The housing market in the Netherlands: a Fun(da) analysis

1.1 Motivation

The dataset is created by Team 2 of the course Online Data Collection and Management (oDCM), on behalf of Tilburg University. There was no financial funding involved in the creation of this dataset.

The housing shortage in the Netherlands in the last few years has led to a serious housing crisis experienced today. Especially young people are encountering major difficulties in finding the right place for an affordable price. Also during the last political party elections, the housing crisis was a very much debated topic. For this reason, Team 2 realized the significance of creating a dataset that can give both house seekers and house sellers more insights into the housing market in the Netherlands. Mainly, the aim of creating this dataset is to make a house's selling duration more predictable. For both parties this can lead to a more efficient housing market: house sellers can better predict for what period their house will be on sale, whilst house seekers get insights into the selling rate in a particular city. On top of historical selling prices and selling durations, the dataset includes all other information that was initially provided on the house product page. This means that information about a house's roof type, number of floors, energy label or even location of garden is also included. This can give house seekers a general idea of what a house with their preferences could cost.

When considering websites to acquire housing market data from, Team 2 looked at different private real estate websites. However, due to the great amount of real estate agencies in the Netherlands, scraping all information would be too time consuming. Therefore, Team 2 decided to scrape www.funda.nl, which is a platform on which real estate agencies from all over the Netherlands present properties that are for sale. The houses that get sold remain on the platform for up to one year. Funda is the most used real estate website in the Netherlands, and has no restrictions such as licenses or usage fees. It has the largest offer on houses for sale and provides a lot of information on each house. Therefore, Funda was chosen over other real estate platforms, such as www.zah.nl. In terms of data availability, houses are listed on city-, province- or even national level. The information that is available for each house on Funda can be accessed by clicking on the URL of the house. When doing so, the product page of the house appears, on which all the different variables can be found.

1.2 Composition

To create the dataset, information from the relevant Funda page with the sold houses is scraped. This is possible on city-, province-, and national level, and can be achieved by adapting the link below for the wanted city or province:

https://www.funda.nl/koop/INSERT_CITY_NAME/verkocht.

Our dataset is focused on the city of Eindhoven. Every data entity in it represents a house in Eindhoven that has been sold in the past year. The maximum retrieval time on Funda is one year, and unfortunately no archival versions of all the data on the sold houses is available. This causes the website to change over time, as sold houses are added daily and others are taken off once their selling date surpasses a year. Whilst our dataset is fully representative in the sense that it geographically covers Eindhoven, it lacks full representation on the time dimension due to the limited maximum retrieval time. Our dataset, formed on March 22, 2021 represents Eindhoven with 2,827 entities. As explained earlier, this research can be extended to any Dutch city, province- or national-level. Regardless of the chosen location, the majority of the entities in the resulting dataset contain information on the variables listed below:

- Address (‘Adres’)
- Offered since (‘Aangeboden sinds’)
- Sale date (Verkoopdatum’)
- Duration time (‘Looptijd’)
- Selling price (‘Laatste vraagprijs’)
- Price per m2 (‘Vraagprijs per m2’)
- Status (‘Status’)
- Kind of house (‘Soort woonhuis’)
- Type of construction (‘Soort bouw’)
- Construction year (‘Bouwjaar’)
- Roof-type (‘Soort dak’)
- Living area (‘Wonen’)
- External storage space (‘Externe bergingsruimte’)
- Percel (‘Perceel’)
- Volume (‘Inhoud’)
- Number of rooms (‘Aantal kamers’)
- Number of bathrooms (‘Aantal badkamers’)
- Bathroom provision (‘Badkamervoorzieningen’)
- Number of floors (‘Aantal woonlagen’)
- General provisions (‘Voorzieningen’)
- Energy label (‘Energielabel’)
- Isolation (‘Isolatie’)
- Heating system (‘Verwarming’)
- Warm water (‘Warm water’)
- Surface area (‘Oppervlakte’)
- Ownership situation (‘Eigendomssituatie’)
- Location (‘Ligging’)
- Garden (‘Tuin’)
- Garden area (‘Achtertuin’)
- Location of garden (‘Ligging tuin’)
- Shed/storage room (‘Schuur/Berging’)
- Type of parking (‘Parkeergelegenheid’)
- Type of garage (‘Soort garage’)
- Boiler (‘CV Ketel’)

However there are instances in which entities only contain a small number of variables, resulting in some missingness in the dataset. The data is available in a CSV file, containing a list of all entities (addresses) with their respective variables. Inspecting the data revealed 2,827 entities, which means that information about 2,827 sold houses in Eindhoven is present in our dataset. Depending on the entity, additional variables can be found in the dataset. However, these are very rare and therefore not listed above (for example: 'Balcony/ Rooftop'). Information missing from all individual entities are the postal codes, the city name and the province. This information was unavailable to scrape from a Funda housepage (because xpath, tag name and class name were too dynamic). When looking to extend this research to multiple cities, provinces or to the Netherlands as a whole, it is essential to add a column with the city and province name in order to distinguish between them. However, our current research is about collecting information about Eindhoven, so we are sure that each entity (address) in this dataset is located in Eindhoven.

