Scraping Huizenzoeker.nl to Analyse the Dutch Housing Market

Contents

Which places in the Netherlands are hit hardest by the Dutch Housing crisis, and which the least? Currently, the housing crisis is one of the most prominent societal challenges in the Netherlands. The Dutch housing market is both very competitive as well as inaccessible as it must deal with a supply shortage, which in turn leads to long waiting lists for social housing. The transaction prices of houses are going through the roof, as CBS (Centraal Planbureau voor Statistiek, 2021) revealed that prices in the first quarter of 2021 were 11.3% higher compared to a year before, which is way above the average increase in Europe. Many prospective buyers have to overbid on the listings in order to ensure a place to live. A huge problem is that the amount of mortage is determined by the appraisal value of the house, which causes many to put their own capital into the purchase. This makes it very challenging for new entrants on the market (think of first-time buyers, young professionals) to succeed in renting or buying a house or appartment, especially as many are still paying off large amounts of student debt. It appears that only 3% of first-time buyers are financially able to buy a home without getting themselves into serious financial trouble.

We have considered a few multiple housing sites to incorporate into our project, where Huizenzoeker.nl appeared to be the most suitable option. This website offers a clear view of the Dutch housing market with a wide range of listings, displaying an extensive amount of information (per listing and neighbourhood). Funda.nl currently is the largest housing provider in the Netherlands, however, the site is not useful for this project. Funda installed secure protection for its data to brace for competitor sites. Similarly, Zoekallehuizen.nl offers a large range of listings too, but could not provide us with important information needed to research the housing crisis, e.g. overbidding percentages. Similarly, Remax.nl, is a large housing website, yet, mainly focusing on houses in other countries, like Spain and Belgium. As we are determined to analyse the Dutch housing market by cause of the severe current crisis, Remax.nl has not sufficed to our needs.

Motivation

1.1 For what purpose was the dataset created? As the seriousness of the housing crisis and the shortage of listings differs across the country, we aim to create a dataset which represents the current housing market for each municipality in every province of the Netherlands. This dataset clarifies which places are hit hardest by the crisis and which the least. We facilitate first-time buyers and young professionals with fundamental information in their search to buy or a rent a house. Furthermore, the dataset provides insights into price developments of listings in certain areas, helping consumers gain relevant information for purchase price-negotiations. This dataset provides consumers with additional data, compared to the information that is offered by their broker (*e.g.* direct information from the Kadaster). There are datasets about the Dutch housing market available already. However, these do not specifically focus on the overbidding aspect of the current crisis, which forms an essential part of our research. Besides that, in addition of narrowly focusing on certain areas of the Netherlands, we preferred to focus on all municipalities in the Netherlands to get a more complete picture of the current state of the housing crisis. By focusing on municipalities, the units we are analyzing are small enough to deeply dive into the housing market of the Netherlands locally (as opposed to only focusing on provinces). Yet, the units are large enough to maintain order and control in our dataset (as opposed to focusing on every house that is for sale in the Netherlands).

1.2 Who created this dataset and on behalf of which entity? This dataset is created for the course Online Data Collection and Management at Tilburg University, as part of the Master's program 'Marketing Analytics'. The contributors of the project are:

- Lesley Haerkens, l.w.g.haerkens@tilburguniversity.edu
- Mila Gargiulo, m.r.e.m.gargiulo@tilburguniversity.edu
- Daniëlle van Bruggen, d.m.vanbruggen@tilburguniversity.edu

1.3 Who funded the creation of the dataset? There was no funding granted for creating the dataset.

Composition

2.1 What do instances that comprise the dataset represent? The instances that comprise the dataset represent all **municipalities** of the Netherlands, as being one type of instance. The entities summarize the data of the housing market (for every house) in that municipality. The values related to the municipalities in the Netherlands represent the averages per municipality. The instances are connected to each other by the province that they are in. Therefore, all municipalities belong to a larger type of instance, the provinces. As a bonus to the datasets that were generated for each municipality and province, we created a scraper that navigates from province to municipality, from municipality to residence, and from residence to street names. Since this would generate an enormous amount of URLS, generating this dataset for each street name in the Netherlands would take a lot of time (which is beyond the scope of this project). Therefore, to give an impression of how data on street-level looks like, we decided to focus on the province 'Noord-Brabant', the municipality 'Tilburg' and the residence 'Tilburg'. From the residence 'Tilburg' we extract all the streets (such as Warandelaan, the street name of Tilburg University!).

