1.Motivation of data collection (why was the data collected?)

This dataset is created by Team 2 of the course Online Data Collection and Management (oDCM), on behalf of Tilburg University. 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 a place for an affordable price. For this reason, Team 2 realized the potential significance of creating a dataset that gives insights into for example: selling durations, historical selling prices, price differences according to seasons and price per square meter in a particular city.

2.Composition of dataset (what's in the data?)

One entity represents one address. For all entities the following variables are in the dataset:

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’)

3.Collection process (how was the data collected?)

The dataset is scraped from the information that is listed on the Funda page for a particular city, namely all the houses that were sold in the last year: https://www.funda.nl/koop/INSERTCITYNAME/verkocht/sorteer-afmelddatum-af/.

4.Preprocessing/cleaning/labeling (how was the data cleaned, if at all?)

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.

5.Uses (how is the dataset intended to be used?)

Firstly, 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. Secondly, 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?’). Thirdly, the code can be used for different cities, or provinces meaning that the collected information can also be used to compare houses among various dimensions on different scales.

6.Distribution (how will the dataset be made available to others?)

This dataset can only be acquired via email, send your request to w.n.r.vanakkeren@tilburguniversity.edu

7.Maintenance (will the dataset be maintained? how? by whom?)

This dataset will be updated monthly by Wouter van Akkeren by scraping newly added houses to www.funda.nl/koop/heel-nederland/verkocht from month to month