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/record represents a house.

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.

I believe the purpose of recording and collecting data on housing was to understand what factors affect the value of a property and why. For instance, age, location, condition and property size are the four important factors that affects the value of a property or a house. The data was collected by Cook County, Illinios.

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.

I believe census tract is one of the feature that contain demographic information because through census tract we can access to population, race, housing occupancy, group quarters population data.

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 1: how would you assess if there is a relationship between the "Sale Price" and the "Land Square Feet"? Answer 1: I would create a scatter plot of "Land Square Feet" vs "Sale Price" in order to see if there is a relationship between these two features.

Question 2: how would you assess the frequency distribution of the "Sale Price"? Answer 2: I would plot the histogram of the "Sale Price" in order to assess the distribution of the "Sale Price" and analyse the SD and mean of the distribution.

1.2 Question 2

1.2.1 Part 1

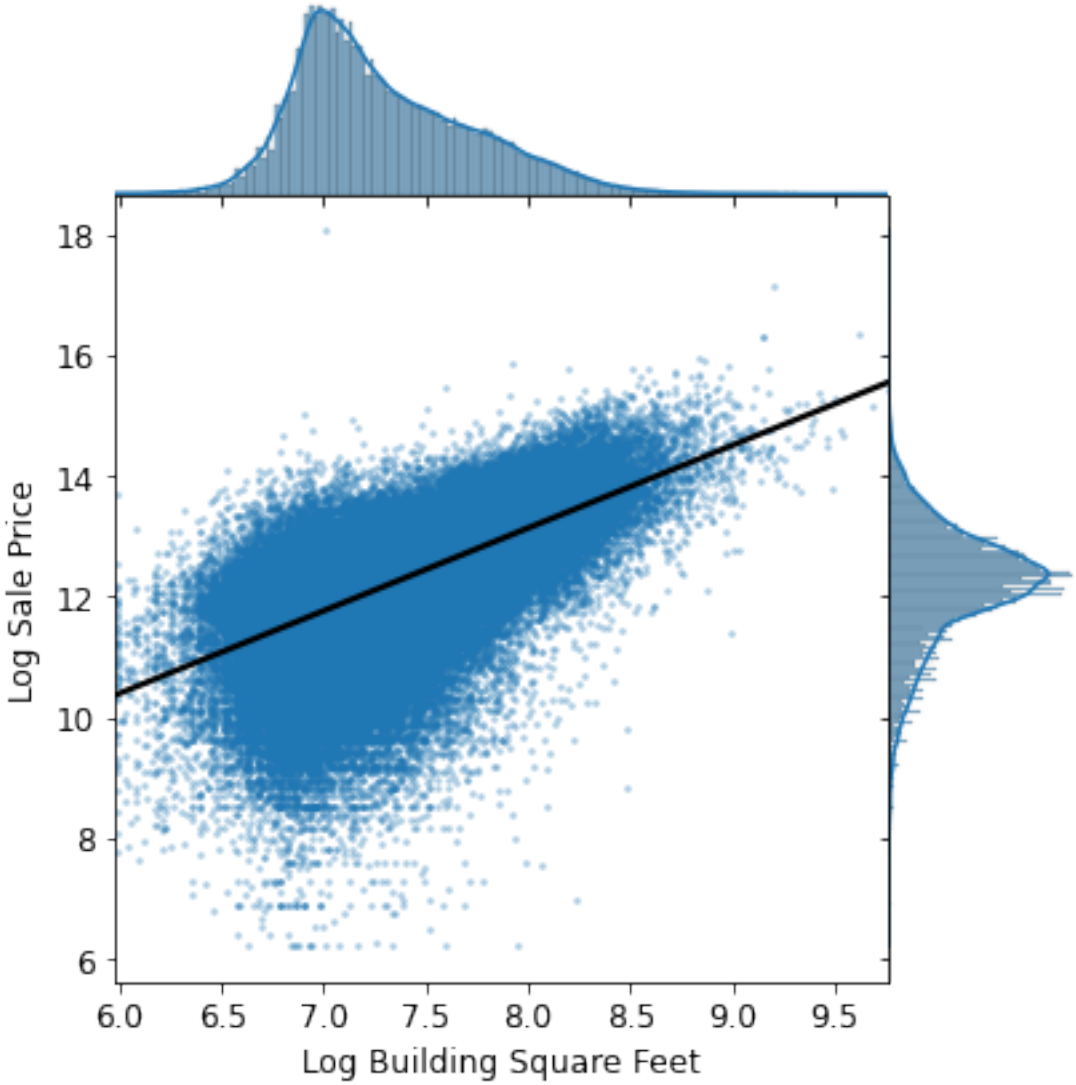
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The "Sale Price" values ranges hugely due to, perhaps, outliers. We can resolve this issue by removing the outliers.

