Final Project Proposal

Moritz Wilksch

Overview

In this final project, I plan to analyze the real estate market in Berlin, Germany. Specifically, I want to study which real estate attributes influence the selling or rental price of residential properties using data set that I scraped from www.immobilienscout24.de, www.immowelt.de, and www.ebay-kleinanzeigen.de, the three largest online marketplaces for properties in Germany (equivalent to www.zillow.com and craigslist.org in the US).

Research Questions

Real estate prices in major cities have been rising for the past years. This project aims to study what attributes are associated with high apartment prices. Moreover, I want to characterize the nature of these relationships to gain a deeper understanding of how real estate owners price their properties. Ideally, this leads to a comprehensive analysis that could inform searches for affordable housing. To my knowledge, there is little previous research based on actual real estate listings, as most research reports on real estate markets are rather high-level and focus on descriptive statistics for different boroughs.

Data

The entire data set contains around 111,000 real estate listing all over Germany. As this data is quite heterogenous, I want to focus the analysis on all properties in Berlin ($n \approx 8{,}600$). Overall, there are approximately 30 different variables for each property, but several of them hardly contain any values as not all platforms and listings publish all information. As the data has been web scraped (and is quite "dirty"), the data cleaning will take some significant effort before modeling. Moreover, it contains free text data (the listings title and the realtors' description of the property), but it will be hard to use these features for this modeling task (maybe the presence of some keywords could be used to model the price). Thus, the analysis will focus on the following variables:

```
location, object_type, price, private_offer, rooms, square_meters, to_rent,
zip_code
```

A sample of the data set is available here:
https://gist.github.com/moritzwilksch/20b2727fe1f47caa244fafb13242b90a

Project Plan

Mainly, I will try to use a linear regression model to regress price on property attributes. It might be wise to build two models, one for rental properties and one for properties for sale. This enables me to answer inferential questions and visualize interesting relationships (e.g., plot the influence of the zip code on price on a map). Should the linear regressions predictive performance be inadequate, I might consider using non-linear but interpretable models like decision trees and compare them against the linear regression.

Date	Milestone
November 8th, 2021	Data cleaning
November 22nd, 2021	Linear modeling
December 6th, 2021	Model interpretation, visualization + additional modeling if necessary
December 12th, 2021	Submit deliverables