Introduction to Cultural Data Science

Jacob Lillelund

Portfolio 1

Concordance - Link to GitHub:
https://github.com/jaco4873/cultural-data-science-AU/tree/main/intro-to-cultural-datascience/concordance

Fun with Pandas - Link to GitHub:
https://github.com/jaco4873/cultural-data-science-AU/tree/main/intro-to-cultural-datascience/fun-with-pandas

Portfolio 2

Correlation and the linear model - Link to GitHub:
https://github.com/jaco4873/cultural-data-science-AU/tree/main/intro-to-cultural-datascience/correlations-and-the-linear-model

Portfolio 3

Analysis of own project data - Link to Github:
https://github.com/jaco4873/cultural-data-science-AU/tree/main/intro-to-cultural-datascience/portfolio-3-housing-prices

Public housing in Aarhus

- Problem: It's hard to get a proper overview of what matters in public housing rent by just looking at listings.

- Question: What factors determine the price per square meter in public housing in Aarhus

What is known?

- Housing prices are influenced by both structural characteristics (e.g., area and number of rooms) and locational factors (e.g., proximity to amenities or urban centers) (Sirmans et al., 2020).

- Smaller apartments tend to have a higher price per square meter due to greater demand for compact & urban-friendly housing (European Commission, 2022)

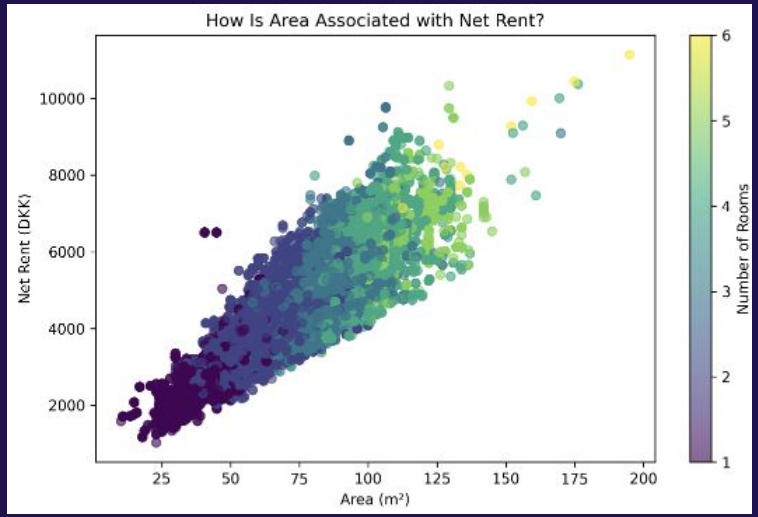
- Location is one of the most significant predictors of housing prices. Proximity to central business districts and desirable neighborhoods is driving up prices (Melecky & Paksi, 2024).

