1. What kind of cleaning steps did you perform?

   The first thing was to inspect the data with .head(), .tail(), .shape, .columns,

   There were 3818 rows and 92 columns. There were too many features, might need to choose some features for a data subset.

   There was sign of something missing values, so I used .info() to see the non-null values.

   I recognized that the non-null was not all 3818,, that meant there were missing values.

   I planned to delete columns with little non-null values.

   Next, I used .describe() to calculate summary stats and since it could only be used on numeric values, we then had to use value_counts for categorical value.

   We had missing values, I set dropna=False. I chose ‘neighborhood’, ‘property type’ and ‘cancellation policy’

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.

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 numerica but object, I first had to change that.

I couldn’t find any outliers in ‘bedrooms’ since it is only from 0-5, for ‘availability_30’ I would consider 30 to be an outlier because that means the listing is pretty new.