2.2 How many instances are there in total (of each type, if appropriate?) If every municipality is seen as an instance, we would say there are 352 municipalities in total, which are spread over the 12 provinces of the Netherlands.

- Groningen = 10 municipalities
- Friesland = 18 municipalities
- Drenthe = 12 municipalities
- Overijssel = 25 municipalities

- Flevoland = 6 municipalities
- Gelderland = 51 municipalities
- Noord-Holland = 47 municipalities
- Zuid-Holland = 52 municipalities
- Utrecht = 26 municipalities
- Limburg = 31 municipalities
- Noord-Brabant = 61 municipalities
- Zeeland = 13 municipalities
- Total = 352 municipalities.

2.3 Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? All municipalities in the Netherlands are included in our dataset (the full population). However, since the data is grouped and displayed as an average for each municipality, our dataset could be interpreted as a sample. Moreover, we created a subset of the streets for the residence Tilburg, as an illustration of the possibilities of scraping data both on province level and on street level. Generally speaking, as every listing is located in a municipality, the sample is certainly representative of the larger set.

2.4 What data does each instance consist of? To clarify, each province knows its own page and is at the same time the parent of several municipality pages (as described in question 2.2). However, the structure of this province and municipality page are almost identical. For this question, illustration is based on one of the municipality pages, Tilburg (Noord-Brabant). For this question, several *screenshots* were taken from the Huizenzoeker.nl site.

First, all pages display a map of the Netherlands and their specific location on the map. Next, all contain a link to all the houses that are for sale followed by a link for all houses that are for rent. Next, each page displays 4 ‘trend’ statistics. Each of the 4 numbers contains a related number, reflecting the percentage difference of the statistic compared to the month before. The first trend refers to the average selling price of a house within the municipality/province. The second trend refers to the number of houses sold in the past month. The third trend refelcts the average selling price per squared meter. And the fourth trend indicates what the average outbidding percentage is within the municipality/province in question. These trends will be of high importance during our project.

Additionally, all pages cover histograms that show price and housing supply trends.Next, a section is shown in which several questions are answered in unprocessed text. The first questions cover the exact same as the first 4 trend statistics. However, the last ones cover the population number and population growth/decline compared to the year before. This population-related information, again, will be of high relevance later in our project.

Furthermore, a pie chart showing the average age distribution in the province/municipality is included. Moreover, a statistic on average disposable income is included, which again will be important later in our project. Finally, at the bottom of the page random houses that are for sale/rent are displayed, followed by links that navigate to a ‘child’-page (e.g. from province page to municipality page).

2.5 Is there a label or target associated with each instance? From each province in the Netherlands, we intend to scrape all corresponding municipalities. For the provinces an associated URL is for example 'https://www.huizenzoeker.nl/woningmarkt/noord-brabant/', which changes to 'https://www.huizenzoeker.nl/woningmarkt/noord-brabant/tilburg/' for Tilburg. So, the instances that we want to scrape correspond to their own URLs.

Moreover, within the code we wrote, we extracted the municipality or province name for each of these URLs, by scraping the title and removing the word 'Woningmarkt' from it. Therefore, we changed the official label to an artificial one for clarity purposes, e.g. now the municipality Tilburg can be identified through the label 'Tilburg', instead of its URL. This 'base' url (*i.e.*, 'https://www.huizenzoeker.nl/woningmarkt/noord-brabant/'), extends even further when navigating to residence and finally to street name. The url of 'Warandelaan', *e.g.*, is 'https://www.huizenzoeker.nl/woningmarkt/noord-brabant/tilburg/tilburg/warandelaan'.

2.6 Is any information missing from individual instances? Some data from Huizenzoeker.nl that we intended to scrape is displayed in charts and graphs, and for that reason, we are unable to grasp this data and extend our dataset according to this information.

