The Berlin Real Estate Market

Moritz Wilksch

Summary

Prices for real estate buyers and renters in Berlin have increased drastically over the past decade making it hard to find affordable housing. In this report, I collect data from the three major German online real estate platforms and use hierarchical linear regression models to assess which characteristics of a real estate object (apartment or house) drive its price. The analysis is split into two separate models for rental properties and properties for sale as the findings can differ between the two. Findings suggest that TBD!!!!!

Introduction

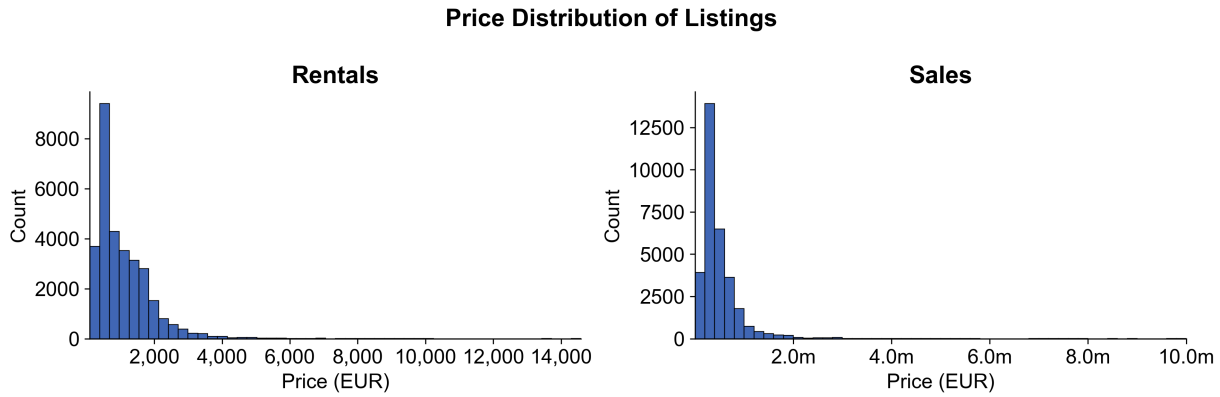
This report aims to identify the characteristics of an apartment or house that are associated with higher and lower prices to shed light on which attributes buyers and renters might look for when searching affordable housing in Berlin. The data used in this study has been collected through web scraping the three most popular online market places for real estate in Germany, `www.immobilienscout24.de`, `www.immowelt.de`, and `ebay-kleinanzeigen.de`. Every listing is characterized using the following attributes: `object_type` (apartment, shared apartment, temporary living or house), `private_offer` (whether the seller is a private or commercial entity), `rooms` (the number of rooms), `square_meters` (the size in square meters), and the `zip_code`. Based on these attributes, I will use hierarchical linear models to model the `price` (separately for properties for rent and sale).

Data and Methodology

The web scraping of the real estate listings was conducted over a period of five moths, from late April 2021 through late October 2021. The raw data for Berlin contains around 72,000 data points. Besides the attributes that are used for this analysis, each listing also has a title and detailed description (both free text). Due to their fromat, they will not be used in this project. As web scraping is a brittle and error-prone process, multiple data cleaning steps are necessary.

EDA

Even after data cleaning the distributions of rental and sales prices are still skewed. This might lead to problems when fitting linear models, but log transforming could be a potential remedy to this issue in the modeling phase. Moreover, the univariate distribution of the number of rooms (*Appendix A*) reveals that properties for sale tend to have more rooms and that rental properties have more missing data in this variable than properties for sale.

