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?

A: one sale of a house in Cook County.

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.

A: By CCAO for the purpose of property taxation.

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.

SOLUTION: Answers should identify at least one of the following:

1. **Census Tract** could be linked to data from the US Census, which contains tract-level statistics regarding household size, ethnicity, income, etc.
2. **Neighborhood Code** and **Town Code** could conceivably be linked to neighborhood- and town-level statistics that would be similar to the Census demographic data.
3. Some other variable with a description of the direct demographic data it embeds or that it could when joined with another data set.

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.

解决方案: 答案会随着下面列出的一些可能性而变化。1. 这段时间库克县的平均房价是多少? 我将计算销售价格的平均值和中位数。2. 一年中哪个月的销售价格最高? 我将创建一个线形图，显示整个销售月份的中间销售价格。3. 机场附近的价格低吗? 我会计算包含奥黑尔噪声的销售和不包含奥黑尔噪声的销售的中间价格。4. 库克县的房屋建设历史如何?T 将制作一张库克县的地图，每个社区都有一个基于平均建筑的颜色阴影。这将需要 L 个独立的数据集

1.2 Question 2

1.2.1 Part 1

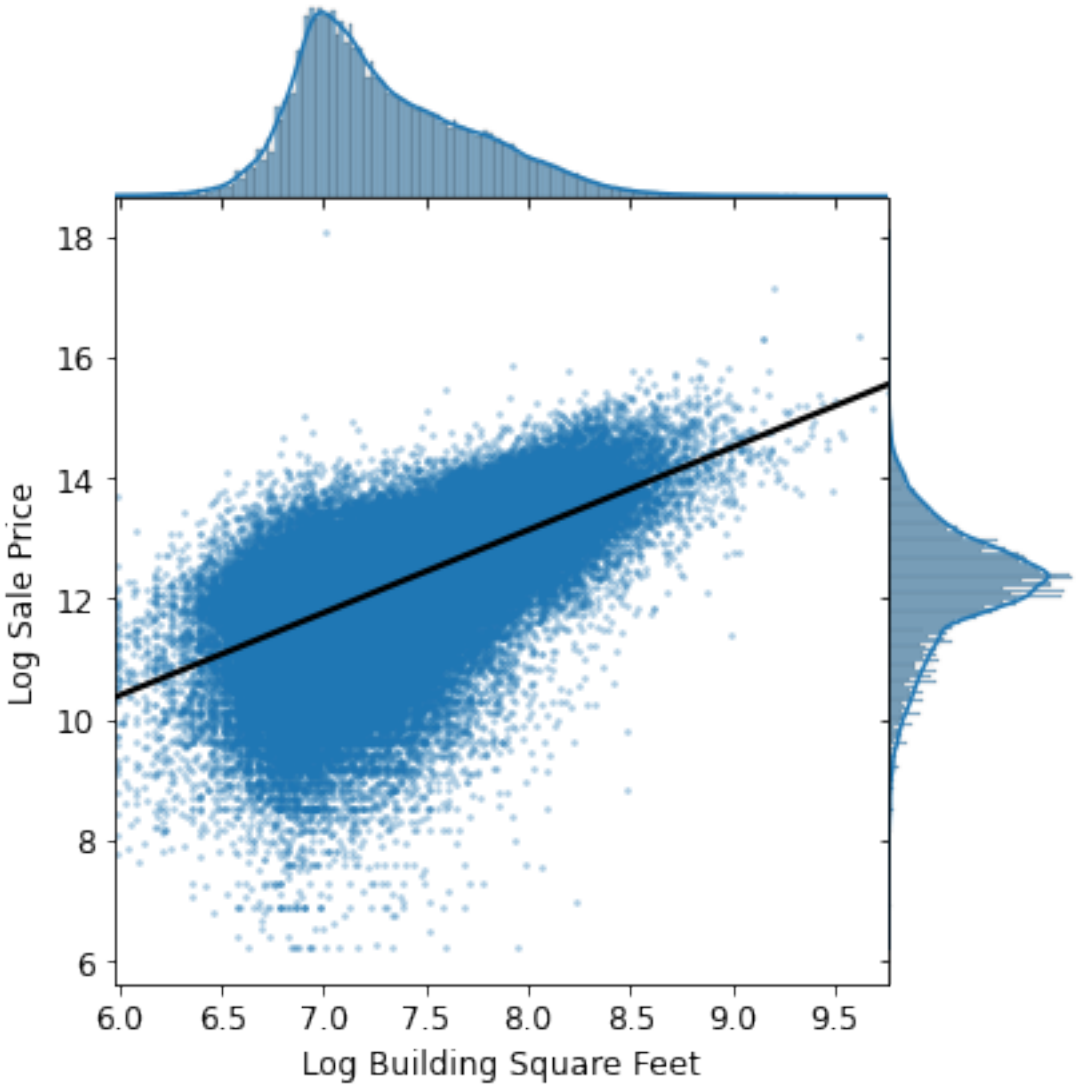
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

