Alyssa Hall
Beck Millet
Brian Maxwell
Sandy Yang

1 Data description

1. Load the data airbnb-seattle-listings-train.csv. Broadly describe the variables you see, their encoding, and discuss if these may be valuable in determining the price. For instance, you may want to tell that house_rules is text, and you may want to check if smoking allowed/not allowed is related to the price.

Looking at the initial data, there are 106 columns of variables most of them being irrelevant to price such as “host_country_code”, “host_id”, “listing_url”, etc. Some that initially stood out as potentially useful were ‘neighborhood’, ‘property type’ and ‘zipcode’, which later turned out to be less than useful due their distributions. A lot of the text fields were irreducibly complex (e.g. “host rules” or “notes”) as strings of text with far too much variability, so we just eliminated those columns.

2. Consider how will you handle missing data. For instance, 95% of the square feet observations are missing, 17% of security deposit observations are missing. You lose too many observations if you just ignore those.

After narrowing the columns, we deleted the three columns with 90% or more missing. For remaining values containing NA values, they were dropped when running the OLS model.

3.Consider which variables you are going to use below. For all of these, create a summary table that contains relevant summary information. In particular pay attention to the missing values. Note that missings may not just be coded as such, they may also be empty strings and values like N/A. You may return to this point repeatedly as you develop your model.

For each of our chosen variables, we first ran value_counts to determine whether or not they contain a reasonable amount of data and their distribution. Further, we removed the variables that had homogenous or grossly skewed distributions. With the remaining 25 or so variables we ran OLS just to see how well each variable predicted price on it’s own.

2 Model

1. Either split your data into training and validation sets, or just use cross validation below.
2. Develop the models. Report all the variables and how do you clean/encode those. While the exact details are visible in the code, explain the broad choices in text.

Cleaning/encoding the data set required changing the variable types of certain columns. Only a few remaining columns were non numeric, like “room_type” and “cancellation_policy” which were both ordinal in nature and therefore reduced easily to numbers. The few boolean values like “host_is_superhost” and “instant_bookable” were easily converted into integer and some of the ratio data(“price” for example) needed to be regex-ed to convert it into integers.

3. Report the final number of observations, the estimated coecient val2 ues, adjusted R , and RMSE set (or k-fold CV) for three models:

(a) a simple one that only contains a few most important variables/best predictors. What do you think are 2-3 best predictors in the data?

“Host_total_listings_count”, “accommodates”, “reviews_per_month”
We chose these variables after trying different variables in the multivariable linear regression model and chose the ones with the largest contributions to the adjusted r squared values. See code***

(b) the full model: everything you consider useful.
See code titled Analysis

(c) something in between.
See code titled Analysis

But please do not report all of the coefficients of the large models, only a small subset (10 or so) of the most important/interesting ones.

4. Interpret the coefficients of the reported models. Again, only interpret the most interesting/important ones, not all of those! Do the coecient values differ between the models? Can you explain why?

If we look at “host_is_superhost”, it turns out that this variable has a negative correlation with price. Which at first glance seems odd. Upon further reflection, this is only odd if we associate quality of host with a higher price. Which is naive and reductionist approach to the complexities of market economics.

Looking at the “longitude” variable, there were some interesting findings in that the adjusted r squared value was 0.003 and it also had a large coefficient. This means that longitude isn’t a good indicator of price based on the model, but the large coefficient indicates that there may be some outliers.

5. Use your models to predict the price. Report RMSE in the table above.

After running all models 500 times we took the mean RMSE of the results:

Arbitrary constant:	195
3 variable model :	127
Comprehensive model:	123
5 variable model:	124

Note: ensure you t the models on your training set and compute RMSE on the validation set. Here we care about overtting. It is less important when you just interpret coefficients above.

3 Think

1. Does your model do a good job in predicting the price?

Relatively, the model does a good job at predicting price when focusing on specific variables that have already been chosen as strong indicators. The range is 1 to 5000, so an RMSE a little over a hundred seems like a good start. However, the arbitrary constant benchmarked at just shy of 200. So our model is better, but whether it is “good” is a different question.

2. can your results be used for something interesting, say for research or commercial purposes? What might it be?

The results could be used either commercially or for research. Specifically looking at commercial uses, investors may be interested in the types of properties/attributes of

successful properties (the more expensive air bnb properties) when purchasing new real estate.

3. you were predicting the price. Did you include any other price-related variables, such as weekly price or security deposit in your model? What would that mean in terms of the model usability?

We did include the security deposit variable in our larger multivariable model, and did not include it in our smaller 3 variable model. The fact that security deposit, is part of the model to predict price may mean that the model is tautological and therefore flawed. However, this is contingent upon what the variable price actually means and whether the security deposit or cleaning fee are included as part of the price. Without documentation we can't know. However, if each dollar of security deposit/cleaning fee caused an increase in a dollar of the subsequent price variable(which would happen if price included those values) then the linear regression model of price ~ security deposit would look very different than it does. This suggests the model is fine including other pricing values.

4. imagine you are developing this work for a local, or for the national government. Why may government be interested in such a job? Do you see any ethical issues that may arise from your work?

This type of information could be used by state or local government to calculate tax rates for certain properties. They may be able to use the model to understand the types of properties and their prices to estimate different percentages/groups of taxes. As for the ethical issues that may arise, it seems like this could be a more ethical and equitable way to tax air bnb properties rather than a flat tourism tax or property tax, and rather focus on the actual expected revenue of the property.

4 Additional task

1. load the testing data arbnb-seattle-listings-test.csv. This has exactly the same structure and variables as the original dataset.

2. compute RMSE on the testing dataset. This is the ultimate goodness measure of your model. Present it prominently in your report.

After running all models with **arbnb-seattle-listings-test.csv** 500 times we took the mean RMSE of the results:

Arbitrary constant:	165
3 variable model :	110
Comprehensive model:	104
5 variable model:	112

3. Do not tinker with the model any more. This was your final test.

Note: you may still have to x coding errors if there is something wrong so you cannot compute RMSE on the test data.