1 Part 1: Contextualizing the Data

Let's try to understand the background of our dataset before diving into a full-scale analysis.

1.1 Question 1

1.1.1 Part 1

Based on the columns present in this data set and the values that they take, what do you think each row represents? That is, what is the granularity of this data set?

Each row of the dataset represents a specific building, with its information conteined in the columns.

1.1.2 Part 2

Why do you think this data was collected? For what purposes? By whom?

This question calls for your speculation and is looking for thoughtfulness, not correctness.

This data might be collected by the real estate agencies, in order to keep track of local housing condition and therefore infer useful business information.

1.1.3 Part 3

Certain variables in this data set contain information that either directly contains demographic information (data on people) or could when linked to other data sets. Identify at least one demographic-related variable and explain the nature of the demographic data it embeds.

The "Use" column contains demographic information, since it states whether the building was used by a single family or multi families. This variable shows the crowdedness of dwelling in a certain building, and could be used for demographic purpose.

1.1.4 Part 4

Craft at least two questions about housing in Cook County that can be answered with this data set and provide the type of analytical tool you would use to answer it (e.g. "I would create a _____ plot of _____ and ***" or "I would calculate the*** [summary statistic] for _____ and _____"). Be sure to reference the columns that you would use and any additional data sets you would need to answer that question.

1. I would create a line plot of "Sale Price" and "Wall Material" to see if there if a relationship between the building material and the valuation of a building.
2. I calculate the average "Sale Prices" based on whether the building has a "Central Air" system, in order to determine customer's preference of housing equipment.

1.2 Question 2

1.2.1 Part 1

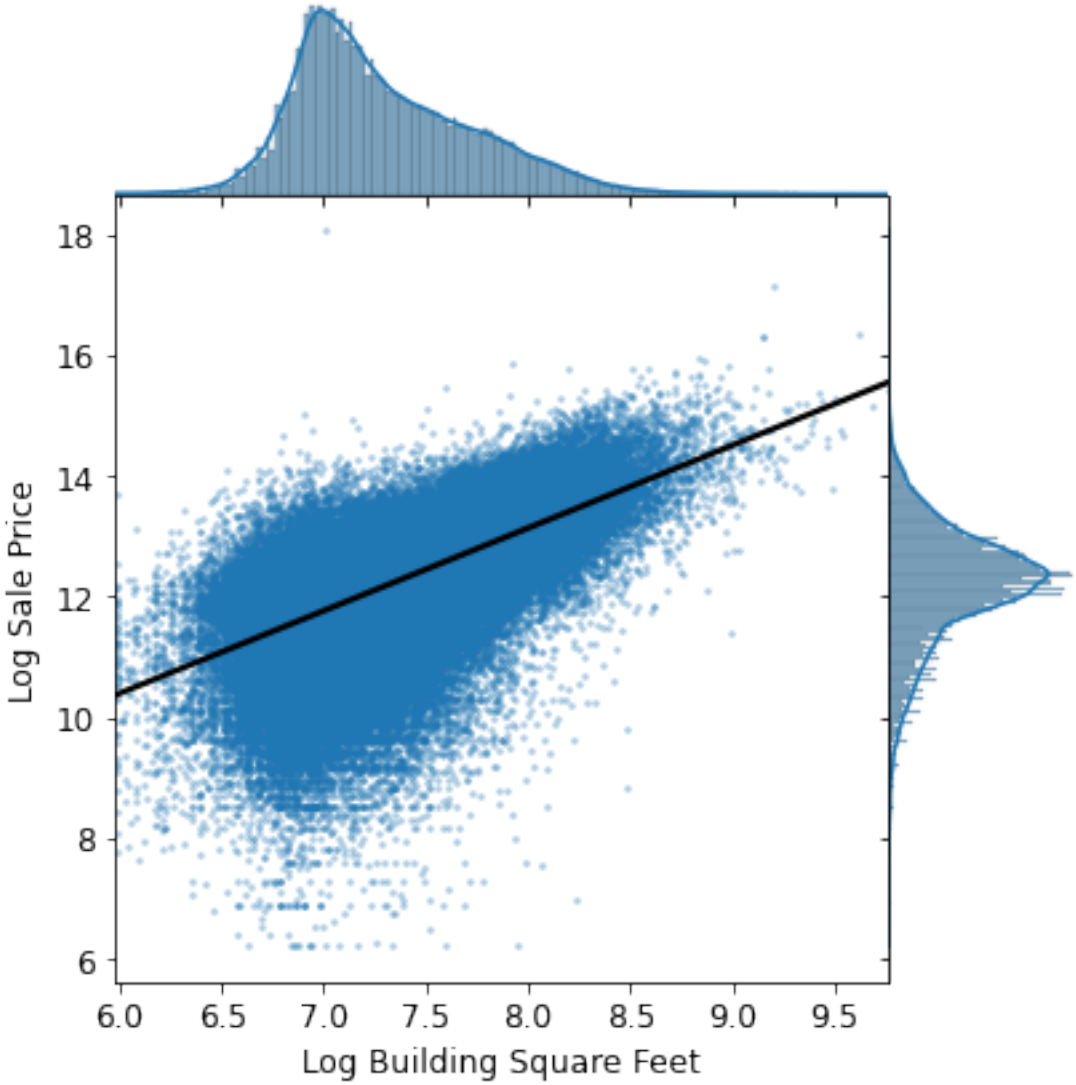
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards as the autograder may not work otherwise.