2.7 Are relationships between individual instances made explicit (e.g., user's movie ratings, social network links)? Yes, the municipalities are related to each other by the province that they are in. When running our code for all municipality URLs in a certain province, all information for every municipality within that province will be the output. The parent URL (province) and child URL (municipality) are **always** connected. This also holds for residences and street names. This is explicitly visible in the URLs. For every municipality, the municipality name in the link follows the province name (see: question 2.5).

2.8 Are there recommended data splits (e.g. training,development/validation, testing)? There are no recommended data splits in our dataset, since our dataset is parsed and collected into one file. We chose to parse and collect the variables in our dataset together in one script to make all the information quickly accessible. We think this is no problem mainly because of the following reason. Huizenzoeker.nl updates its content each month automatically. The structure and the URLs stay exactly the same, yet, the statistical numbers change per month. We designed our code in a way that it captures every number that is present at the moment, regardless of whether we run it, *e.g.*, in September versus October.

2.9 Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g. websites, tweets, other datasets)? If it links to or relies on external resources:

a) are there guarantees that they will exist and remain constant over time; Huizenzoeker.nl states that, for over 10 years, they have made every effort possible to ensure that this website functions properly and is kept permanently accessible for reputational reasons. Huizenzoeker.nl edits the information offered on its site with the greatest possible care and devotes the same care to the composition of the site. However, it legally cannot guarantee the correctness and completeness of the data shown as a result of imperfections that may occur. Moreover, Huizenzoeker is able to adapt the website where and whenever they please. No restrictions hold. This information has been retrieved from the disclaimer section on the Huizenzoeker.nl website.

b) are there official arhival versions of the complete datasets (i.e. including the external resources as they existed at the time the dataset was created). Possibly for own utilization. However, no official archival versions of the complete datasets are available to us as the public of Huizenzoeker.nl. Huizenzoeker.nl displays real-time monthly data, and not so much archival data for the data we scrape to answer our research objective (there is for instance data on the 'prijsontwikkelingen' over the last couple of years, which means data for the previous years must be available too). The data we scrape from the municipality pages is data that is updated every month, so when scraping this page you do not get direct access to the figures or averages for the previous months.

c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? The external resources include JAAP.nl and Huislijn.nl, who in turn extract data from other sites such as Funda.nl. These sites are all available for free, thus, no restrictions are present in the form of licenses and fees for future users. There is a premium (under 'Abonnementen') part of Huizenzoeker.nl, for which you do need to pay to access it. For our project, the premium information was irrelevant.

2.10 Does the dataset contain data that might be considered confidential (e.g. data that is protected by legal priviledge or by doctorpatient confidentiality, data that includes the content of individuals non-public communications)? No, the data is not confidential. Therefore users do not have any rights to remove listings from the Huizenzoeker site. Only if their house is no longer for sale/rent on JAAP.nl, their listing will be removed. However, information on the house itself such as its value, year of construction, property size, will remain available. This information is considered as public.

2.11 Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? No the data can in no way be perceived as offensive, insulting, or threatening.

2.12 Does the dataset relate to people? If not, you may skip the remaining questions in this section. The dataset relates to people in the sense that data is available for variables such as the average disposable income and the age distribution in a certain municipality or residence.

2.13 Does the dataset identify any subpopulations (e.g., by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset. Our dataset does not relate to sub populations, as we scrape houses and not people, so there is no need to identify sub populations.

2.14 Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset? This question is not applicable to our dataset.

2.15 Does the dataset contain data that might be considered sensitive in any way (e.g. data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? This question is not applicable to our dataset.

Collection process

3.1 How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? Via the selenium package we accessed ChromeDriverManager. Using this webdriver we scraped all the possible municipality URLs for the entire Huizenzoeker website.

