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 1a

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 an individual property sale in Cook County, Illinois.

1.2 Question 1b

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 would imagine that some kind of local government entity would collect this kind of data - either the county or state level. It could be used by the govenment for a variety of reasons including tax planning, zoning, city planning, affordable housing assessments, and other governmental tasks. However it looks like this data set could also be used by a justice group like a nonprofit to prove inqueity and the presence of systemic issues. Or a bank mightfind this useful when calculating loans or an insurance company for calculating coverage rates.

1.3 Question 1c

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.

What month of the year is the most popular for home sales in this county? I would calculate the mean number of homes sold per month across all years of the data set and then find the month with the max. I would use the 'Sale Month of the Year' column and 'Sale Year' to do so.

Do homes with an attached garage have a higher sale price than homes without? I would define a house with an attached garage as having either "Garage 1 Attachment" or "Garage 2 Attachment" or both marked as a yes. Then I would make a side by side box plot showing the sale price data (gathered from the "Sale Price"column) of those properties with an attached garage compared to those without.

1.4 Question 1d

Suppose now, in addition to the information already contained in the dataset, you also have access to several new columns containing demographic data about the owner, including race/ethnicity, gender, age, annual income, and occupation. Provide one new question about housing in Cook County that can be answered using at least one column of demographic data and at least one column of existing data and provide the type of analytical tool you would use to answer it.

Do people with lower incomes own older homes? I would make a scatter plot of age of the home vs income of the owner and look for a correlation. I would use the columns "Age Decade" and "Annual Income".

1.5 Question 2a

Using the plots above and the descriptive statistics from `training_data['Sale Price'].describe()` in the cells above, identify one issue with the visualization above and briefly describe one way to overcome it.

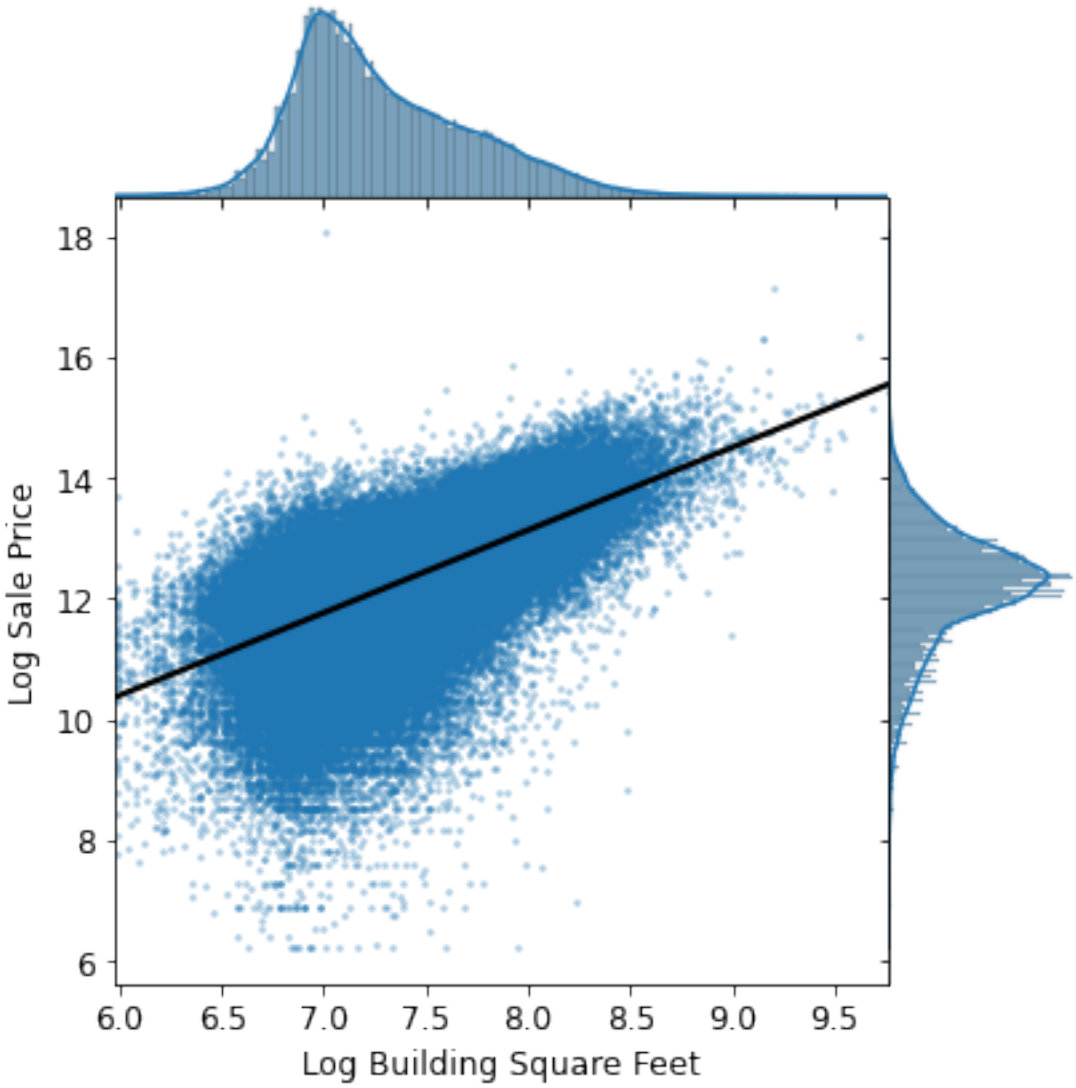
It is difficult to see any of the information because it is visually so scrunched together. The scale of the visualizations is not effective. One way to overcome this would be to remove the extreme outliers. It looks like the max sale price was 70 million and the min is 1 which is throwing our scale way off. We could consider running the same cells as above but using data with a sale price between 1000 and 1 million to exclude the outliers.