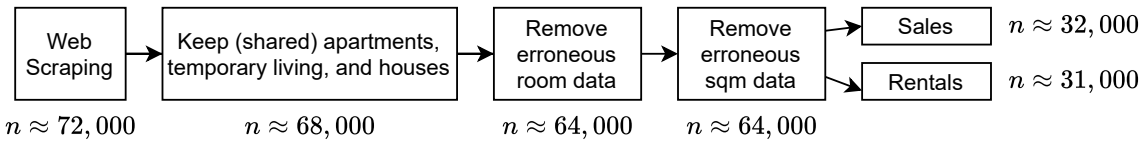


Data Cleaning

After scraping the data, I first focus the analysis on apartments, shared apartments, temporary living and houses. This excludes commercial and retail properties as well as nursing homes and retirement homes, as this analysis is primarily motivated by the scarce market for individual's homes. After conducting some exploratory data analysis, it is evident that some values in the predictor `rooms`, are erroneous. It contains more than 80 levels most of which are wrong, e.g. 990 or other large numbers. This ist most likely the result of the web scraper picking up a different number as the number of rooms. Based on a frequency table of the `rooms` levels, I keep properties with 5 or less rooms and merge the text-based entries into the correct categories (e.g. "single room" $\rightarrow$ 1, "shared room" $\rightarrow$ shared, "not given" $\rightarrow$ missing). Next, I clean the `square_meters` variable. The EDA has shown significant skew with properties listed as having up to 10,000,000 m^2 which once again is the web scraper picking up wrong numbers (maybe the purchase price?). After manual inspection, the problem with skew in `square_meters` originates from an incorrect classification of a properties `to_rent` attribute: Some listings have "rental" prices of up to 1,000,000 EUR and others have "sales" prices of a few hundred EUR. I employ the following heuristic to re-classify these listings based on their price per square meter (`ppsqm`):

Current `to_rent`	`ppsqm`	Current `price`	Reclassify as...
`TRUE`	> 100	$> 10,000$	For sale
`FALSE`	< 250	$< 10,000$	For rent

This heuristic assumes that a) "rentals" that cost more than 10,000 EUR/month and 100 EUR/sqm are actually properties for sale (because no rental property is this expensive) and b) properties that are "for sale" with a listing price of below 10,000 EUR and a `ppsqm` < 250 EUR are actually rental properties. Some properties for sale listings have an unrealistic price of exactly 21,474,836 EUR. Since this is the exact upper boundary of the `int32` data type used to save prices in the database, I will have to assume the price caused an integer overflow and remove these listings from the analysis. This affects 14 listings in total. Finally, to remove data entry errors like unreasonably large apartments, I remove outliers in the `square_meters` variable at the 99.9th percentile for rental properties and properties for sale separately. This only removes a hand full of data points which severely skew the distribution due to erroneous data. The entire data cleaning process and its effect on the sample size n is displayed in the flowchart below.



After this data cleaning process, the only variable that contains missing values is `rooms`. Most other missing values were removed with the removal of the erroneous room data, i.e., most rows that had missing values in other variable before the data cleaning were removed anyways as part of the data cleaning process displayed in the figure above. A potential reason for this correlation in missingness is a malfunction in the web scraper, where a faulty page load or incorrectly defined HTML tags messed up the collection of an entire listing. However, as we can see from the change in sample size throughout data cleaning, this only affected a minority of the data. To cope with the missing data in rooms, I encode "missing" as a separate category to potentially identify characteristics of listings with no given number of rooms later in the modeling phase.

Methodology

To model the relationship between prices and object attributes, I will use separate linear models for rental properties ("rentals") and properties for sale ("sales") to gain interpretable insights. I start by fitting a linear regression model regressing price on all mean effects. I will subsequently explore interactions and assess the validity of the model's main assumptions. In the next step, I will introduce the hierarchical level location (here: `zip_code`). The prices of real estate are expected to vary significantly by area, so this multi-level model should deliver insights into how prices vary by neighborhood. TBD?????