The next step was to extract all the variables desribed in question 2.8. One code has been created for this step. At the beginning of this code, we made sure all the output got saved into a json file. Next, one big for-loop is created that will loop through all of the munucipality pages, making sure the code gets 5 seconds of sleep. Within the loop we loaded the BeautifulSoup package. We then defined the first variable intended identification purposes: the municipality name. These were all the steps that needed to be completed in order to scrape all the variables we wanted from the municipality pages. These variables have been created using multiple 'if' - 'else' statements, tailoring each variable to its corresponding html output that can be accessed when inspecting the municipality webpage. Furthermore, irrelevant characters/words have been dropped to enhance clarity. At the bottom of the code all the variables are appended into a list.

Using the 'pandas' package we were able to convert this list with variables into a large table (dataframe) containing all variables per municipality. This dataframe, in turn, is converted into a csv file. All the data we scraped from the Huizenzoeker platform was data that was directly observable in the form of raw text.

3.2 What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated? We scraped the data using Python's programming software in Jupyter Notebooks. By loading the packages BeautifulSoup, Selenium, requests, re, pandas, time, webdriver manager, and json, we were able to use functions allowing for our specific webscraping steps.

Huizenzoeker.nl does not provide an official software API (anymore), so we scraped the data by writing code ourselves.

3.3 If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)? Technically, we have taken the entire population, and no sample, to conduct our project with. We took all the municipality pages as input, an not a portion of them.

Yet, logically, we have taken a sample. Namely, a single unit would represent a single house in logical terms. However, as the statistics we were after were only available on an average-level on the municipality pages, we took the municipality pages as single units. A municipality page consists of average numbers from all the single houses present in that region. Thus, that is the sampling strategy applied.

3.4 Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)? In the data collection process the team members of this project were involved. In addition, Hannes Datta, professor at Tilburg University and program creator of the Online Data Collection and Management course was involved in the data collection process in the form of guidance and supervision.

3.5 Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? Huizenzoeker.nl covers the housing market data of September 2021. This is the most recent housing market data. Huizenzoeker.nl shows this most-recent data because the housing market changes every month (e.g., houses are sold, new houses are offered, the asking price may be more extremely outbid in one month than in the other month, etc.).

3.6 Were any ethical review processes conducted (e.g., by an institutional review board)? This question is not applicable to our dataset.

3.7 Does the dataset relate to people? If not, you may skip the remaining questions in this section. This question is not applicable to our dataset.

3.8 Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)? This question is not applicable to our dataset.

3.9 Were the individuals in question notified about the data collection? This question is not applicable to our dataset.

3.10 Did the individuals in question consent to the collection and use of their data? This question is not applicable to our dataset.

3.11 If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate). This question is not applicable to our dataset.

3.12 Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted? This question is not applicable to our dataset.

Preprocessing, cleaning, labeling

4.1 Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? First of all, all the values of the variables have been cleaned in a way that they only give a certain numeric value or percentage as output (no additional words, and only consistent punctuation). This means removing the HTML tag words, stripping out unnecessary characters and retaining relevant substrings only. To achieve this, we made use of regular expressions (regex) to pre-process the textual data. When no numeric value exists for a specific municipality, we encoded that 'NA' will result as output for the variable in question. Furthermore, all the variables have been assigned a clear label, such that the numeric values are given a meaning. For example, we identified values as provinces, municipalities, streets, etc. Additionally, all the variables have been displayed in a table against all the municipalities/provinces as a small start in preprocessing.

In addition, we use pagination to be able to scrape data on street-level. Interaction (*i.e.*, in the form of clicking buttons) is required since the large amount of streets for each residence is displayed on several page numbers.

4.2 Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? Yes, the raw output is being saved in a json file automatically, as part of our coding script. The json file can be accessed by running our Scraping Woningmarkt (Final Code) jupyter script. In addition, we stored our dataset locally, being able to monitor the process of our data collection process more easily (Boegershausen et al., 2021). Our data is stored in files rather than databases. The dataset is structured (*i.e.*, it is downloaded as a clean pandas Dataframe) and the code can be run on one computer (*i.e.*, the size of our dataset is not exhaustive).

4.3 Is the software used to preprocess/clean/label the instances available? We decided to use self-developed code that interfaces with high-level scraping libraries (e.g. Selenium and BeautifulSoup), as a software tool for data extraction. We did not choose to use a ready-made scraping toolkit like Monzenda, or packages that only require some coding like Scrapy for Python, as for the complexity of our data collection the self-developed code method seemed the most desirable. Developing the code ourselves in Python required quite some time and effort, however in the end it is a better way to actively manage the data quality and reproductibility than through the other methods.