1.3 Technical extraction plan

To collect the data, the following link was adapted accordingly: https://www.funda.nl/koop/INSERT_CITY_NAME/verkocht/. It can be changed to the desired city, province, or the Netherlands as a whole. The resulting dataset contains the addresses of all the houses sold in the past year. The directly observable data (from adapting the above link accordingly) were the addresses and individual URL links for each house. When clicking on the respective URL, more information variables became readily available. Because Funda is a dynamic website, we used Selenium in order to find elements by their tag name, class name and xPath. Using Selenium, a new chromedriver was launched that navigated to the respective city page (for Eindhoven) and could accept cookies automatically. A problem that arises here is the reCAPTCHA verification, which we could not bypass by code. To create some clarity for future users, Jupyter Notebook refreshes the page and prints 'please fill in the reCAPTCHA' when the warning message appears. When scraping the list of URLs, we found that all URLs were surrounded by <a> tags, of which we acquired the 'href' tag. Using a while loop with condition= True, our code scrapes all 'href' instances on a page, and clicking the 'next' button on the page, until condition= False is reached. This means that there is no 'next' button on the page anymore (so you have reached the final page). After storing all URLs in the variable my_urls and verifying it with the length of my_urls with the total number of houses sold in a city (which should be equal) (found by it's xpath), we looked at house-specific variables. When inspecting the house page we found that all values were surrounded by <dd> tags and all elements were surrounded by <dt> tags, which made it easy to acquire them. However, some unwanted information was also stored by <dd> and <dt> tags, which we eventually removed by 'if' conditions (if 'X' is in list: remove 'X'). Additionally, we appended the values list by adding the address of a house (found by class name) and adding 'address' to the elements list. Eventually, we stored the two lists in a .json dictionary. The last step in the extraction process was to execute the house page extraction function for all URLs stored in my_urls. We used a for loop to do so, which opened a new .json file (cityname.json) and stored all .json dictionaries per URL in it. To check the progress of the for loop we used tqdm, which complemented our scraper with a super cool progressbar! Finally, we stored the newly acquired .json file in a dataframe using Pandas, which we converted to a .csv file (cityname.csv).

In terms of potential legal and/ or ethical concerns: Funda explicitly prohibited web scraping. Moreover, there have been some lawsuits against companies trying to scrape Funda[1]. However, because this is an educational project, without any further future actions/ implications, no laws or ethical boundaries will be crossed.

1.4 Preprocessing

At first we encountered misaligning problems in the obtained CSV file, resulting in an unreadable final dataframe. To tackle this, we cleaned the data on the fly in our code. For example, if a scraped value was 'Onbekend' or ' ', it was removed. Additionally, we created a list with the values that caused misalignment. Then a condition was put in place that automatically removed these variables if they were present in the scraped values. From the elements list we deleted the element 'Gebruiksoppervlakten', which was one of the main ones causing misalignment. Throughout the process of data collection, we decided not to save the 'raw' data, because it could not be interpreted due to this heavy misalignment. Renaming columns and deleting irrelevant ones are steps we pursue in the dPrep analysis of our dataset.

As mentioned in 1.2., the dataset contains some missing values spread across the variables. However, because the level of missingingness is very low and often seen in the less obvious variables, entities with values for the first 10 variables will remain included in the dataset regardless of subsequent missing values. This is opted for because the first 10 variables are regarded as the more obvious and influential variables when making a housing decision or comparing houses.

[1] http://sync.nl/andermans-site-scrapen-wanneer-mag-dat/3

1.5 Uses

The dataset has not been used for any external tasks yet. However, it has been used by Team 2 to perform its own research about the Dutch housing market. Firstly, an app/website can be created which gives users insight in the average selling duration based on the price of their house. In other words, what selling duration time can they expect when they want to sell their house for €X in Eindhoven? This could even be extended to other variables ('if my house has 4 floors, what is the average price it has been sold for in the last year?'). Secondly, exploratory research on the housing market within Eindhoven can be conducted by looking at house prices, prices per square meter, selling duration, number of rooms etc. After all, the code can be used for different cities, or provinces meaning that the collected information can also be used for comparisons across different regions.

The Dutch housing market is very volatile, which means that the assumptions made from the acquired dataset should be taken in with high consideration. The conclusions that can be drawn from the results are not constant throughout the timeframe of one year. For example, whilst a house with 4 floors might have been worth €235.000 in October 2020, it could be €280.000 in October 2022. Therefore house-sellers can base the price they want to ask on this dataset, but not solely as there are always external factors that can play a role. This dataset should not be used to generate financial rewards. Funda is a protected data bank and content from this data bank should not be displayed on other websites or be used to create own housing search engines. Violation of this has previously resulted in legal consequences with the Dutch Database right (Databankenrecht) and may happen again if done so in the future.