Predicting Airbnb Listing Price | Statistical Analysis
Claire Miles

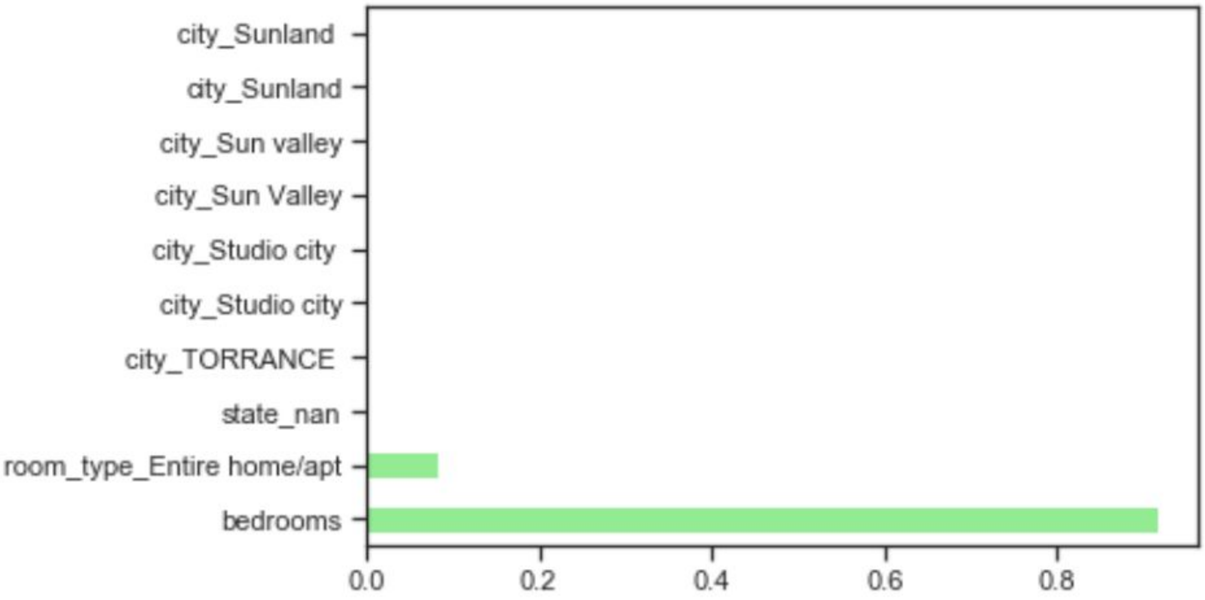
After cleaning the data and starting to craft the beginning of my data story, it was time to enrich my exploration with inferential statistics. My exploratory data analysis included the following steps:

1. Run a quick random forests analysis of a subset of the data to get a sense for which features stand out as important in the analysis
2. Look for correlations in the most important features, including the target feature.
3. Run confidence intervals and significance tests on variables that are related to the project question.
4. Re-evaluate the original project question based on statistical evidence.

Once again, I focused on the listings dataset in this notebook, as the listing data hosts all numerical features that the dataset currently has. The reviews data, which contains the review text for different listings, will be interpreted using NLP later on in the project process, as will other text data in the listings file.

Random Forests:
I briefly prepared the data for the random forests training by converting certain attributes of the dataset to categorical data type, turning booleans into 0/1 values, and selecting only the numeric, datetime, and categorical datatypes in the dataframe. To get rid of NaN values, I used pandas interpolate function to fill in the values linearly (I'll give more thought to interpolation in the training of the final model, but here I'm just looking for a crude first pass). Then, I trained a RandomForestRegressor classifier on one-third of the data and plotted the features by their importances.