One problem of the plot was that the scale of Sale Price was too large so that the distribution could not be presented clearly. One way to solve this was to adjust the scale to smaller unit based on the descriptive statistics of "Sale Price".

1.2.2 Part 3

As shown below, we created a joint plot with `Log Building Square Feet` on the x-axis, and `Log Sale Price` on the y-axis. In addition, we fit a simple linear regression line through the bivariate scatter plot in the middle.

Based on the following plot, does there exist a correlation between `Log Sale Price` and `Log Building Square Feet`? Would `Log Building Square Feet` make a good candidate as one of the features for our model?



Based on the plot, there is a correlation between the Log Building Square Feet and Log Sale Price, since

in general these two variables have a positive correlation. Thus, Log Building Square Feet makes a good candidate as one of the features for the model.

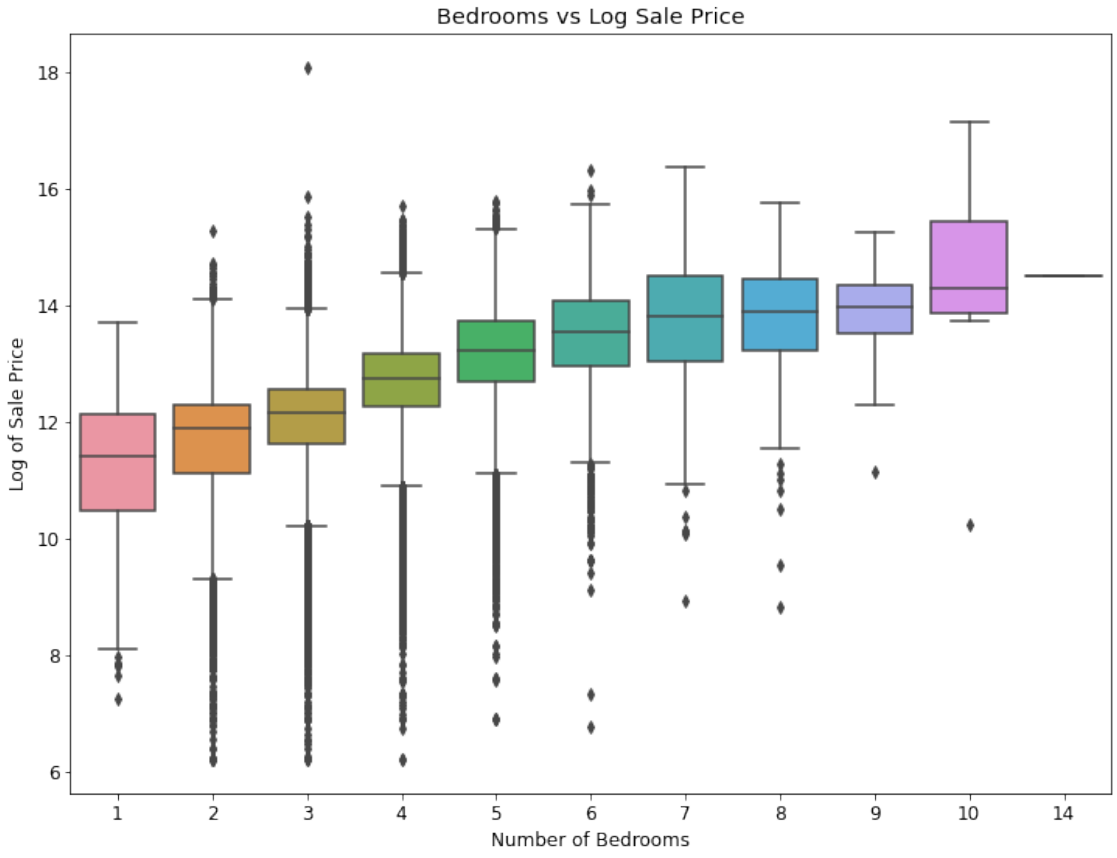
1.2.3 Part 3

Create a visualization that clearly and succintly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and succinct title. - It should convey the strength of the correlation between the sale price and the number of rooms.

Hint: A direct scatter plot of the sale price against the number of rooms for all of the households in our training data might risk overplotting.

```
In [146]: sns.boxplot(training_data['Bedrooms'], training_data['Log Sale Price'])
          plt.title('Bedrooms vs Log Sale Price')
          plt.xlabel('Number of Bedrooms')
          plt.ylabel('Log of Sale Price')
```

```
Out[146]: Text(0, 0.5, 'Log of Sale Price')
```



1.2.4 Part 3

It looks a lot better now than before, right? Based on the plot above, what can be said about the relationship between the houses' `Log Sale Price` and their neighborhoods?

Based on the plot above, Log Sale Price does not have a clear correlation among different Neighborhoods, since the average of Log Sale Price were generally around the same horizontal line.