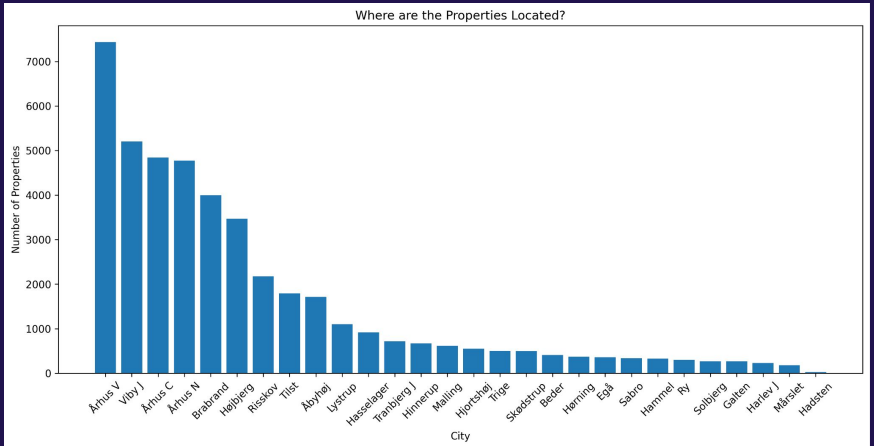
Slide: Jonathan Laursen

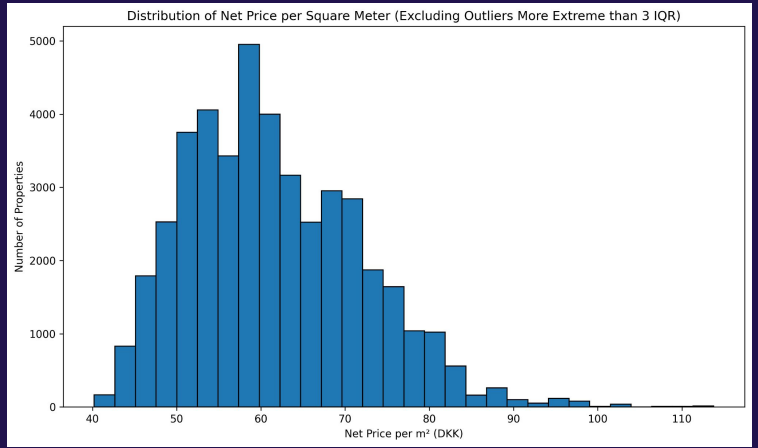
Methods & Data

- Data source:
  - Aarhus Kommune, 2024
  - All public housing in Aarhus: 44090 data points
  - Data cleaning and transformation

- Descriptive Analytics
- Linear Models (linear regression and Bayesian)
  - Outcome variable: price pr. sqm
  - Predictors: Area, rooms, deposit, city







Slide: Jacob Lillelund

Results

- Smaller properties - higher price pr. sqm
- Area and rooms correlate
- Area and number of rooms is associated with some variance in data
- Deposit have no effect
- Postal code is associated with price/sqm

Linear Regression Model

```
Call:
lm(formula = net_price_per_sqm ~ rooms + area + deposit + city,
    data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-51.533  -5.548  -0.601   4.618  84.507

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.075e+01  2.566e-01 275.762  < 2e-16 ***
rooms           -1.774e-01  9.837e-02  -1.803   0.0713 .
area            -2.462e-01  4.381e-03 -56.190  < 2e-16 ***
deposit          3.577e-04  5.561e-06  64.329  < 2e-16 ***
cityÅrhus C      1.203e+00  2.546e-01   4.724 2.31e-06 ***
cityÅrhus N      2.266e+00  2.556e-01   8.867  < 2e-16 ***
cityHarlev J     2.130e+01  6.318e-01  33.712  < 2e-16 ***
```

Slide: Jonathan Laursen

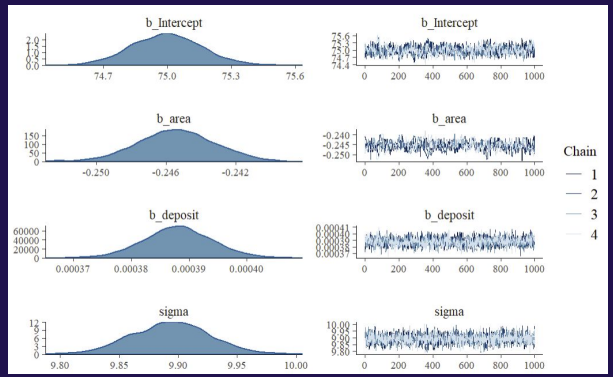
Bayesian Generalized LM

```
prior = c(
set_prior("normal(100, 30)", class = "Intercept"),
set_prior("normal(0, 4)", class = "b", coef = "area"),
set_prior("normal(0, 4)", class = "b", coef = "deposit")
```

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: net_price_per_sqm ~ area + deposit
   Data: d (Number of observations: 44090)
  Draws: 4 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 4000

Population-Level Effects:
          Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
Intercept    75.01      0.16    74.71    75.33 1.01      580     1140
area         -0.25      0.00    -0.25    -0.24 1.01      412      492
deposit       0.00      0.00     0.00     0.00 1.00     1632     1912

