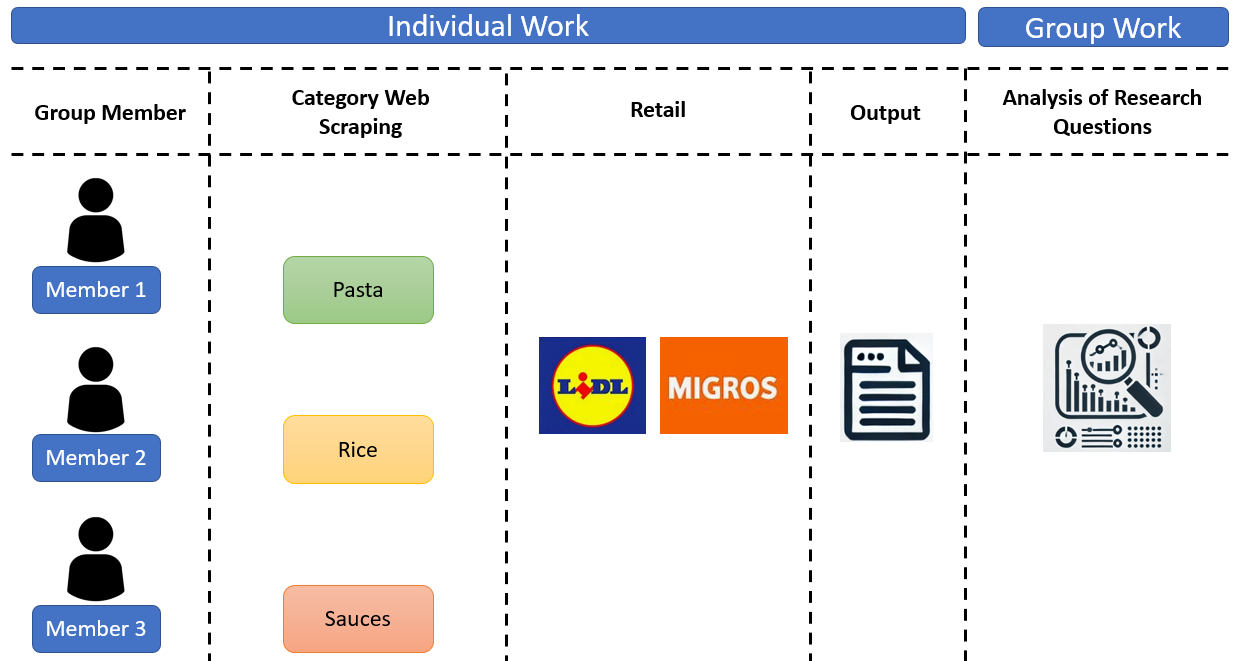


Figure 1: Work Process

* How is the output looks like?
* Structure of the DataFrame (What we added, calculated, etc)

## **3.2 Transformation Techniques**

* Cleaning, formatting and structure of the scraped data for the analysis (Martina – Cata - Fatima)
* Missing Values (strange things) (Fatima)

## **3.3 Analysis**

This section explains how the group will address the research questions, and an exploratory data analysis is conducted to gain a more comprehensive understanding of the dataset.

**3.3.1 Address of Research Question**

1. *Average Prices Comparison (Fatima)*

* How we compared the prices (Average Regular prices – Actual prices)

1. *Number of brands offered by category and competitor (Martina)*
2. *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)