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?



There is a positive correlation between "Log Sale Price" and "Log Building Square Feet". Also, "Log Building

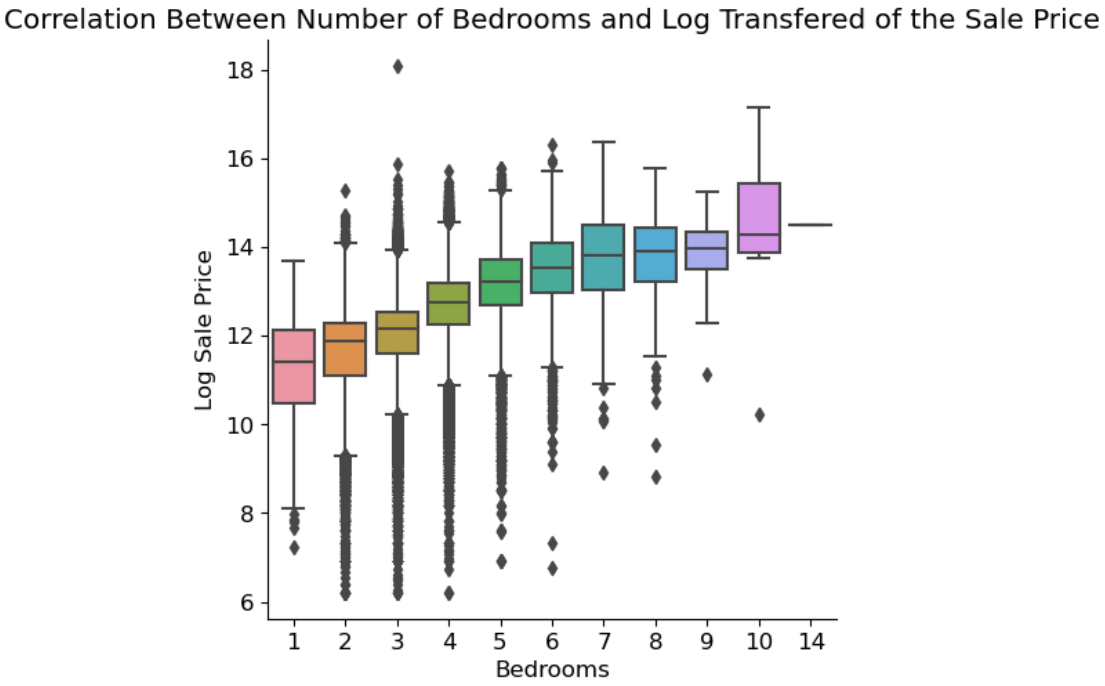
Square Feet" is a good candidate for our model.

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 [28]: #training_data_copy = training_data
         #training_data_copy['Log Bedrooms'] = np.log(training_data_copy['Bedrooms'])
         (sns.catplot(data=training_data, x = 'Bedrooms',
                      y = 'Log Sale Price', kind='box')
                     )
         plt.title('Correlation Between Number of Bedrooms and Log Transfered of the 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?

It's an empty plot and I don't think there is a relationship between the "Neighborhood Code" and "Log Sale Price"