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     9.89      0.03     9.83     9.96 1.00     2670     2107
```



Discussion

The typical market dynamics are restricted in public housing schemes

Imbalanced data with regards to postal code - careful interpretation
- E.g.: Harlev J - high prices, but only recently built apartments are listed

Handling of identical or near identical listings

Interesting variables for future work:

- Year build/latest renovation
- Distance from city centre
- Amenities in area (child care, groceries, parks etc.)
- Other structural characteristics
- Length of waiting lists for each apartment complex

Determinants of Public Housing Rent in Aarhus

Introduction

The primary aim of this project is to explore the determinants of public housing rent prices in Aarhus, Denmark. Specifically, we focus on price per square meter (“price/sqm”) and how it relates to structural characteristics like area, the number of rooms, and deposit requirements. Public housing rents deviate from typical market dynamics due to regulatory factors and location-specific features. Previous research highlights the following trends:

1. Housing prices are influenced by structural characteristics and proximity to urban centers (Sirmans et al., 2006)
2. Smaller apartments often command higher price/sqm due to greater demand for urban-friendly living spaces (European Commission. Directorate General for Economic and Financial Affairs., 2022).
3. Location remains a crucial determinant, with proximity to amenities driving up prices (Melecky & Paksi, 2024).

Given these findings, we investigate how these variables affect rents in Aarhus public housing and evaluate the utility of Bayesian analysis for this purpose.

Methods

Data Collection and Cleaning

The dataset after cleaning consists of 44,090 public housing entries from Aarhus Kommune (2024). Data cleaning steps included:

1. Removing missing or zero values for essential variables such as area, rent, and deposit.
2. Calculating derived variables: net_price_per_sqm and gross_price_per_sqm.
3. Handling outliers, e.g., properties with unrealistic rents (<1000 DKK or price/sqm <40 DKK).

The cleaned dataset showed imbalances in variables like postal codes, limiting generalizability.

Operationalization

Key variables:

- Dependent Variable: Net price per square meter.
- Independent Variables: Area (m²), number of rooms, and deposit amount.

Model Description

We employed two modeling techniques: linear regression and Bayesian generalized linear modeling (Bayesian GLM). The Bayesian approach incorporated weakly informative priors:

- Intercept: Normal(100, 30).
- Area coefficient: Normal(0, 4).
- Deposit coefficient: Normal(0, 4).

Data preprocessing was performed using Python, while model fitting was done in R.

Results

Descriptive Analysis

- Smaller apartments had a higher price/sqm.
- Area and number of rooms were positively correlated (0.62).
- Deposit amount showed minimal variation and no significant effect on rent.

Bayesian GLM Output

The Bayesian GLM confirmed the findings of the standard linear regression:

- Intercept**:** 75.01 (±0.16), representing the baseline price/sqm when all other variables are zero.
- Area coefficient: -0.25 (±0.01), indicating a decrease in price/sqm with increasing area.
- Deposit coefficient: 0.00 (±0.01), suggesting no impact on rent.
- Model fit diagnostics ($\widehat{R} \approx 1$) indicated convergence.

Visualizations

We created multiple visualisations of the data to aid in the exploratory data analysis.

Figure 1: Area vs. Gross Rent

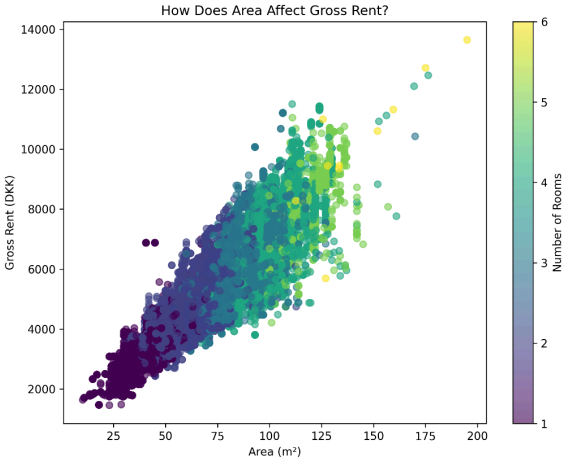


Figure 2: Area vs. Net Rent

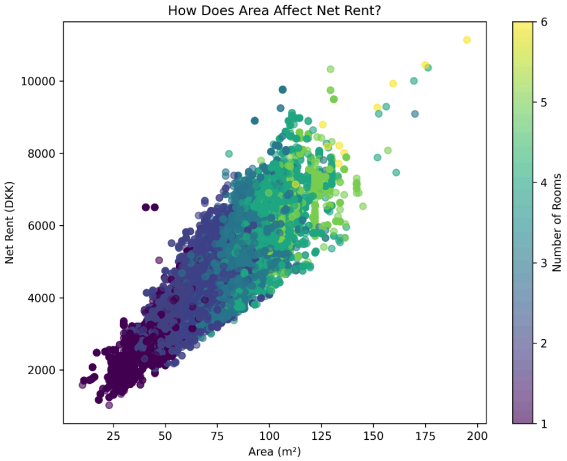


