Own Data – Analysis Report

Samuel William Nehrer

**Public Housing in Aarhus**

**Introduction**

Figuring out what drives the cost per square meter in public housing can be challenging. Listings often give just the basics—location, size, perhaps the building’s energy rating—but they don’t always explain why one apartment costs more per square meter than another. To get a better sense of what really matters, this brief report looks at a dataset of public housing listings in Aarhus. Existing research on European housing markets shows that rental prices are influenced by a combination of structural characteristics and a variety of location-based factors. For example, housing prices are influenced by proximity to urban centers and structural characteristics such as area and number of rooms (Zietz et al., 2006). The compact and urban-friendly character of smaller apartments frequently leads to a higher demand, which in turn drives an increase in price per square meter (Dubois & Nivakosk, 2023). Moreover, the significance of location has been found to be best predictor of price per square meter (Melecky & Paksi, 2024). This report looks at these patterns in Aarhus, Denmark, with the aim of getting a basic overview about which factors influence public housing prices there. The data for this overview come from the City of Aarhus’s (Aarhus Kommune) open data portal and consist of the latest master data from all housing organizations in Aarhusbolig (*Aarhus Residence - Dataset*, 2024).

**Methods**

Our dataset included 44,090 records, with each one representing an individual public housing unit in Aarhus. Each unit was assigned identifiers that connecting it to a particular flat, department, and company. This included details about property and flat types, financial attributes like deposits, net and gross rents, and the related net and gross prices per square meter. Additionally, it covered structural features such as the number of rooms, total area, and BBR area (the officially registered area of the property). Each record also came with a postal code and city, which allowed us to look at how housing prices varied by location.

As our dataset required data processing before we could run any analysis, we first carried out minor data wrangling and subsequent descriptive analysis in Python. The data wrangling included conversion of Danish column names to English, processing of the numeric fields to maintain consistent decimal formatting and removed entries with missing data or unrealistic values (e.g., apartments with zero area). Additionally, we recalculated the price per square meter. For the data analysis part, we first ran basic descriptive analysis and created visualizations (histograms, boxplots, and correlation matrices) to identify important trends and relationships in our data. Next, we fitted two regression models in R with deposit, city, area, and rooms as predictors to find out which of these are the best predictors of net price per square meter.

**Results**

1. *Descriptive Statistics*

The housing units included in our dataset (n = 44,090) have a mean of 2.71 rooms (median = 3.0, std = 0.99) and a mean area of 76.91 m² (median = 79.0, std = 23.05). The mean deposit is 14,608.12 DKK (median = 13,420.0, std = 8,929.56). The net price per square meter averages 61.79 DKK (median = 60.2, std = 11.16), while the gross price per square meter averages 74.29 DKK (median = 72.9, std = 12.12).

1. *Exploratory Analysis*

The created scatterplots (Figure 1.) show a general positive trend in gross and net rent as a function of area and number of rooms. However, although both increase with area and the number of rooms, the gross rent appears to rise at a different rate. This might suggest that the additional charges (e.g., utilities) not included in the net rent do not grow proportionately with the apartment’s area.

A group of images of different colors

Description automatically generated with medium confidence

Figure 1.

1. *Regression*

The results of the first fitted linear regression *(formula: net\_price\_per\_sqm ~ rooms + area + deposit + city)* were as follows:

The range of the residuals was from -51.533 to 84.507, with median of -0.601 and an interquartile range of -5.547 to 4.618. We found a strong statistical significance of area (β=−0.2462, SE = 0.0044, t = -56.190, p<2x) and deposit (β=0.0003577, SE= 0.0000056, t = 64.329, p<2x). The coefficients for city variables were significant for all locations except for Tilst (β=-0.3048, SE= 0.3059, t = −0.996, p= 0.3190). The number of rooms was not significantly associated with our outcome variable (β=−0.1774, SE=0.0984, t=−1.803, p=0.0713).

