Analysis on used cars

Zsombor Hegedus

2020-10-25

Introduction

The purpose of this document is to summarise the work done to complete two homeworks. * The first one is Data exercise 2 in Chapter 1 (p28) * The second one is Data exercise 2 in Chapter 2 (p55)

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

- The process of web scraping
- How many observations I collected
- Encountered difficulties
- Cleaning the data
- Describing the cleaned dataset

All codes and files used for this analysis are avaialbe in this github repo.

Web-scraping

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 - 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. 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 *page* 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.

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 to github.

Challenges faced and data cleaning

The challenges that I needed to face can be grouped into the following categories:

- 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 NAs.

Variables and README file is available in the github repo.

When I started data analysis, there were a few issues that I needed to address and clean my data before I could proceed. First of all Prices variable contained characters such as 'Ft' which stands for Forint. This could be solved easily by replacing the non-number characters using the *gsub* function. However some observations had taxation related information that I needed to filter out with more sophisticated means. At the end of such observations, there was always a string followed by an integer, which marked the final price that someone has to pay for the car, taxes included. I used the separate function and used that specific string **fizetendo** as my separator.

I did further smaller cleaning steps as well namely:

- Dropping observations where price were not available
- Changing numeric variables to be recognised as numeric with as.numeric function
- I converted Engine variable to factor variable, so that I can use it for further analysis

Describing the data

Table 1 shows a summary of the kez features of used car prices. We have 1147 observations (we deleted 3 where price was NA). Our median is much lower compared to the mean, and the price distribution has a positive skewness with a thick right tail as you can see from Figure 1. Giving a quick overview on prices, the earlier the car was made, the higher it's price will be.

mean	median	std	iq_range	min	max	skew	numObs
1572108.30	900000.00	1689725.70	1365000.00	35000.00	13899000.00	2.54	1147

Table 1: Summary statistics used Renault Megane cars

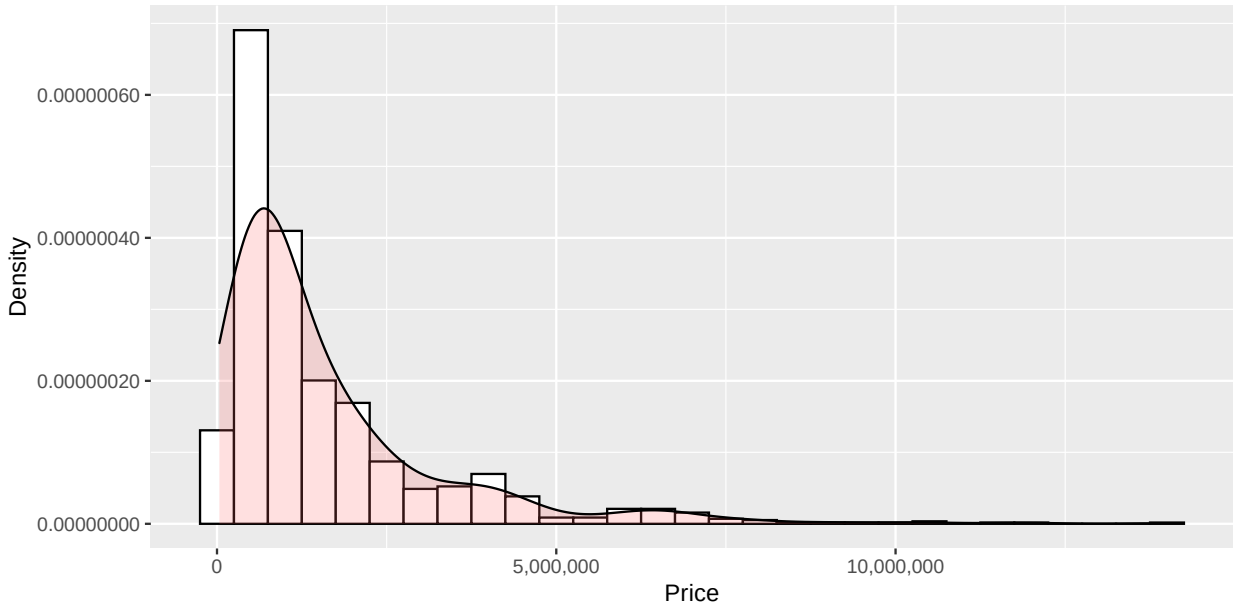

Figure 1: Price distribution uf used cars

It is worth taking a look at Table 2 and Figure 2 as well. We can see that most of our cars are either run on Patrol or Gasoline, and looks like Petrol run cars on average cost more.

