OLX.PL SCRAPER

Description of the topic and the web page

Olx.pl is a Polish website with various advertisements. One of the types of advertisements posted on this website are flat rental advertisements. The project consists in making scrapers that will download data on advertisements for renting apartments in Warsaw. The data collected are: the rental price of the flat, the floor on which the flat is located, whether the flat is furnished, the type of building in which the flat is located, the area of the flat, the number of rooms in the flat, and additional rent. The obtained data may allow for an assessment of the flat rental market in Warsaw

Short description of scraper mechanics

All three scrapers have very similar mechanics. Initially, they download links to individual ads based on links to ad websites. Then, using the downloaded links to advertisements, they download the following variables for each advertisement: flat rental price, floor on which the flat is located, whether the flat is furnished, type of building in which it is located. , apartment area, number of rooms in the apartment and additional rent. Comparing the three scrapers: soup and selenium take about 80 lines of code, and the scrapy scraper about 70 (no comments). Regarding the execution time - soup scraper needs about 1 minute and 45 seconds, scrapy scraper only needs about 15 seconds, while selenium scraper needs about 3 minutes (for taking 100 observations, measured with a stopwatch). The scrapy scraper seems to be the fastest and requires the least lines of code. However, in my opinion it is the most difficult to understand. On the other hand, selenium scraper, despite the fact that it is the slowest, personally gave me the most fun when I was doing it. Its interesting advantage is that you can observe its operation directly in the browser window. However, its disadvantage, apart from speed, is that I had to give a 3 second pause in the code because the page was loading too slowly in relation to the code execution and thus the scraper was downloading data from previous ads for some links. This created duplicates and reduced the overall number of sightings as some of the ads were not downloaded. However, adding a 3 second pause fixed the problem.

Short technical description of the output

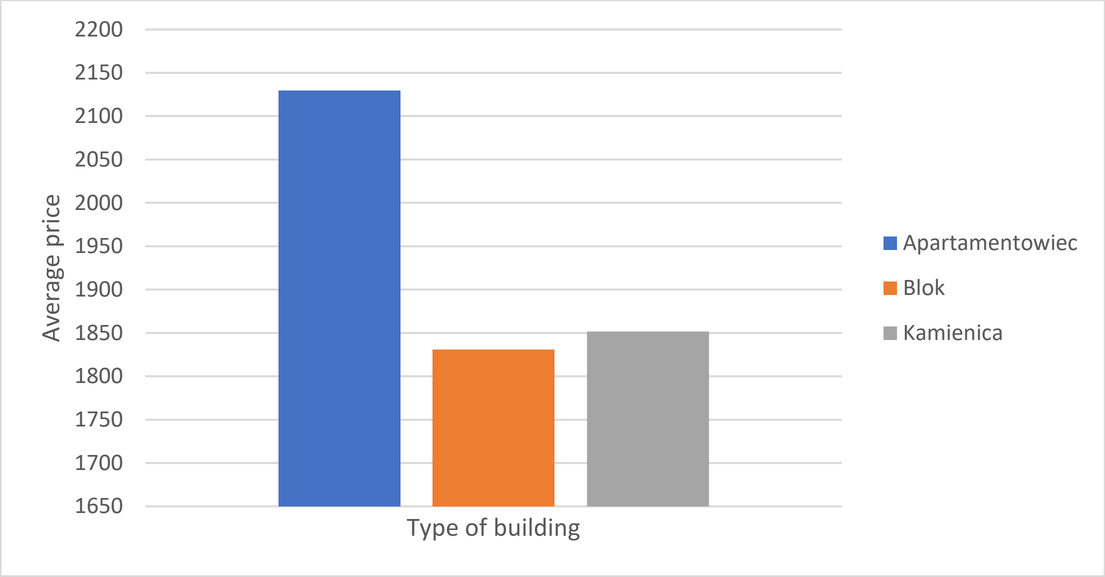
As an output I got eight variables - one binary, three continuous, two discrete and one qualitative. Unfortunately, the data received is a little dirty and needs some cleaning up. In some cases, you need to remove units to keep the numeric type of the variable, etc.

Division of work on the project

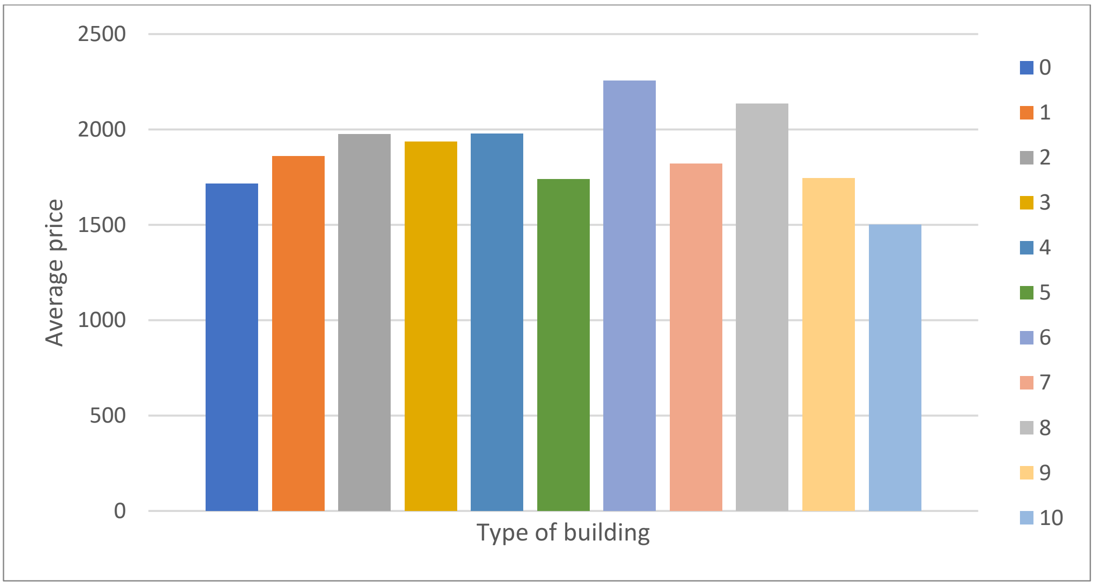
I made the whole project myself.

Data Analysis

The average price of renting a flat amounted to 1,893.28 PLN. The average flat was located on the 3rd floor with an area of approximately 39 sq m. Each flat was furnished.



As expected, the most expensive apartments were in an apartment building. Interestingly, the average price of renting an apartment in a block of flats was slightly lower than the price of renting an apartment in a tenement house (a greater difference was expected).



On average, the most expensive apartments were located on the 6th and 8th floors. Surprisingly, the cheapest were the apartments on the 10th floor. The second cheapest level was the ground floor. There is no significant difference in the price of apartments on floors 1-5.

In addition, the obtained data can be used for more advanced analysis and for the construction of several interesting models regarding the apartment rental market in Warsaw.