General data cleaning

Much of the data available required minor cleanup – stripping out dollar signs from numerical data, converting simple data such as profile picture to a binary “hasProfilePic,” etc. Null values were handled on a case-by-case basis. For some features, such as “cleaning fee,” a null value has been assumed to be a 0 and replaced by such. For others, such as “review scores rating,” a null value has been assumed to be null for lack of information, and has been replaced by the mean value of that feature over all collected data. For instance, the mean for one dataset’s “review scores rating” is 95.074. For observations where this rating is missing, the null value has been replaced by 95.074, assuming a neutral response. Treating this as a baseline, then, our model assumes that a rating above 95.074 (for this dataset) to be significant in one direction, while below this to be significant in the other direction.

At this time, values have not been normalized directly, but doing so (and replacing nulls with 0 in that case), may aid in improving the model.

Several data features, such as host response rate, are categorical in nature, with Airbnb providing a dictionary of possible outcomes. These are retained as categorical features, but analysis and experimentation may lead to a quantitative approach where such data has a clear ordinal nature.

Lastly, the “amenities” data has been treated in two separate manners, with the two approaches yielding similar results as of yet. Amenities are also provided as a dictionary of possible inclusions, the final result containing any combination of 0 to all amenities.

In one method, the data cleaning process is used to create one binary feature for every amenity. The other method simply condenses all amenities into a single quantitative variable which represents the total number of amenities listed. As both these approaches have thus far yielded similar results, the latter is preferred for its ability to retain accuracy while helping reduce the total number of features the model requires. Additionally, it allows for more potential scaling of the model approach, as regions which offer different amenities (or which would require language translation) may be directly compared.

Handling of locational data

Our intuition is that the AirBnB rental’s location plays a role in determining its optimal price, however working off of defined neighborhoods presents possible bias, as well as limits the modeling approach from future scaling.

Rental longitude and latitude data are available, and when treated as separate quantitative variables performed well for individual cities. It again introduces a problem of scaling of the modeling approach. When applied to smaller regions, the lack of variability in lat/long reduces their effectiveness as price predictors, while when applied to larger regions, their dependence becomes a greater issue.

In order to lessen this impact, we have devised a k-means approach to transform the quantitative lat/long variables into a single, k-level categorical variable. In testing, a K of 8 appears adequate to match the performance of using lat/long directly for a large city-sized region, and this same approach is likely to scale well both to a neighborhood level (single Chicago suburb for instance, given adequate observations) or upward to state, country, or global models.

