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

Feasibility Study

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: 11 October 2024
Final Project hand in Date: 15 November 2024

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.
Its direct competitor, *Coop,* is the second largest food retailer in Switzerland, after Migros. Finally, *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: r*ice, pasta, and sauces (tomato and pesto)* of the chosen supermarkets.

2. Research Questions

The research question that will be addressed through web scraping are:

1. *Which supermarket has the most competitive prices?*
   It is necessary to extract the regular prices from each competitor and calculate the average price for every category (rice, pasta, and sauce).

2. *Which competitor offers more brands across distinct categories?*
   The brands of the different products will be extracted for each competitor.

3. *How much more expensive are own brands compared to traditional brands for each competitor?*
   A comparison will be made between traditional brands (e.g., Barilla, Alnatura) and own brands (e.g., Migros Bio, Migros Budget, Coop) to analyze the pricing convenience.

3. Sources

The project will focus on the three above-mentioned supermarkets that offer online shopping platforms (dynamic web pages): **migros.ch**, **coop.ch** and **lidl.ch.**
The scraping process will focus on three product categories: **pasta**, **rice,** and **sauces**.
For each category, a considerable number of products is available in expanding lists.

Migros
https://www.migros.ch/en/category/pasta-condiments-canned-food/pasta-rice-semolina-grain/pasta
https://www.migros.ch/en/category/pasta-condiments-canned-food/pasta-rice-semolina-grain/rice
https://www.migros.ch/en/category/pasta-condiments-canned-food/spices-sauces/pasta-sauces-pesto

Coop
https://www.coop.ch/de/search/?text=pasta
https://www.coop.ch/de/lebensmittel/vorraete/grundnahrungsmittel/reis/c/m_0143
https://www.coop.ch/de/lebensmittel/vorraete/pastasaucen-warme-saucen/pesto-pastasaucen/c/m_0165

Lidl
https://sortiment.lidl.ch/de/catalogsearch/result/?q=pasta
https://sortiment.lidl.ch/de/catalogsearch/result/?q=reis
https://sortiment.lidl.ch/de/catalogsearch/result/?q=konserven

4. Methodology

4.1 Project Methodology

Fig.1 shows the sequential stages of our project.

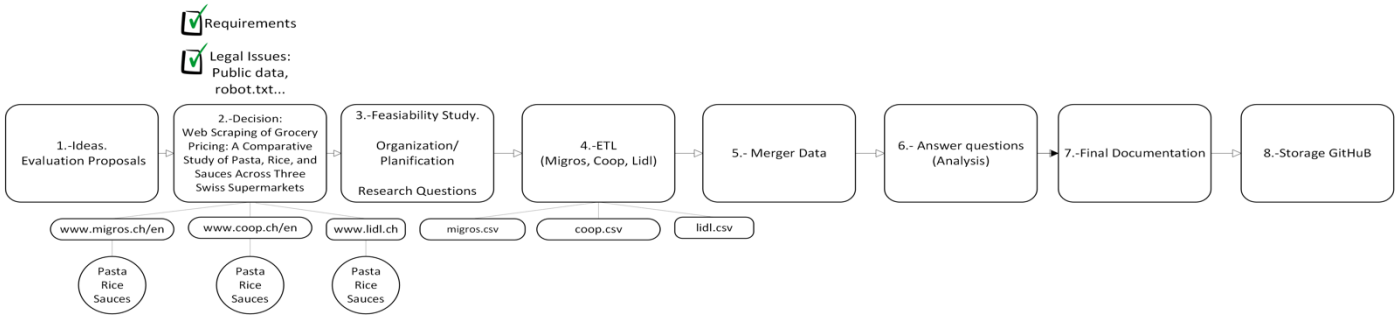

Fig 1. Project Stages Flowchart

4.2 ETL Methodology

Fig.2 below explains in detail the ETL process and the steps that follow.

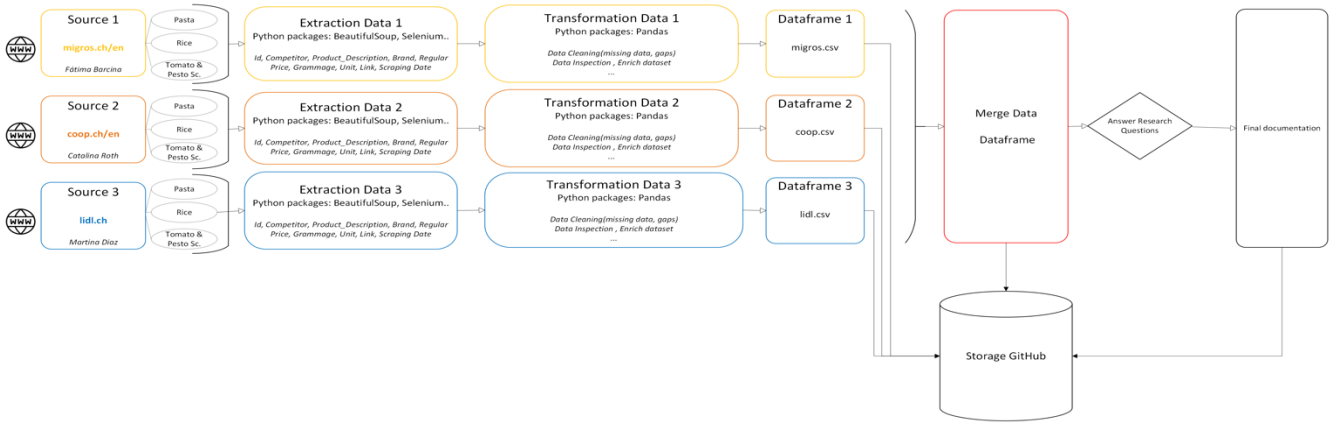

Fig 2. ETL Flowchart

5. Expected Output

A CSV file from each group member is expected. The idea is to merge the three outputs into one file to create a final data frame for data manipulation and analysis. The description of each column is explained in the table below:

Column	Description	Data Type
"Id"	Unique identifier for each row	String
"Competitor"	Name of the competitor	String
"Category"	Category of the product (rice, pasta, sauce)	String
"Description"	Description of the product	String
"Brand"	Brand of the product	String
"Grammage"	Weight of the product	String
"Unit"	Unit of measurement of the grammage	Float
"Price"	Price of the product	Float
"Discount"	Price of discount (if the product has)	Float
"Link"	URL of the source	String
"Date"	Date of the web scraping	Datetime

Table 1: Columns description of the final CSV

6. Potential Risks and challenges

Potential risks might concern the website permissions, the data quality, and a weak answer to some of the research questions. Regarding permissions, checking the robots.txt is the first necessary strategy to confirm the exploit of the selected project's sources. The most relevant issues related to the data quality might be missing values or incoherent product descriptions between the websites. In the worst scenario, this might lead to replacing the original source; otherwise, it can be overcome through the data transformation phase. Eventually, if the results will provide dubious answers, we will provide suggestions to improve the workflow to address more solid proofs.