WEB SCRAPING PROJECT

OLX SITE: PROPERTY OFFERS IN WARSAW

LINK TO PROJECT REPOSITORY: https://github.com/wmotkowska-stud-412081/ws_project_2022

TEAM MEMBERS:
Weronika Motkowska (412081)
Karolina Kowalska (412009)
Magdalena Pruszyńska (443737)

DESCRIPTION OF THE TOPIC AND THE WEB PAGE

The main goal of the project was to scrap and save information about properties (such as name or price).
Those are valuable information that can be used by business analyst to evaluate the current market or prepare an econometric model, compare it or to speed up the process of finding a property to buy.
The used page is OLX, which is a page that accumulate many different offers posted by people. For the purpose of this project, the focus was only on the properties for sale and only in Warsaw. Each page contains photos, location, important characteristics and long description. However, there were also some limitations on what could be scrapped. For example, contact details were disallowed in robots.txt page. Detailed description on what was scrapped is in the next part of this file.
The procedure was conducted in three different ways: using Beautiful Soap, Scrapy and Selenium. In the end, a data analysis image was prepared to show, how scrapped data can be used.

DESCRIPTION OF THE SCRAPER MECHANICS AND THE OUTPUT

Our program scraps links to property offers listed on https://www.olx.pl/nieruchomosci/mieszkania/sprzedaz/warszawa/. If the boolean parameter is set as „True”, then only 100 links are considered. It was made sure that all links used were from "olx" domain. The links which led to "otodom" site were deleted, because a different group chose this page.

For each link with offer the program scraps:
- name,
- price,
- price for m^2 ('price_m2'),
- m^2 ('m2'),

- how many rooms ('rooms'),
- which floor the apartment is on ('floor'),
- where is the property located ('map').

Name and price are scraped from the header. Other properties are scraped from detail table. We were only able to obtain map variable in selenium and this property was vital for data analysis.
All three methods are similar and provide similar outputs. The main difference is the time of the processing. The fastest is Scrapy, it takes only a couple of seconds to process. Beautiful soup is slower: about 2-3 minutes. Selenium scraper takes the longest time because of time limitations, which are crucial for page to get its full html.

The output is in a form of a dataset (csv format). The map variable is a string with 'Warszawa' being the first word and district following it. Price and price m2 are in PLN. URL is the link to the offer. 'Parter' in floor means 0.

DATA ANALYSIS

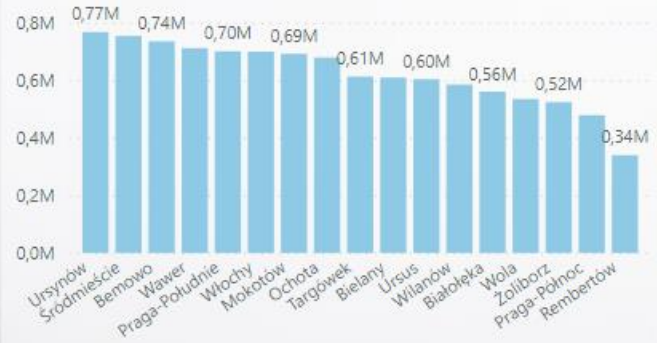
A short visualisation of the characteristics of property offers can be seen on data analysis image on the next page. The bigger the circle on the map, the higher the average price in the district. The brighter the colour of the circle, the higher is the price for m^2.

DIVISION OF WORK

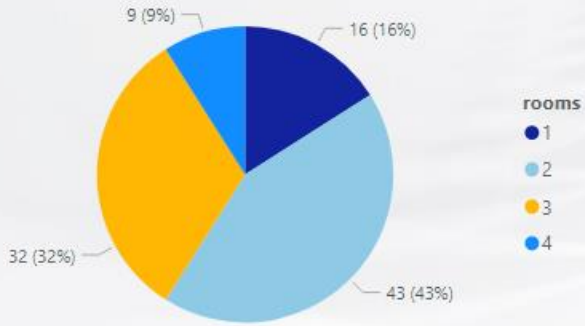
Karolina Kowalska was working on Beautiful Soup and collaborated on description file.
Magdalena Pruszyńska was working on Scrapy and prepared data analysis image.
Weronika Motkowska was working on Selenium and collaborated on description file.

648,22K	30K	1958K	51,38	2	Warszawa, Ursynów
Average of price	Min of price	Max of price	Average of m2	Typical number of rooms	Most expensive location

Average of price by Location



name	price	m2	floor	rooms
Mieszkanie 3/4 pokoje 78 m2 Warszawa Bemowo Wola metro komórka parking	1958000	155,70	4	6
Mieszkanie 172.2 m2 Warszawa Ursynów	1670000	172,20	2	4
Mieszkanie 64m2. Warszawa Wola	1500000	64,00	3	3
Rodzinny Apartament, 4 pokoje, 2 x Garaż, 4 parki	1480000	108,00	1	4
Sprzedam Mieszkanie 76m + 2 miejsca, bezpośrednio, ul. Magazynowa	1350000	76,00	7	3
Mieszkanie 114 m2 z	1260000	114,62	Parter	4



price by m2