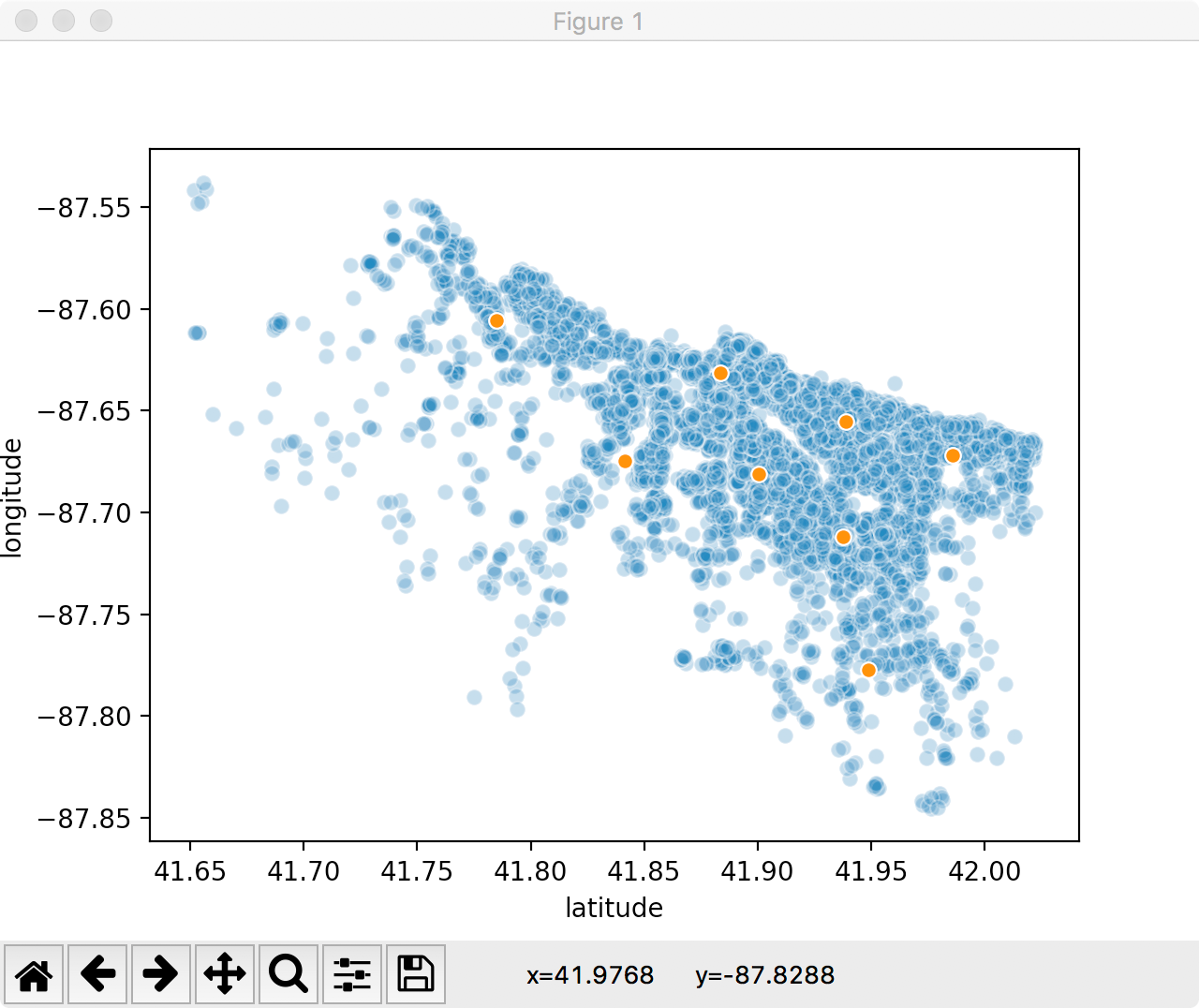


Fig: The 8 primary neighborhoods of AirBnB rentals in Chicago, as predicted by a k-means approach, to be used as a categorical predictor.

Linear-based Models

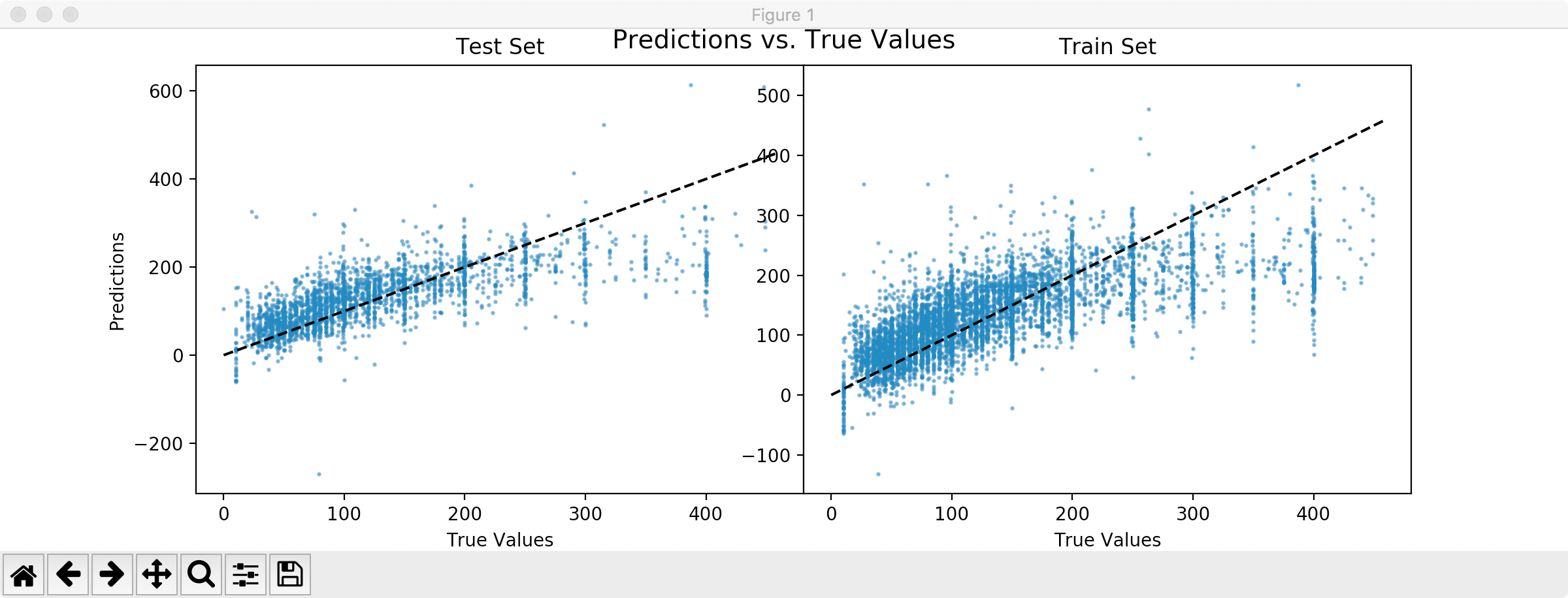


Fig: Predicted prices vs. actual prices on training and validation data for AirBnB rentals in Berlin, with an Y/X=1 ab1 line.