Engine	mean	median	std	iq_range	min	max	skew	numObs
Benzin	1372072.68	790000.00	1682747.12	1250000.00	35000.00	13899000.00	3.33	560
Benzin/Gáz	1047222.14	920000.00	743842.60	621222.50	198000.00	2499000.00	1.01	7
Dízel	1631209.70	1097000.00	1449953.52	1458890.00	80000.00	7999000.00	1.90	553
LPG	597500.00	690000.00	239078.09	217500.00	250000.00	760000.00	-0.96	4
	5350782.61	6499000.00	2752481.59	2885000.00	500000.00	8990000.00	-0.83	23

Table 2: Summary statistics used Renault Megane cars

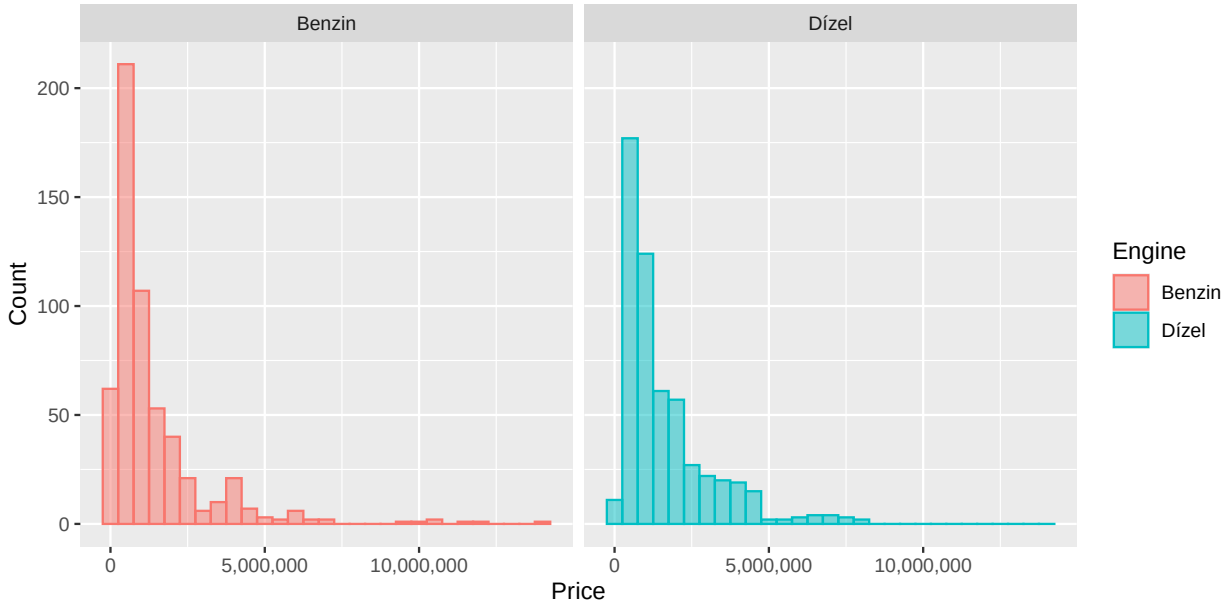

Figure 2: Price distribution of used cars with Gasoline or Petrol engine

Let's uncover further interesting factors - the correlation matrix based on our numeric variables. Since correlation was 1 for HorsePower and Power_Km (they are basically the same, only they are expressed in a different unit of measurement), I dropped the Power_Km variable.

There are three points we can make by looking at Figure 3, all of which are quite intuitive:

- There is stronger positive correlation between price vs horsepower - people like to speed up fast
- There is no correlation whatsoever between displacement vs prices - looks like people don't price in the cylinder volume
- Strog negative correlation between km vs prices - People percieve cars that already run a lot to be less reliable perhaps

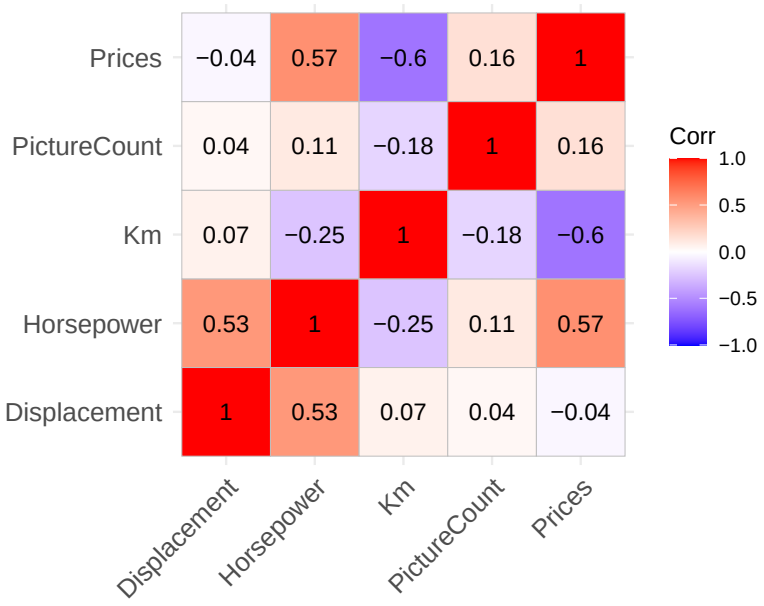

Figure 3: Correlation matrix of scraped numeric variables