Web scraping report

Homework for Data Analysis 1 || Origins of Data – Data Exercise 4)

Exercise

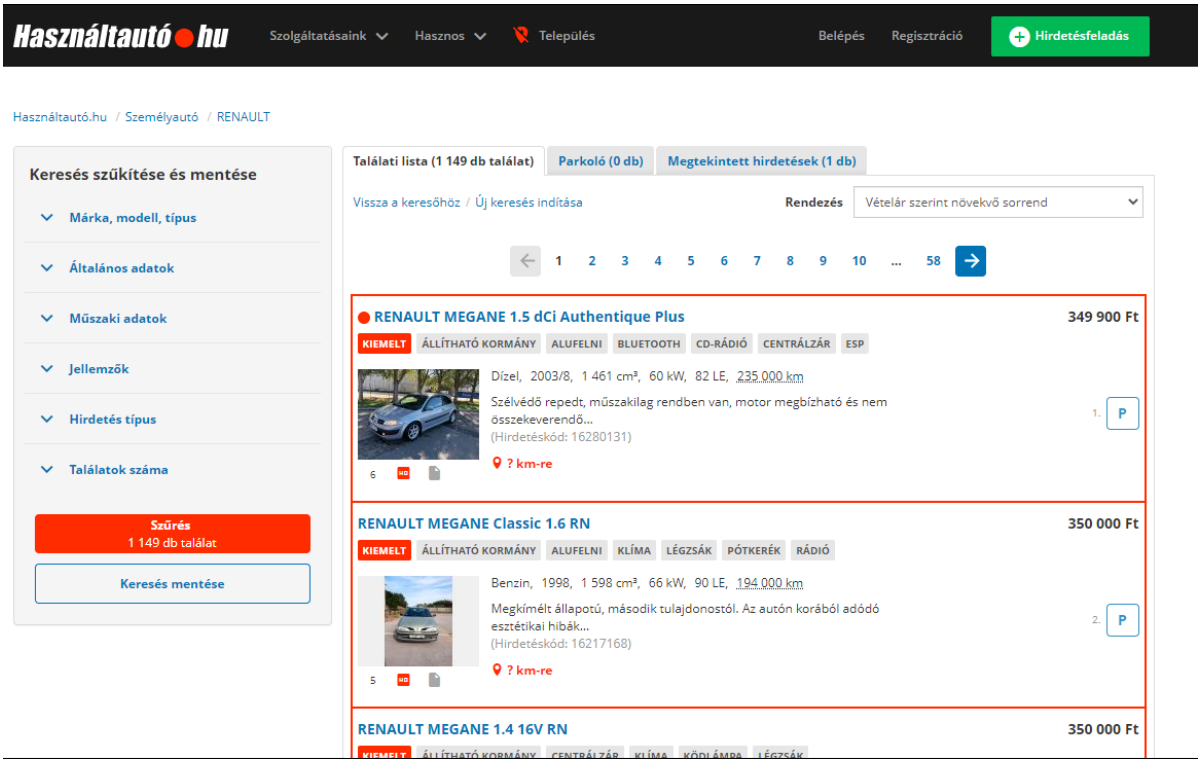
The task at hand was to collect data on used cars of a specific model form the Web using web scraping and then describe:

1) The process of web scraping
2) How many observations I collected
3) Encountered difficulties

Section 1 – the web scraping process

The website of my choice was hasznaltauto.hu, as this is one of the biggest Hungarian sites advertising used cars. On the opening page of the site, there is a search prompt that asks you what type of car you are looking for. There is a lot of options to choose from in the dropdown where a small number in brackets indicates how many of such vehicles are advertised currently. I briefly looked through all brands and decided to go with a car that has a relatively high number of adverts. This is the Renault Megane.

After choosing the car, I was directed to a site with 20 adverts/page (a snapshot of this is available in Figure 1) – I will call this a subpage. My first task for the scraping was figuring out a way to scrape every subpage that is related to the Megane advertisements. In Figure 1 you can see that in this task I needed to go through 58 of such webpages. The way how these distinct pages are represented in URL is by a string in the end with '/pageX' where X denotes the order of a given subpage. To go through each one of them, I generated all 58 URLs with their respective '/page' appendage and put them in a list.



1. Figure: Overview of hasznaltauto.hu

After the list of URL links was created, I took one of the subpages and wrote a scraper in Python. The html code behind these subpages was relatively clear and simple and there were a lot of features that I could collect. The Python package, called BeautifulSoup was used for the parsing, and the most common functions that I took were the *.find()* and *.findAll()* functions. The result of the scraping was a Pandas dataframe that I downloaded as a csv file and uploaded in github.

2) Describing the data

Figure 1 is a good representation to show the amount of information available for the user on each subpage. Table 1 is the summary of Variables collected from such sites with web scraping:

Variable	Data type	Description
ID	Primary key, Integer	Unique identifier for each advert generated by the site
ShortDesc	String	A short description for given advertismenet
Prices	Integer	The price of the car expressed in HUF
Engine	Categorical	The type of the engine either Diesel, Gasoline or in rare cases, LPG
Year	Date (YYYY/MM)	The year this model was introduced
Displacement	Integer	The measure of the cylinder volume in m^3
Power_kW	Integer	Power expressed in Kilowatts
Horsepower	Integer	Power expressed in Horsepower
Km	Integer	How many kilometers the car travelled so far
PictureCount	Integer	How many pictures were uploaded in the ad

1. Table: Variables scraped from hasznaltauto.hu

Overall, I collected 1,150 observations.

3) Challenges faced

A few challenges that I faced with this exercise.

- **Dispersed nature of data** - only 20 observations were available for each subpage.
- **Hidden layers** - price appeared twice due to hidden layers in the html code which I realized at a later stage only making my price list twice as big as the other lists
- **Unnecessary characters** – some characters like spaces, or units of measurement were scraped along with the necessary information
- **Missing data** – Some ads had missing data that I needed to replace in my dataframe with NaNs