# 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

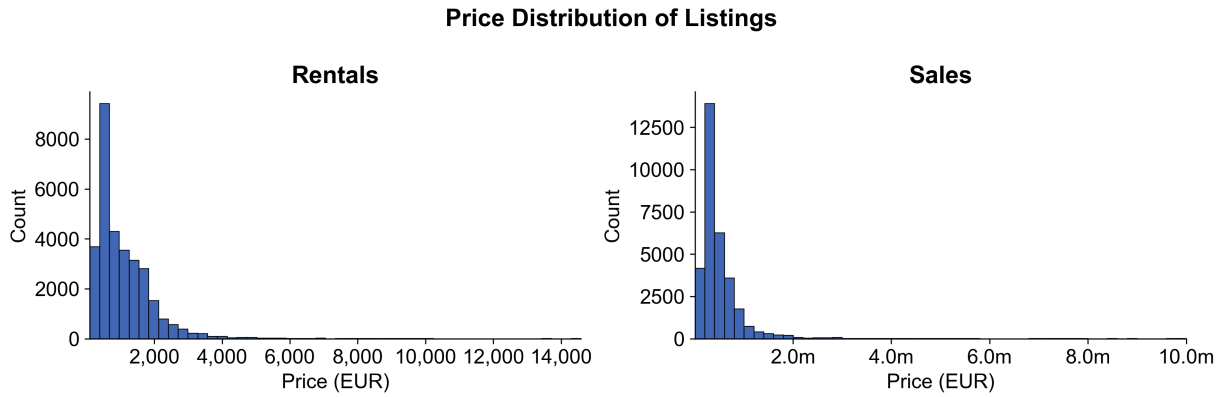
This report aims to identify the characteristics of an apartment or house that are associated with higher and lower prices to shed light on which attributes buyers and renters might look for when searching affordable housing in Berlin. The data used in this study has been collected through web scraping 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.
- todo?

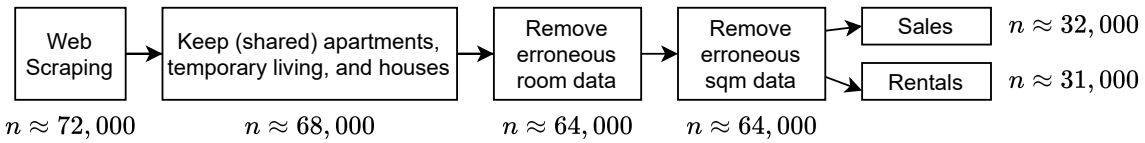



## Data Cleaning

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 is 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. This final hierarchical model will be interpreted to answer the research question of which object characteristics drive price and how one can potentially find affordable housing (not just necessarily in terms of absolute price but also in terms of what you get for your money).

## 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 $R^2 \approx 0.65$ for rentals ($R^2 \approx 0.54$ for sales) after removing three (12 for sales) high-leverage points. Thus, it already explains a significant portion of the variance in `log(price)`. All predictors turn out to be significant for modeling both rental and sales prices. 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 (for both rentals and sales respectively) suggests that adding the interaction to the model is helpful in explaining more variance in the dependent variables as the models with interactions have an $R^2 \approx 0.68$ and $R^2 \approx 0.62$ respectively. We can see that the model is able to explain slightly more variance in rental prices compared to sales prices and that including the interaction - while valuable for both models - helps more for modeling sales prices than it does for rent. This might indicate that forecasting sales prices of properties is a harder task than forecasting rental prices. The coefficients for both 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 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.

- VIF

The random intercepts by zip code are can also be visualized on a geographic map (grey indicates non-significant random intercepts). The model has clearly learned that areas close to the city center (Mitte) are more expensive than the outskirts of Berlin (with some exceptions) and that cheaper rentals can generally be found in the north-east and north-west parts of Berlin.

