Participants:

1. Michał Niegierewicz - 397243
2. Piotr Nowicki – 397253

Description of the website

The chosen website is https://www.euro.com.pl/. It is a store of one of the largest Polish retail chains that sells electronics, computers, phones, laptops, TV and the like. A user can navigate to different categories and look through products and their prices, technical details, reviews. It is also possible to make purchases after logging in.

The scraped web page was https://www.euro.com.pl/laptopy-i-netbooki.bhtml which contains laptops and netbooks. On a single page, limited number of products is displayed. By navigating to the next pages, a user can look over all available laptops. Clicking on one of them carries the user to a page with more details about laptops.

Description of the topic

The main goal is to scrape features of the laptops. That is: price, brand, battery, RAM, processor, graphic card, laptop type, SSD memory, screen, operation system. Such data can be further used to make comparisons between prices of different laptops with similar features, differences between brands, or making regression analysis. The task is microeconomic in nature – to look at the market of laptops.

Scrapers mechanics

1. Beautiful Soup

The BeautifulSoup scraper consists of three main parts. In the first of them, links to laptops from the first page and the number of the last page are scraped (because the first page contains information only about 5 other pages). Then, the last page number is used to construct a list of links with the pages needed in the second part. In that part, all links with laptops are gathered using a loop. Finally, the collected links are used to collect the data.

2. Scrapy

There are 3 spiders. First is called “pages” - it obtains links to all pages with laptops. The second spider “laptops_links” scrapes links to every laptop on the pages. The third one - “laptops” - scrapes information of the laptops and put them together into one file.

3. Selenium

The idea for the Selenium scraper is very similar to the Beautiful Soup scraper. First, after accepting cookies, links to laptops and the number of the last page are collected. These results are used in scraping the remaining pages. Finally, in a loop relevant data are collected for each of the laptops.

Output

The final result contains following information: price, brand, battery capacity (Watt-hours), RAM memory (GB), processor name, graphic card, laptop type (e.g. gaming, office), SSD memory (GB), screen (which is split into size, resolution), operating system (Windows 10, Windows 11, Mac OS, no system).

Performance between the scrapers

Time elapsed (in seconds; approximations) for Beautiful Soup: 1105; Scrapy: 28 ; Selenium: 1210. The time.sleep() function was taken into account in calculations (i.e. it was subtracted from the total time). The Scrapy scraper is overwhelmingly faster than the others. Beautiful Soup and Selenium are comparable.

Data analysis

The data analysis was performed after excluding factors with less than 10 observations (e.g. if there is not at least 10 laptops made by brand X, the brand is excluded). Let's start with an average price by brand. Table 1 contains the results. It can be seen that, on average, Apple laptops are more expensive than the others. MSI laptops also come with higher prices. The cheapest laptops are made by Toshiba and Huawei.

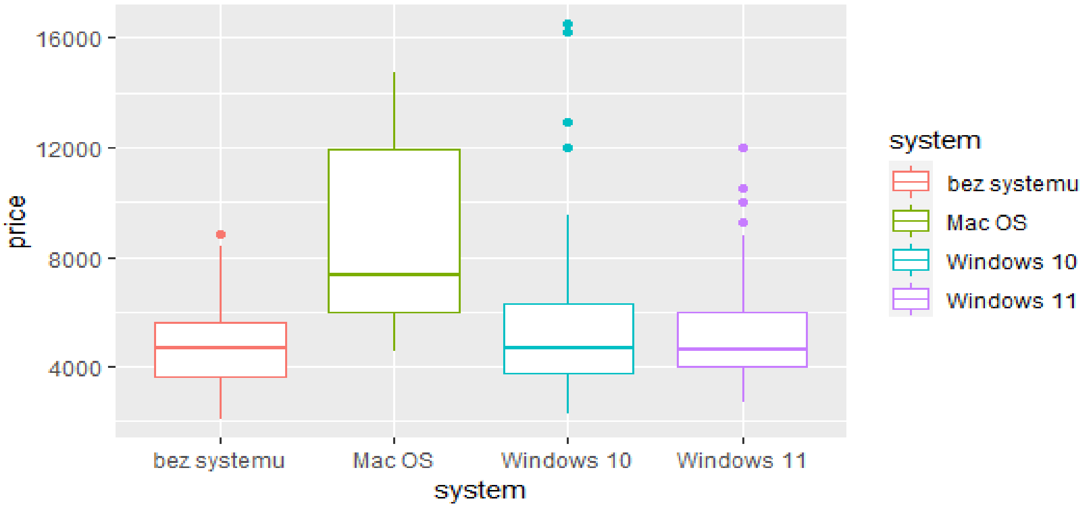
Table 1. Average price by brand.

	Acer	Apple	ASUS	Dell	Gigabyte	HP	Huawei	Lenovo	LG	MSI	Toshiba
Average price	5425	8588	4585	5167	5742	5551	3843	4680	5894	6830	3854

To corroborate that the average price is different between brands, we can perform ANOVA test or Kruskal-Wallis test. Both of them return p-values close to 0, hence the average prices differ at least between two brands.

Figure 1 shows boxplots of price conditional on an operating system. There are little differences in distribution of prices between 'no system' (in Polish: bez systemu), Windows 10 and Windows 11. However, Mac OS laptops are usually more expensive than the others.

Figure 1. Boxplots of prices with regard to operating system.



One can also run a regression analysis. Regressing ‘price’ on all available features, returns a model with adjusted R^2 equal to 0.9241. The resolution and size of a screen seems to not have a significant effect. Preliminary results suggest that graphic cards, RAM and processors have the largest effect on prices of the laptops. It also seems that some features can be highly associated with each other as well as some instances within a feature. It might be possible to reduce the number of variables by merging them with each other.

Participants’ contribution to work

Michał Niegierewicz – writing Beautiful Soup and Selenium scrapers, writing instructions for Beautiful Soup and Selenium in README.md, writing technical description of the Beautiful Soup and Selenium scrapers in description.pdf.

Piotr Nowicki – creating github repository, writing the Scrapy scraper, performing data analysis, writing description.pdf file, writing an instruction for Scrapy in README.md, adding minor changes to the Selenium scraper (chunk of code that expands technical details on a web page; 2 corrections to XPATHs), calculating performance of the scrapers.