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 represents a household, such as home, apartment, condo, etc., and its descriptions, such as features and details.

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 could have been collected by an authority such as the city, or a real estate business in order to get details of hosueholds for business or policy-making use. People that would be interested in this dataset would be people looking to buy/sell a home in Cook County, real estate firms, and local government. The information in the data could be used to determine remodeling or rebuilding of certain houses if the city or real estate business decides it is appropriate.

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.

Variables such as 'Census Tract', 'Most Recent Sale' and 'Sale Price' are demographic-related variables. Variables such as 'Census Tract' embeds information that could relate to issues like redlining and segregation, and variables such as 'Most Recent Sale' and 'Sale Price' could be related to make discoveries such as average price of homes around certain areas in order to find any patterns that relate to the social classes of the demographics.

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.

Question: Are there any correlations between the wall material and the cost of the house? I would answer this question by creating a bar plot of houses for each wall materials and seeing their average price. I will need to calculate the average price of homes but will have to consider any outliers that might skew the data.

Question: Do the costs of houses differ for different locations? I would answer this question by solving for the average price of houses for different locations and creationg a scatter plot to see how much deviation individual houses have from the average to see if there are any patterns in costs of houses for locations.

1.2 Question 2

1.2.1 Part 1

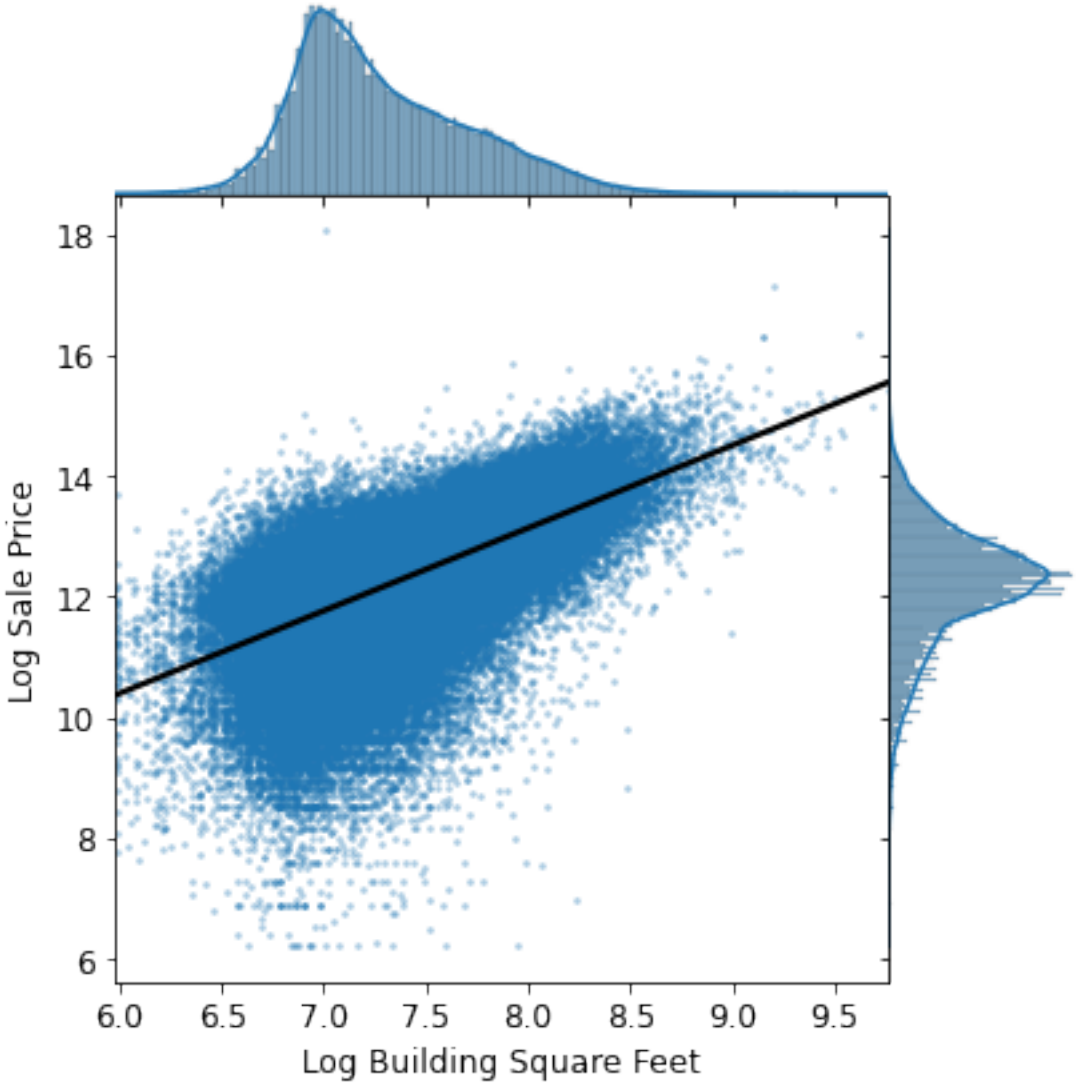
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The big issue with the visualization above is the scale in which the x-axis (Sale Price) was set. It is impossible to read the data because the x-axis is scaled in 1e7. One way to solve this issue would be to set change the scale of the x-axis to a smaller value, where the data would be spreadout throughout the whole graph, which would help us visualize the data better.