Plotting predicted/true values, there are two things of note: There is a rather substantial decrease in the number of available observations as price increases (also see figure below), and as price increases, our model tends to undervalue rentals (evident from the imbalance of points above the ab1 line on the lower price range, and majority of points falling below the ab1 line in the higher price range).

There is the possibility of features not found by the model which could account for these discrepancies, based on reviews or similar linguistic data not currently modeled, however it is also likely that a non-linear model and/or a more localized model will more accurately capture the shifting trend across all price ranges.

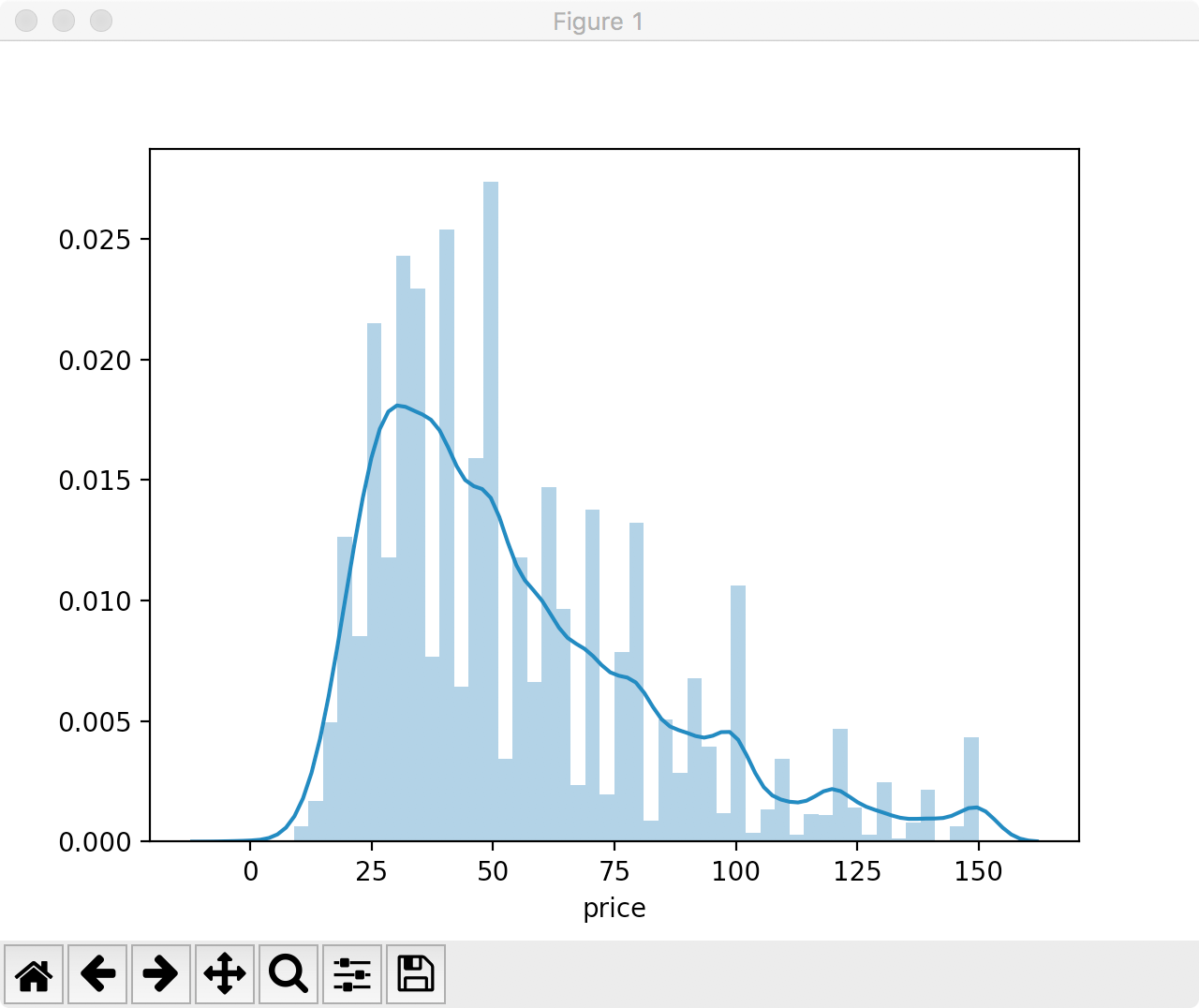


Fig: Price distribution of AirBnB rentals in Berlin. Available data quickly tapers off for higher price ranges. (This data has been trimmed to π0.95, removing outliers for illustrative purposes).