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.

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, I keep properties with 5 or less rooms and merge the text-based entries into the correct categories (e.g. "single room" => 1, "shared room" => shared, "not given" => 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. 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. The entire data cleaning process and its effect on the sample size n is displayed in the flowchart below.

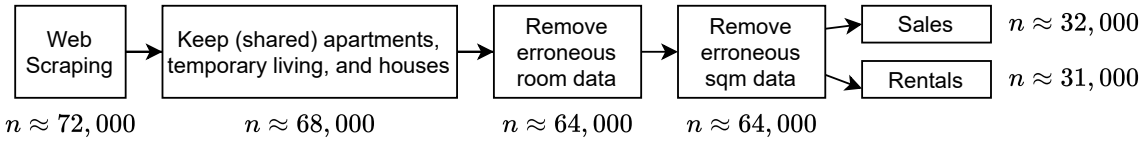

Figure 1: Flowchart of data preprocessing

Model

Conclusion