Figure 3: Distribution of Apartment Types

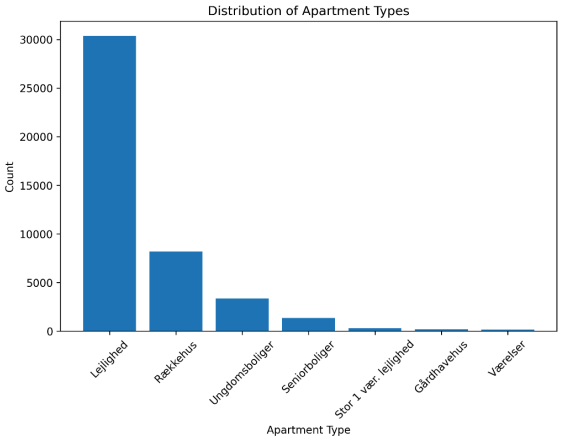


Figure 4: Distribution of Number of Rooms

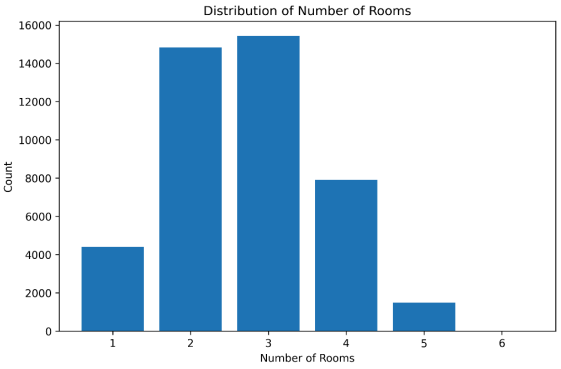


Figure 5: Distribution of Net Price/Sqm

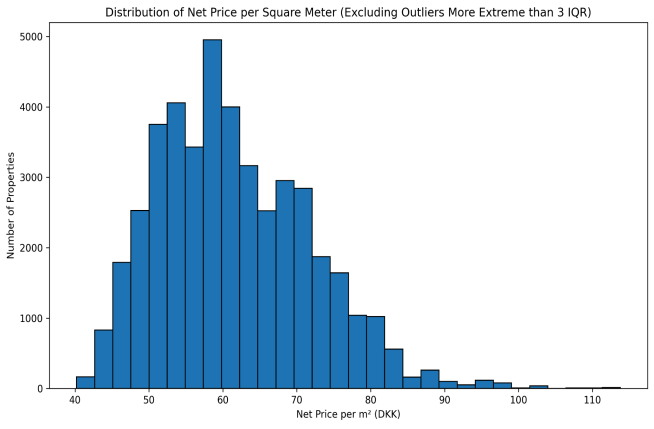


Figure 6: Correlation Matrix of Variables

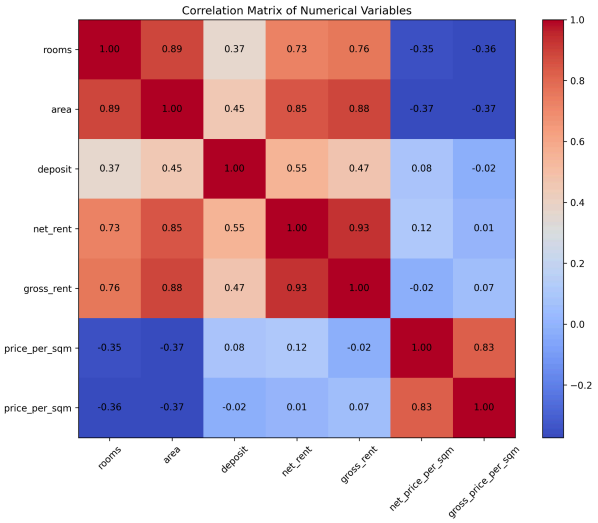


Figure 7: Distribution of Net Rent

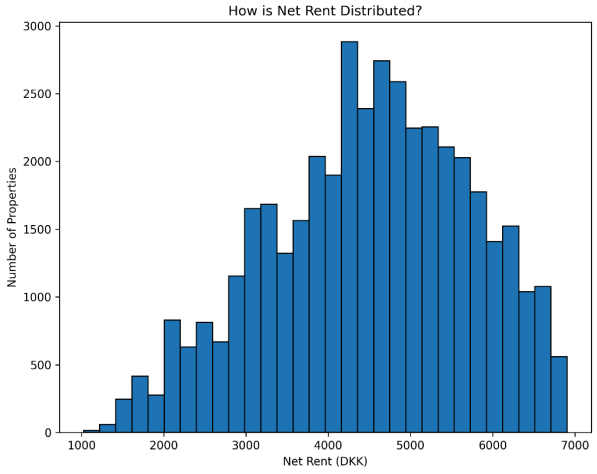
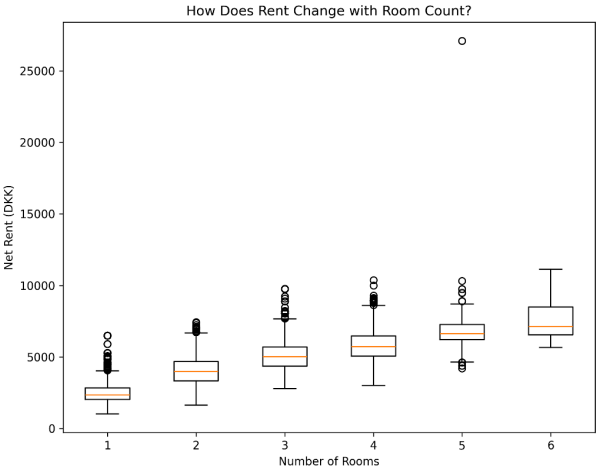


Figure 8: Box Plot of Net Rent Based on Rooms



Discussion

Our findings support the hypothesis that structural characteristics like area and the number of rooms significantly influence public housing rent in Aarhus. However, deposit amounts and smaller cities within Aarhus Kommune showed no substantial effects. These results align with prior studies emphasizing the role of apartment size and urban demand.

Limitations

- Data Quality: Imbalanced representation of certain postal codes (e.g., Harlev J).
- Operationalization Issues: Key factors such as distance to city center and renovation year were not included.

Conclusion

The Bayesian approach offered inferences about rent determinants. Future research should incorporate locational and temporal features to better predict rent variability in public housing markets.

References

European Commission. Directorate General for Economic and Financial Affairs. (2022). *Housing market developments in the euro area: Focus on housing affordability.* Publications Office. https://data.europa.eu/doi/10.2765/74242

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

Sirmans, G. S., MacDonald, L., Macpherson, D. A., & Zietz, E. N. (2006). The Value of Housing Characteristics: A Meta Analysis. *The Journal of Real Estate Finance and Economics*, *33*(3), 215–240. https://doi.org/10.1007/s11146-006-9983-5