1.6 Question 3c

In the visualization below, we created a `jointplot` 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, would `Log Building Square Feet` make a good candidate as one of the features for our model? Why or why not?

Hint: To help answer this question, ask yourself: what kind of relationship does a "good" feature share with the target variable we aim to predict?



“Log Building Square Feet” could make a good candidate for a model feature. This seems like a good feature since there is an apparent linear relationship between these two variables. The regression line goes also goes relatively through the middle of the data so there is not a major area that is being neglected.

1.7 Question 5c

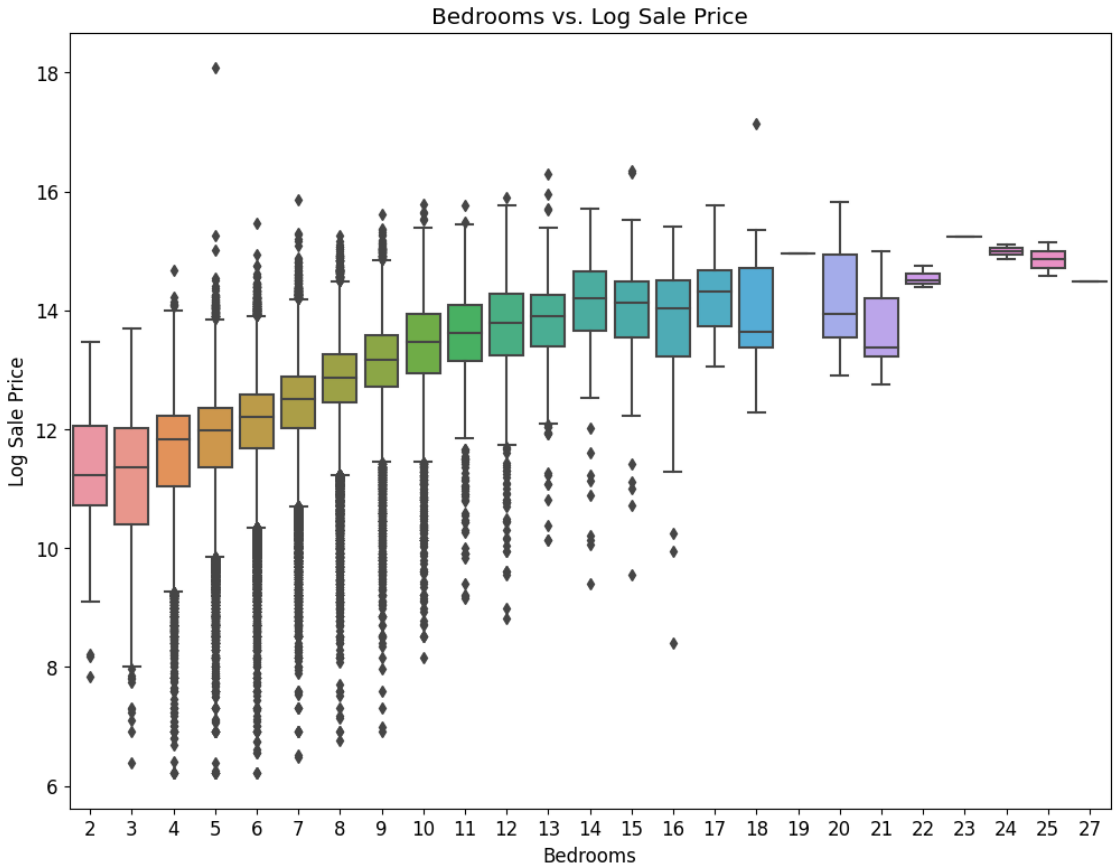
Create a visualization that clearly and succinctly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and a succinct title. - It should convey the strength of the correlation between `Sale Price` and the number of rooms: in other words, you should be able to look at the plot and describe the general relationship between `Log Sale Price` and `Bedrooms`

Hint: A direct scatter plot of the `Sale Price` against the number of rooms for all of the households in our training data would result in overplotting (since there are only a small discrete number of bedrooms) - so **don't use a scatter plot**.

```
In [29]: x = training_data["Bedrooms"]
         y = training_data["Log Sale Price"]

         sns.boxplot(x = x, y =y)
         plt.title("Bedrooms vs. Log Sale Price")
```

```
Out[29]: Text(0.5, 1.0, 'Bedrooms vs. Log Sale Price')
```


Bedrooms vs. Log Sale Price
Log Sale Price
Bedrooms
18
16
14
12
10
8
6
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
27