0.1 Data cleaning

The scraping of Boliga left us with 65,950 observations. However, big data is often dirty and requires tidying before it can be used in any meaningful statistical context (Salganik 2018). In the following sections we describe our data cleaning efforts.

0.1.1 Initial clean-up

The initial clean-up sorted out rows that contained an illogical value. Here we focused our efforts on the removing any and all observations that contained a municipality code of 0. Additionally, we chose to exclude any real estate valued below 100 DKK, while these listings existed on Boliga's website, they were all listings, where the seller wanted to sell to the highest bidder. We excluded these listings on the basis that they were unrealistically priced compared to the property's real market value.
Boliga subdivides its listings into ten real-estate types. We chose to exclude the typification "other", as there were only 17 houses listed in the category - too few for training our machine learning model. Furthermore, we also removed any listing without coordinates as these were vital for the scraping process on hvorlangerder.dk.
We dropped all observations with an unreasonably high "liggetid", as we saw these instances to not represent reasonable pricing or demand. We have thus set an arbitrary limit of 3 years (1,095 days), and omitted any observations that has been on the market for longer than that. This results in the omission of approximately 5% of our dataset. The highest mean "liggetid" on a municipal level was roughly 600 days, so as not to discriminate against the observations from the municipalities with a longer average 'sales time' we set the cut-off somewhat higher than the highest mean. Another option was to set the limit according to each municipality's mean, a solution that was a bit more time consuming, and would have yielded approximately the same result, hence we opted not to do it. In figure XX the number of rows dropped at each step of the cleaning can be perused.

Feature	n
Municipality code	231
Price	11
Coordinates	275
Type	17
Liggetid	3,940
Total[1]	4,332

0.1.2 Ejerlejlighed's lot size

Inspecting our data we found that from 8,028 apartments 1,176 of these had a lot size. These listings included the apartment complex' common area, whereas the rest did not. To overcome this discrepancy we chose to set all non-zero values to zero, as not to confuse our model with an inconsistent feature (Raschka & Mirjalili 2017).

0.1.3 Final touch-ups

After merging the housing info with the demographic data and the distances from hvorlangterder.dk, we removed a single duplicate and deleted all columns not relevant for our analysis. After the clean-up of the data we were left with a dataset consisting of 61,618 observations and 37 features and 1 target variable.