解决方案: 在分布的最右端存在极端的异常值，这将过度延伸图表的范围，使得大多数数据几乎不可能可视化。克服这个问题的一种方法是通过移除数据中的异常值来缩小范围。

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?



解决方案：log 销售价格和 log 建筑面积之间似乎有很强的相关性。由于它们之间有很强的相关性，Log

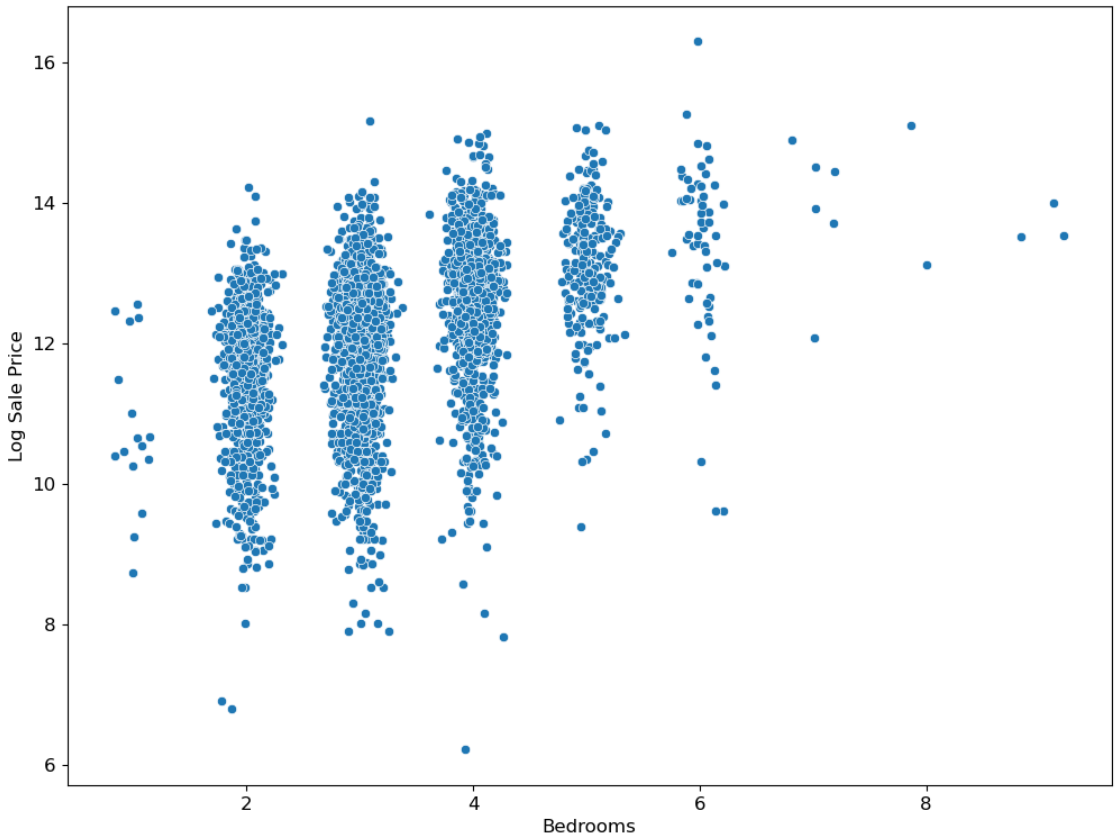
Building 平方英尺确实是我们模型的关键特性之一。

1.2.3 Part 3

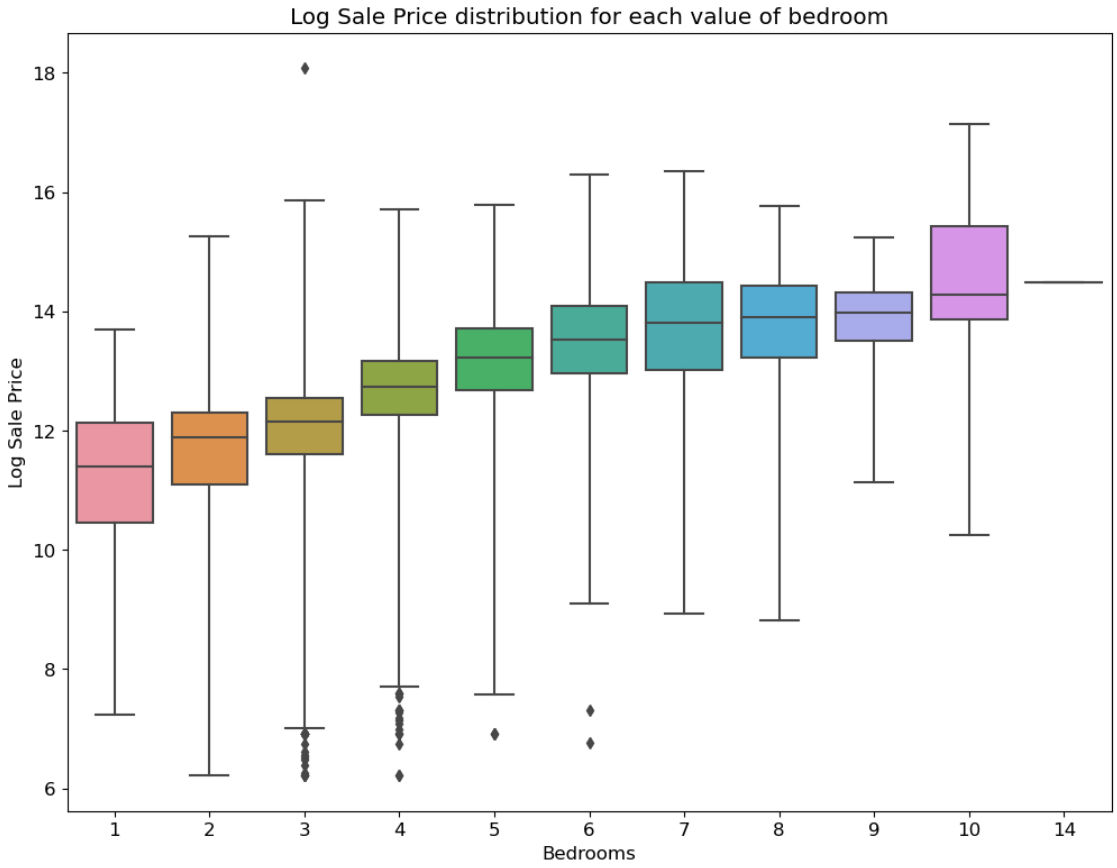
Create a visualization that clearly and succintly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and succinct title. - It should convey 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]: plt.figure()
         sampled_data = training_data.sample(5000, replace=False)
         sns.scatterplot(
             x = sampled_data['Bedrooms'] + (np.random.normal(size=5000)) / 10,
             y = sampled_data['Log Sale Price']
         )
         plt.show()
```



In [29]:
```
# SOLUTION
plt.figure()
sns.boxplot(
    x = 'Bedrooms',
    y = 'Log Sale Price',
    data = training_data,
    whis = 5,                # 异常值比例
)
plt.title('Log Sale Price distribution for each value of bedroom')
plt.show()
```



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?

解决方案: 很明显，除了 12 号和 120 号街区外，各个社区的房价变化一般都不是太大。此外，可用的数据量并不是在各个社区之间均匀分布的。例如，30 号小区共有 8751 户，而 71 号小区只有 30 号小区的 27% 左右。