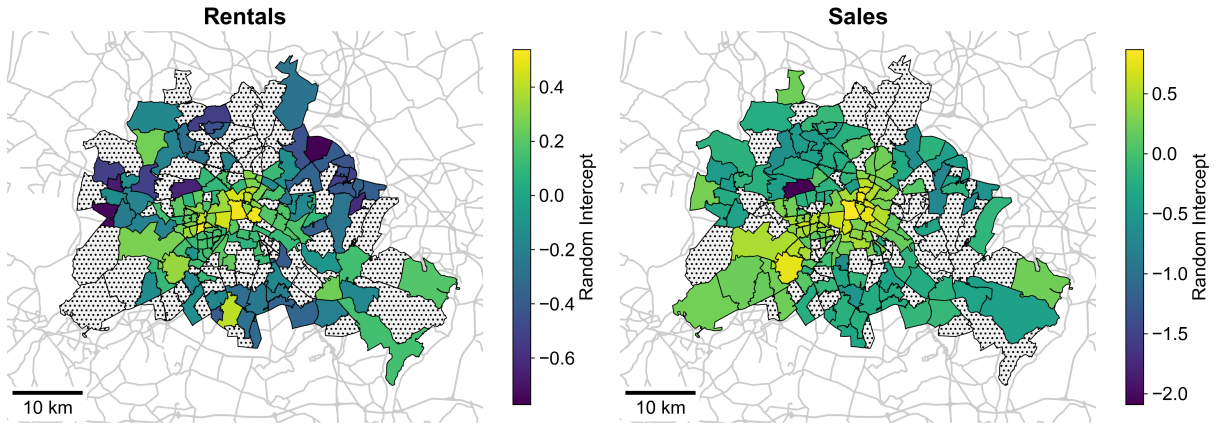
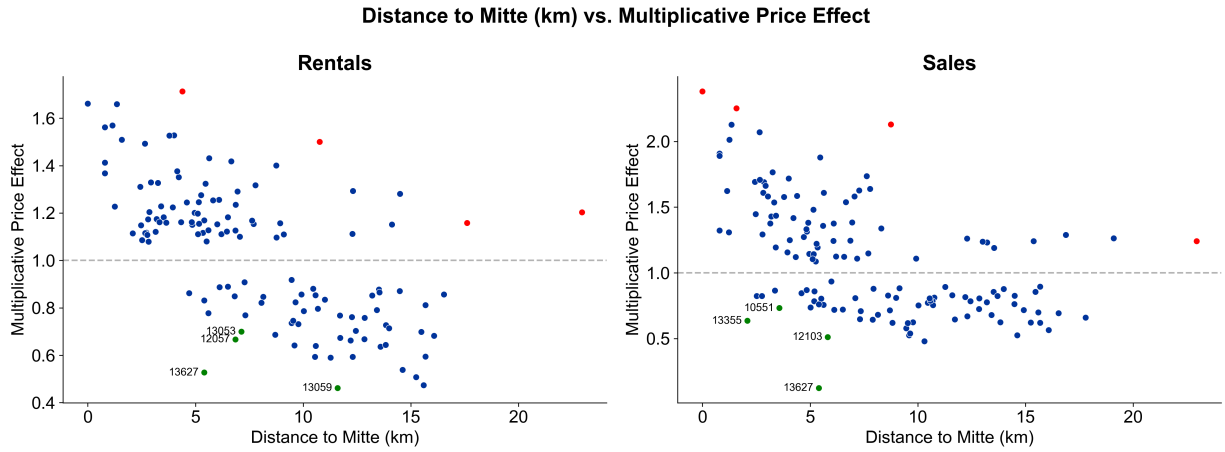



Figure 1: Map of Random Intercepts by zip code

To further analyse how the distance from the city center affects the random intercept (and therefore the predicted price of the listing), the following plot graphs the geographical distance of each zip code are to the city center against its random intercept. Generally speaking, we see a negative linear relationship, where zip code areas that are far from the city center tend to be less expensive. However, we can also see that some zip codes are especially cheap (or expensive) given their distance from the center. Highlighted in red (expensive) and green (cheap) are the points with the top 2.5% largest absolute difference from a best-fit linear model in either direction (where the linear model regresses the multiplicative price effect on the distance to the city center).