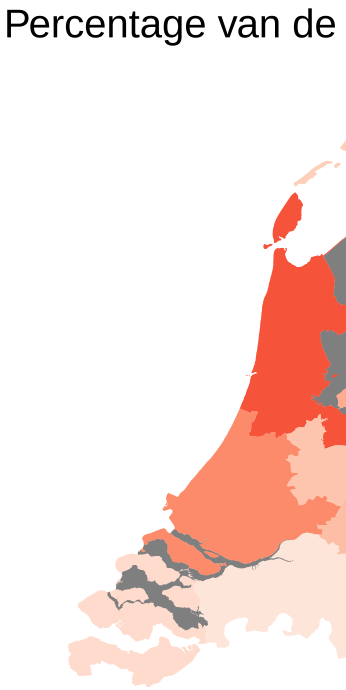
After preprocessing, cleaning, and labelling the data in Python, we exported the dataset to RStudio where we transformed the dataset into one ready for analysis. Python and Rstudio, and the libraries Selenium and BeautifulSoup are all publically available. So, together with our code, you can replicate our scraping efforts easily.

Uses

5.1 Has the dataset been used for any tasks already? The dataset has not been used for scientific research yet. However, to illustrate how one could use our dataset we provide several engaging figures.

```
library(ggplot2)
library(maps)
library(viridis)
library(tibble)
library(sp)
library(extrafont)
library(readr)
library(tidyr)
library(tinytex)
library(dplyr)
huizenzoeker <- read_csv("../data/huizenzoeker_province.csv")
```

```
ggplot(final_map) +
  theme_minimal() +
  geom_polygon(aes(x = long, y = lat, group = group, fill= perc_overboden)) +
  labs(x="", y="", title="Percentage van de vraagprijs overboden") +
  theme(plot.title = element_text(hjust = 0.5,size=15),
        axis.title.x=element_blank(),
        axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        axis.title.y=element_blank(),
        axis.text.y=element_blank(),
        axis.ticks.y=element_blank(),
        legend.position = "bottom") +
  coord_map()+
  scale_fill_distiller(name = "% overboden",
                       palette = "Reds",
                       direction = 1) +
  theme(legend.position = "right",
        panel.grid.major = element_blank(), panel.grid.minor = element_blank(),
        panel.background = element_blank())
```



In which provice of the Netherlands are houses currently most outbid?

5.2 Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point. There are not many (if any) papers or systems that use this dataset, so there is not really such repository. On Github, we found some respositories for:

- A simple python wrapper for the Huizenzoeker API (but last updated in oct 2013) = https://github.com/bpeschier/huizenzoeker
- Using the Jaap API to look for rentals in Rotterdam = https://github.com/thomasvt1/HuizenZoeker

5.3 What (other) tasks could the dataset be used for? Broadly speaking, a suitable task this dataset can be used for is helping (future) inhabitants of the Netherlands find their ideal home. By accessing our data, a person could find the best municipality to live in for this person's specific circumstances (e.g. specific disposable income level), find a region where the value-for-money seems to be of high standard, to help them in negotiations on the price, to help them find out what is the norm in terms of overbidding for each municipality, and more.

5.4 Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? The only harms that could be done in the case of Huizenzoeker relates to financial harms, e.g. when one is misinformed about housing prices due to our dataset.

However, as long as Huizenzoeker does not incur drastic changes, no undesirable harms will arise. *also when there would be drastic changes on Huizenzoeker.nl then our code likely won't work anymore so in that case our scraper won't result in undesirable harms either, but just won't work)

Future users of the dataset could decide to implement more variables, or kick certain variables out. As long as this is done following the same steps as in our coding script, no harm can be done.

5.5 Are there tasks for which the dataset should not be used? If so, please provide a description.
The dataset can be used for any matters regarding the housing market in the Netherlands, at municipality level as well as province level. For anything outside of this topic, the dataset has no use.