1. What kind of cleaning steps did you perform?

   The first thing was to inspect the data with .head(), and .info()

   There were 3818 rows and 92 columns. There were too many features, that needs a feature selection.

   There was sign of something missing values, as there are not all non-null values.

2. How did you deal with missing values, if any?

   The columns that had a lot of missing values are license and square feet. This meant two things: Seattle didn’t require a license for airbnb, and that people didn’t know their square footage. That meant if airbnb could offer a way to measure the square footage, it would be a great feature.

   My original plan was to use mean to fill these missing values, however, after rethinking, this might creates some outliers in the pricing predictions.

   I decided to leave the missing values alone, since it didn’t affect much.

3. Were there outliers, and how did you handle them?

   In order to see if there were any outliers, I proceeded with visualizing single variables. Here I chose: ‘price’, ‘bedrooms’, and ‘availability_30’

   I didn’t want to chose scaling variables here, since there are scaling, there are no outliers to test.

   Since price wasn’t in numerical but object, I first had to change that into floats.

   Along with price, I also converted percentage to floats and date to datetime.

   I couldn’t find any outliers in for price vs bedrooms in my swarmplot.

for ‘availability_30’ the frequency of 30 is high because that means the listing is pretty new or empty.