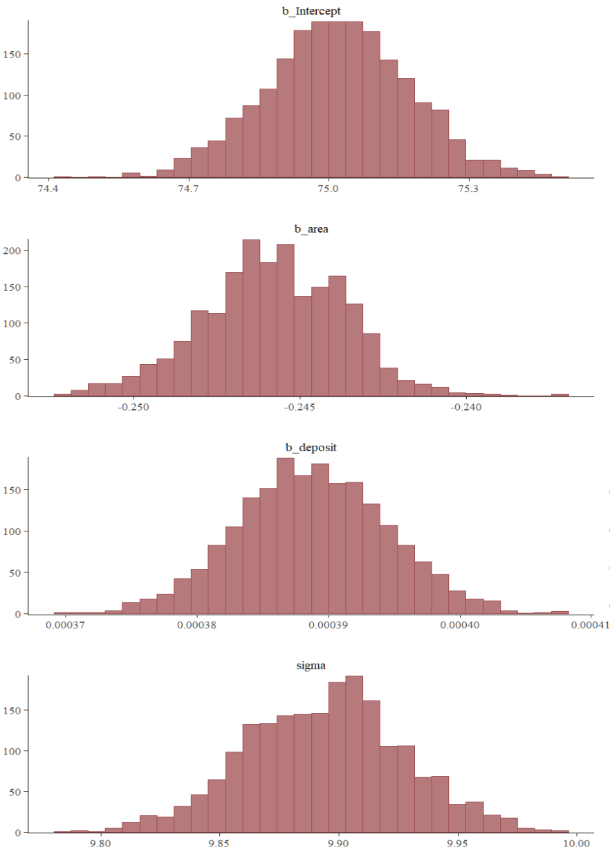
In the second part of our regression analysis, we fitted a Bayesian regression model (*formula: net\_price\_per\_sqm ~ area + deposit).* We assumed a Gaussian distribution for the outcome variable with identity link function. The results of our fitted model showed credible evidence for area having a small effect on net price per square meter, with each additional square meter decreasing the net price per square meter by -0.25 (95% CI: [-0.25, -0.24]). We found no evidence for an effect of deposit on our outcome variable. (See Figure 2. for posterior distributions of the estimates.)

Figure 2 – Posterior Estimates

**Discussion**

Our analysis of the data shown here suffered several methodological limitations. First, our analysis does not generalize over all housing properties in Aarhus since our dataset contained only the listings by housing associations. Second, our analysis might have been skewed by the presence of identical or near-identical listings present in our dataset. These identical listings were practically difficult to handle due to the size and complexity of our dataset. Third, our dataset was imbalanced in terms of locations since several areas only listed a small number of houses with disproportionately higher prices than average.

When it comes to our Bayesian regression model, the Gaussian family was not an ideal choice given that our outcome variable exhibits a distribution closer to Gamma or Log-normal. However, we opted for the implementation of the Gaussian family due to the poor chain convergence in the fitted Gamma and Log-normal models.

In conclusion, the results shown here can only be understood as complementing the results of earlier studies. Among the more relevant factors we could have included would be the relative impact on the net price of the stratification of the location to hedonic (e.g., proximity to nature, beach, forests) or functional (amenities in the area like groceries, train stations).

**References**

*Aarhus residence—Dataset*. (2024). Retrieved December 7, 2024, from https://www.opendata.dk/city-of-aarhus/aarhusbolig?fbclid=IwZXh0bgNhZW0CMTEAAR0ra7fv\_hHwOB\_lj1e4rwdPfQNy3C0n0Vqf2M3Ska\_gAjQPXodSzwH35-g\_aem\_JGFGaEINLyZiEa-jfv2g1w

Dubois, H., & Nivakosk, S. (with European Foundation for the Improvement of Living and Working Conditions). (2023). Unaffordable and inadequate housing in Europe. Publications Office of the European Union. https://doi.org/10.2806/715002

Melecky, A., & Paksi, D. (2024). Drivers of European housing prices in the new millennium: Demand, financial, and supply determinants. Empirica, 51(3), 731–753. https://doi.org/10.1007/s10663-024-09611-5

Zietz, E., Sirmans, S., & Sirmans, G. (2006). The Value of Housing Characteristics: A Meta Analysis. The Journal of Real Estate Finance and Economics. https://www.academia.edu/21853990/The\_Value\_of\_Housing\_Characteristics\_A\_Meta\_Analysis