Results

Non-hierarchical Linear Model

The initial model is regressing the `log(price)` on the predictors `object_type`, `private_offer`, `rooms`, and `square_meters`. The log transformation of the dependent variable is necessary, as the price is log-normally distributed, so not transforming it leads to severe assumption violations. This initial model achieves an $R^2 \approx 0.66$ after removing three high-leverage points. Thus, it already explains a significant portion of the variance in `log(price)`. All predictors turn out to be significant. As the EDA suggests a potential interaction between `rooms` and `square_meters`, I build a second version of the model using this interaction. Comparing the two versions of the model with an ANOVA test suggests that adding the interaction to the model is helpful in explaining more variance in the dependent variable. The coefficients for this model are shown in *Appendix B.* - ANOVA Table? - Model assessment

Hierarchical Linear Model

Since the price is expected to vary heavily by geographical area, it makes sense to include the location of a listing in the model. The location for this data set can only be determined on a zip code level, however, this should suffice to tell apart cheap from expensive neighborhoods as zip code ares in Berlin are rather small. Thus, I construct a the same linear model as described above (including the significant interaction term), but this time add the zip code as a level of random intercept. This way, each zip code area is allowed to have a different "baseline price" from which apartment features determine the final price.

Conclusion

Appendix

(A) EDA plots

Distribution of `rooms`

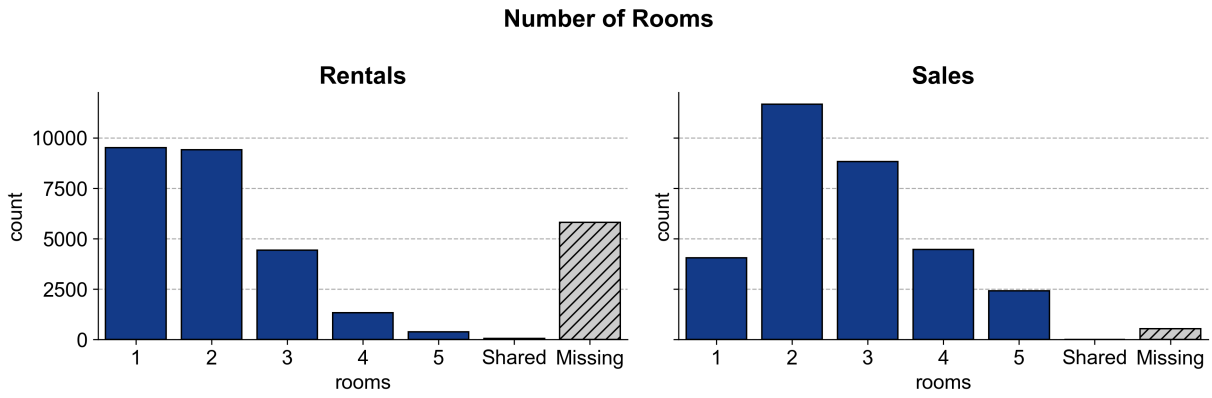


Figure 1: Distribution of `rooms`

(B) Coefficients for non-hierarchical linear model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.0485	0.0089	677.08	0.0000
object_typeHOUSE	0.9385	0.0531	17.69	0.0000
object_typeSHARED_APARTMENT	0.2049	0.0078	26.12	0.0000
object_typeTEMPORARY_LIVING	0.8298	0.0055	151.76	0.0000
private_offerTRUE	-0.1425	0.0068	-20.81	0.0000
rooms2	-0.0897	0.0191	-4.71	0.0000
rooms3	-0.1282	0.0231	-5.54	0.0000
rooms4	0.1626	0.0393	4.13	0.0000
rooms5	0.2081	0.0729	2.85	0.0043
roomsShared	-0.4476	0.0818	-5.47	0.0000
roomsMissing	-0.0890	0.0133	-6.68	0.0000
square_meters	0.0040	0.0002	22.25	0.0000
rooms2:square_meters	0.0060	0.0003	18.14	0.0000
rooms3:square_meters	0.0069	0.0003	22.72	0.0000
rooms4:square_meters	0.0046	0.0004	12.17	0.0000
rooms5:square_meters	0.0043	0.0005	8.21	0.0000
roomsShared:square_meters	-0.0030	0.0018	-1.71	0.0874
roomsMissing:square_meters	-0.0041	0.0002	-19.51	0.0000