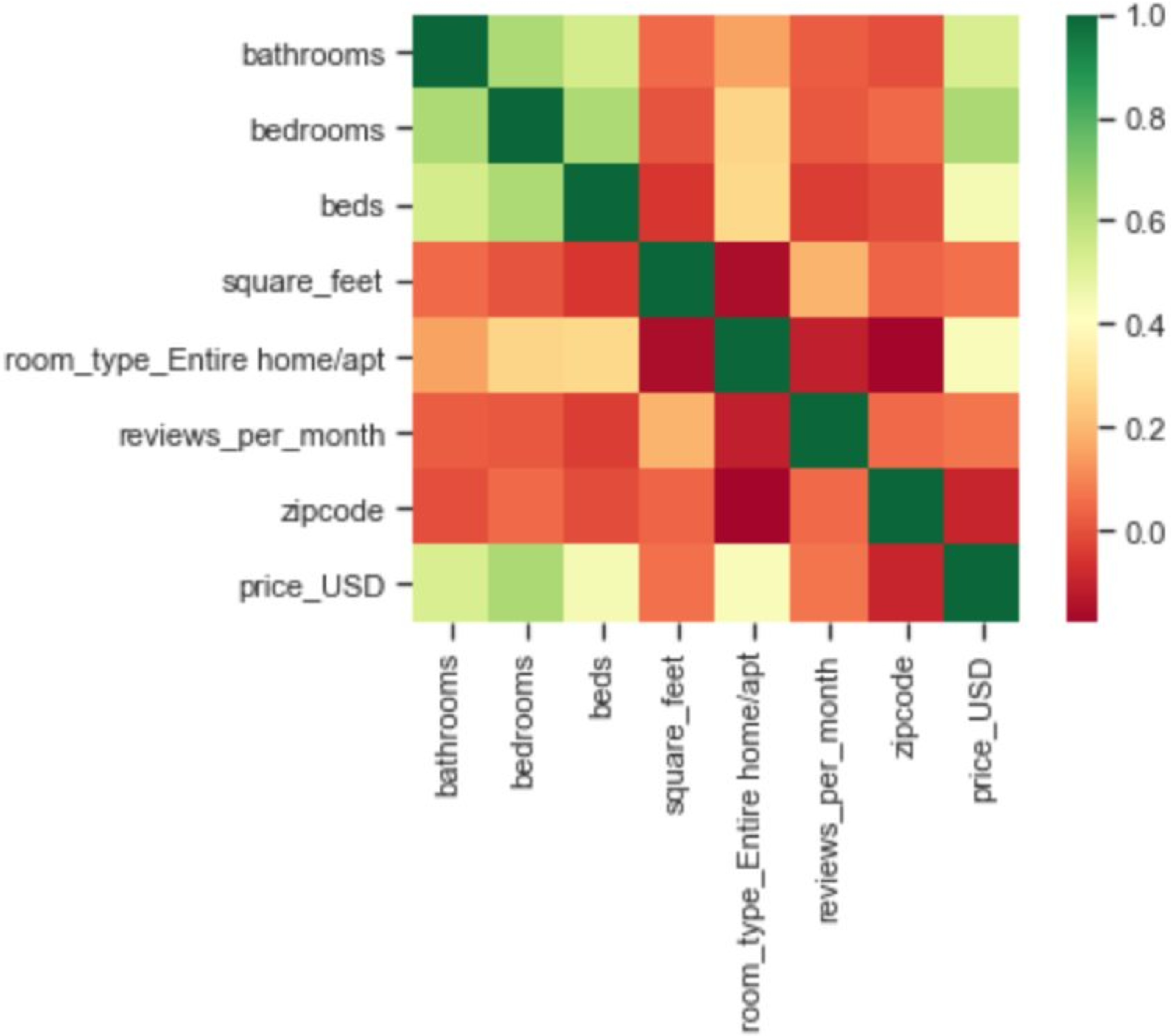
Random Forest Feature Importances



The most important features to our question seem to be bedrooms and if the room type is an entire home/apartment. It's important to note, however, that this was a very simplistic first-pass at the data, and that the feature importance algorithm for random forests shows some bias towards continuous features or high-cardinality categorical variables. Therefore, these results should be taken with a grain of salt.

Correlation-finding:
Using the feature importance information from the random forest algorithm, I plotted these features, as well as some other I've previously identified as important, on a heatmap along with price. The heatmap shows the correlations between the features.

Heatmap of Select Features



From looking at the 'price_USD' row of the heatmap, it looks like 'bathrooms', 'bedrooms', 'beds', and 'room_type_Entire home/apt' all have some sort of correlation with price.

Statistical Testing:
Another way to check for correlation is by running hypothesis tests of correlation between a particular feature and price. By permuting one of the features and computing the Pearson

correlation coefficient, we can see if the correlation we see between features is most likely by chance, or statistically significantly correlated. Therefore, I performed hypothesis tests for correlation for the three features that showed correlation with price: bathrooms, bedrooms, and room type (Entire home/apartment). The null hypothesis for these tests was that the two features are NOT correlated, with the alternative hypothesis being the opposite. All three tests concluded to reject the null hypothesis, giving us further evidence that bathrooms, bedrooms, and room type are correlated with price.

Re-evaluation of project question:

The biggest finding of my statistical analysis was that it appears that the 'bedrooms' and 'room_type_entire Home/apt' features will most likely be important to the our machine learning model. Additionally, we had statistically significant confirmation that many of the features we thought were correlated indeed are.

Moving forward, the machine learning analysis will likely offer more complex insights than this statistical analysis, especially as I incorporate information from the reviews data. It may be the case that some features that are not significant in this analysis play a larger role down the line.

This analysis does not alter the initial project question of trying to decipher a listing's price from its other features.