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 following plot, there exists a positive correlation between Log Sale Price and Log Building

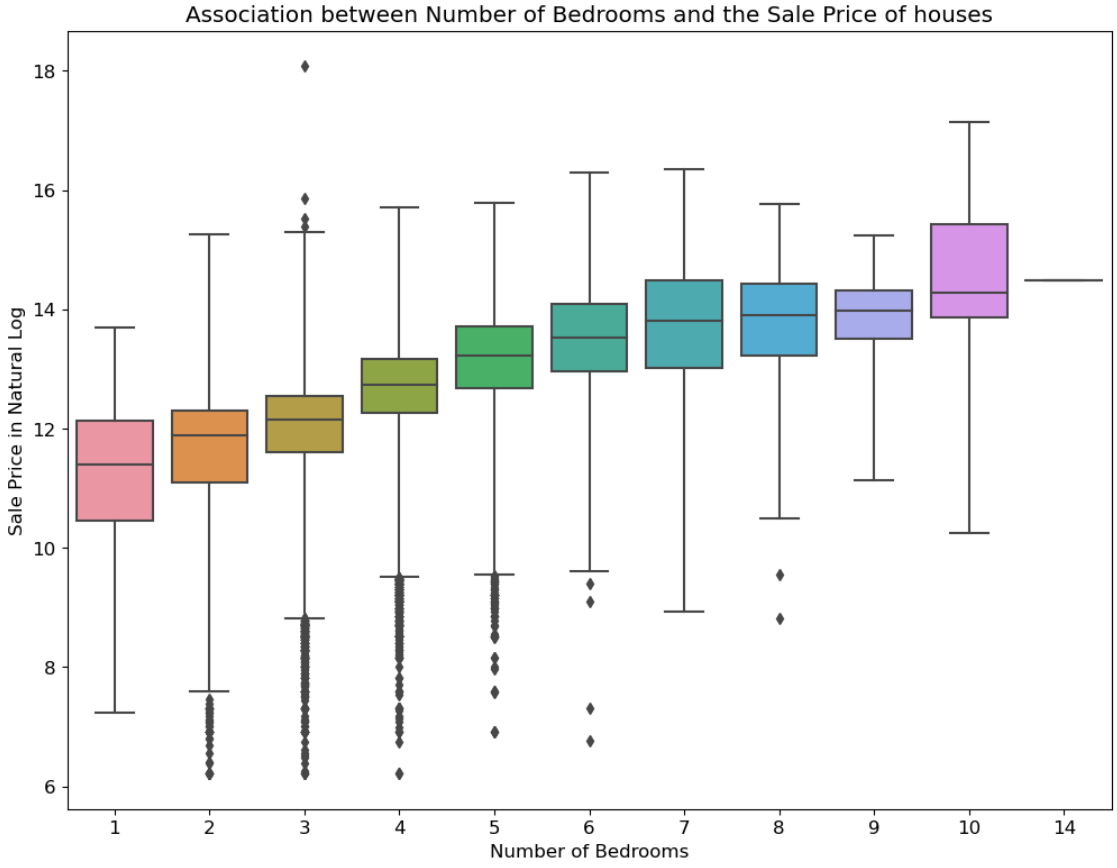
Square Feet. As Log Building Square Feet increases, Log Sale Price also increases. Therefore, Log Building Square Feet would make a good candidate as one of the features for our model.

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 [206]: sns.boxplot(x='Bedrooms', y='Log Sale Price', data=training_data, whis=3)
          plt.xlabel('Number of Bedrooms')
          plt.ylabel('Sale Price in Natural Log')
          plt.title('Association between Number of Bedrooms and the Sale Price of houses');
```



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, there seems to be no significant relationship between the hosues' log sale price and their neighborhoods. We don't see any significant difference in house prices for different neighborhoods, besides maybe neighborhood code 120. So we can't say there is a correlation between log sale price and their neighborhoods.