Chapter 1 - Data Exercise 4

Attila Serfozo

2020.10.15

Purpose of the data collection

The purpose of my exercise is to collect data on a selected car type from the largest Hungarian used car website (https://www.hasznaltauto.hu/) with web scraping in order to analyze how the price varies through the different variables. To be able to conduct the price comparison, I selected Ford Focus as a car type and the key Characteristics of the used cars I want to include. The description of these variables can be found in the Variables.xlsx file.

For the task I used a publicly available point and click interface web scraper which can be added as an extension to Google Chrome at the following website (https://webscraper.io/).

Selection Criteria

I wanted to collect data only on used cars, therefore advertised new cars or demo cars on the website are excluded. I also wanted to collect data specifically on Ford Focus Hatchback to be able to provide a better price comparison in the next task where I will do a price comparison. What this means is that I excluded:

- the Ford Focus Sedan vehicles as sedans tend to follow different pricing.
- the Ford Focus Estate cars, mainly because their pricing is different as most of these cars are used taxis, police cars or company cars, which have larger amortization, but also estate cars are another market segment.
- the Ford Focus C-Max, which belongs to another market segment as it is a minivan.

Also I wanted to examine the price of the cars registered from 2010, so maximum 10 years old. My motivation behind the age selection is that I also have an approximately 8 years old Ford Focus and I was interested in the price of these cars less than 10 years. Based on these criteria I ended up with 244 Ford Focus cars.

Difficulties during the task

One of the barriers was that the 244 cars were spread on multiple webpages and it was hard to set up the scraper, because I wanted it to be able to move between pages and collect the data in one exercise. By understanding the working method of the scraper and with some practice I could cross this obstacle and collect the data with one scraping.

It was also a difficulty that many times when I wanted to rerun the scraping after modifying something, the number of cars available on the website were varying during the time, so it sometimes made cross-checking hard. (Note: I ran the web scraping algorithm at 2020.10.13 15:32.)

Also another difficulty is regarding the data quality. In the beginning I decided I want to collect data on hatchbacks, however many owners clicked hatchback as the category of their estate or sedan cars, so there are some observations which are not part of the intended content. Also there are many typos in the different fields which should be carefully observed how to handle.