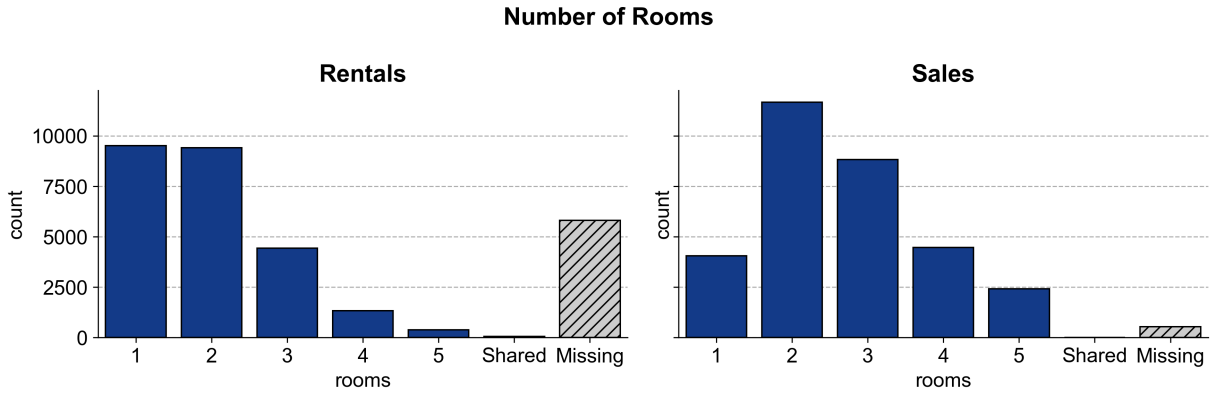

Distance to Mitte (km) vs. Multiplicative Price Effect
Rentals
Sales
Multiplicative Price Effect
Distance to Mitte (km)
13053
12057
13627
13059
10551
13355
12103
13627

# Conclusion

# Appendix

## (A) EDA plots

### Distribution of `rooms`




## (B) Coefficients for non-hierarchical linear model

### Rentals

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 6.0398 | 0.0089 | 679.06 | 0.0000 |
| object_typeHOUSE | 0.9115 | 0.0527 | 17.31 | 0.0000 |
| object_typeSHARED_APARTMENT | 0.1927 | 0.0079 | 24.39 | 0.0000 |
| object_typeTEMPORARY_LIVING | 0.8269 | 0.0055 | 149.86 | 0.0000 |
| private_offerTRUE | -0.1366 | 0.0069 | -19.77 | 0.0000 |
| rooms2 | -0.0737 | 0.0189 | -3.90 | 0.0001 |
| rooms3 | -0.1060 | 0.0230 | -4.61 | 0.0000 |
| rooms4 | 0.1822 | 0.0385 | 4.73 | 0.0000 |
| rooms5 | 0.1481 | 0.0648 | 2.29 | 0.0222 |
| roomsShared | -0.3439 | 0.0820 | -4.19 | 0.0000 |
| roomsMissing | -0.0868 | 0.0132 | -6.56 | 0.0000 |
| square_meters | 0.0044 | 0.0002 | 24.36 | 0.0000 |
| rooms2:square_meters | 0.0056 | 0.0003 | 17.07 | 0.0000 |
| rooms3:square_meters | 0.0065 | 0.0003 | 21.46 | 0.0000 |
| rooms4:square_meters | 0.0042 | 0.0004 | 11.40 | 0.0000 |
| rooms5:square_meters | 0.0044 | 0.0005 | 9.78 | 0.0000 |
| roomsShared:square_meters | -0.0041 | 0.0018 | -2.32 | 0.0202 |
| roomsMissing:square_meters | -0.0042 | 0.0002 | -20.14 | 0.0000 |

### Sales

| | Estimate | Std. Error | t value | Pr(>\|t\|) |
|---|---|---|---|---|
| (Intercept) | 12.0372 | 0.0092 | 1313.74 | 0.0000 |
| object_typeHOUSE | -0.2369 | 0.0092 | -25.65 | 0.0000 |
| private_offerTRUE | -0.1473 | 0.0099 | -14.90 | 0.0000 |
| rooms2 | -0.2348 | 0.0178 | -13.18 | 0.0000 |
| rooms3 | -0.2147 | 0.0187 | -11.50 | 0.0000 |
| rooms4 | 0.3492 | 0.0241 | 14.50 | 0.0000 |
| rooms5 | 0.7571 | 0.0320 | 23.67 | 0.0000 |
| roomsMissing | 0.6933 | 0.0250 | 27.73 | 0.0000 |
| square_meters | 0.0043 | 0.0002 | 27.23 | 0.0000 |
| rooms2:square_meters | 0.0098 | 0.0003 | 33.75 | 0.0000 |
| rooms3:square_meters | 0.0098 | 0.0002 | 41.14 | 0.0000 |
| rooms4:square_meters | 0.0049 | 0.0002 | 20.76 | 0.0000 |
| rooms5:square_meters | 0.0019 | 0.0002 | 7.93 | 0.0000 |
| roomsMissing:square_meters | -0.0018 | 0.0002 | -10.75 | 0.0000 |