* 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

* Boxplots for Prices per Category

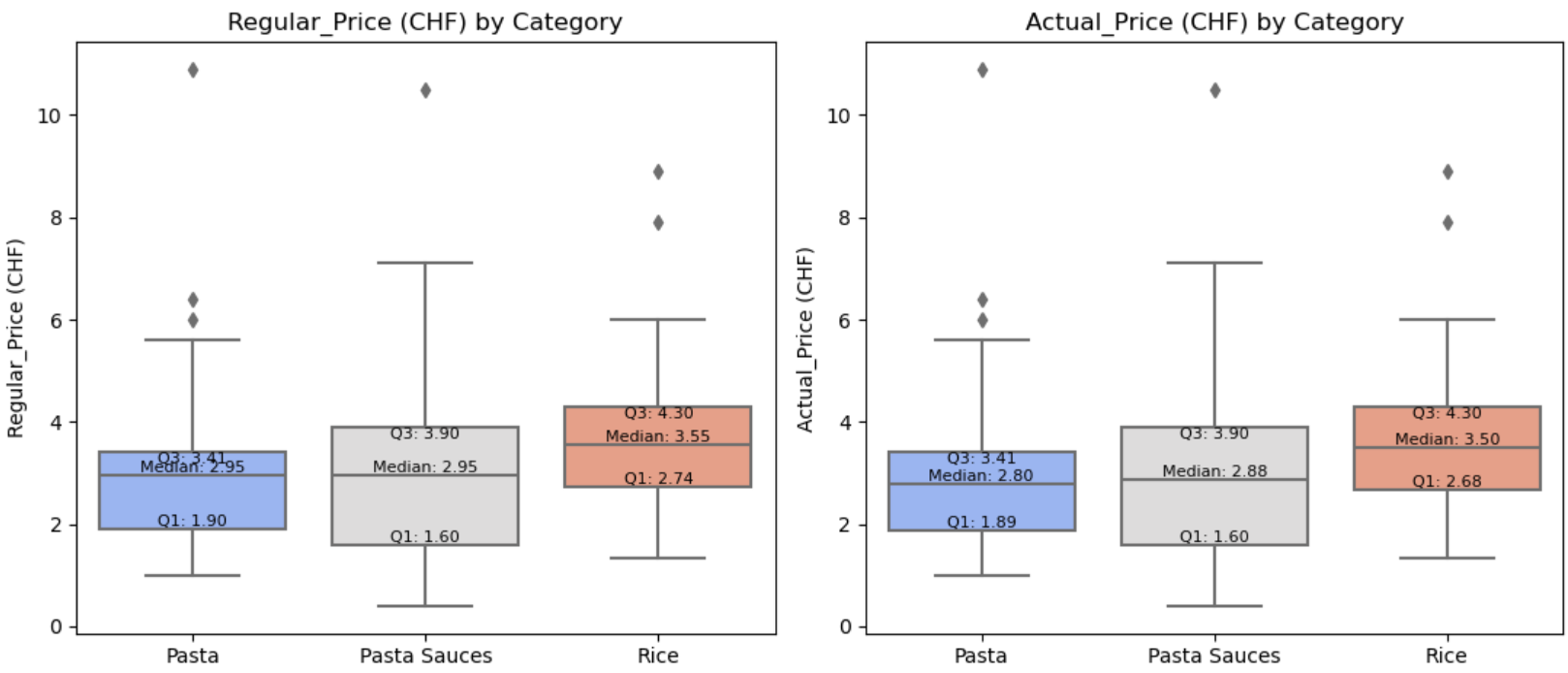


Figure 2: Boxplot for prices

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 will be 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 -🡪 Plot (Fatima)**

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

**4.1.3 Gap Price between Private Label and Third Brands 🡪 Plot (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.

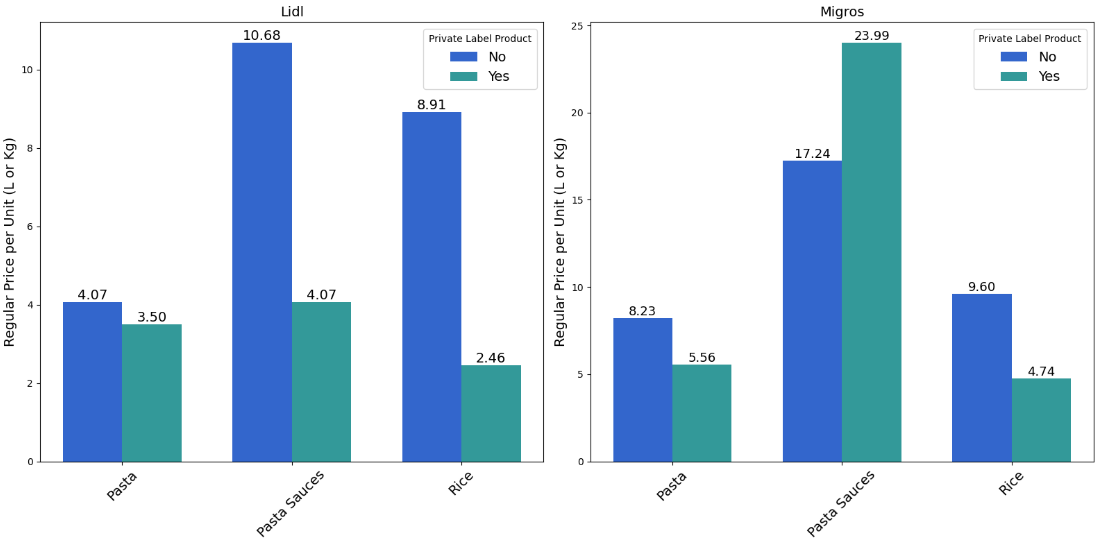
****

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)

### **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 Limitations**

* Challenges Faced (Martina)
* Example: We were blocked, finding buttons, cookies messages, etc

## **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.