0.1 Question 1a

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

Each row represents various features per property in Cook County.

0.2 Question 1b

Why was this data collected? For what purposes? By whom?

You should watch Lecture 15 before attempting this question.

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

This data was collected to make an inference about sales price based on various features of the properties. Government could have collected this data in order to manange their finance. They need to get price information about each people's properties to add taxes. Also, this data could have been collected by real estate workers. It is because this daya would be really useful to people who want to sell their properties. They can predict what the sales price of their properties might be by looking at the other properties which have similar features to theirs.

0.3 Question 1c

Craft at least two questions about housing in Cook County that can be answered with this dataset 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 datasets you would need to answer that question.

The relationship between the Land Square Feet and the Sale Price could be answered with this dataset by creating a scatter plot of the log of both variables. The relationship betweeen the garage size and the sale price can also be answered. I would create a box plot for each garage size and see the distribution of the sale price across different garage sizes (dropping values of 7 which corresponds to garage size of 0)

0.4 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.

I think we can use scikit-learn to perform a correlation analysis. We will employ the 'Race/Ethnicity' column as a categorical variable and the 'Age' column as a continuous variable. By using a suitable tool for categorical vs. continuous data correlation, such as point-biserial correlation, we can assess whether there is a statistically significant relationship between the race/ethnicity of property owners and the age of the properties they own.

0.5 Question 1e

Look at `codebook.txt` to see some of the unique regional features CCAO utilizes, such as `O'Hare Noise`. Now imagine you were in charge of predicting the `Sale Price` of houses in **your hometown** (your actual real life hometown/city - not the data provided). Propose a feature that you would want to collect specific to your location and hypothesize why it might be useful in predicting the sale price of houses.

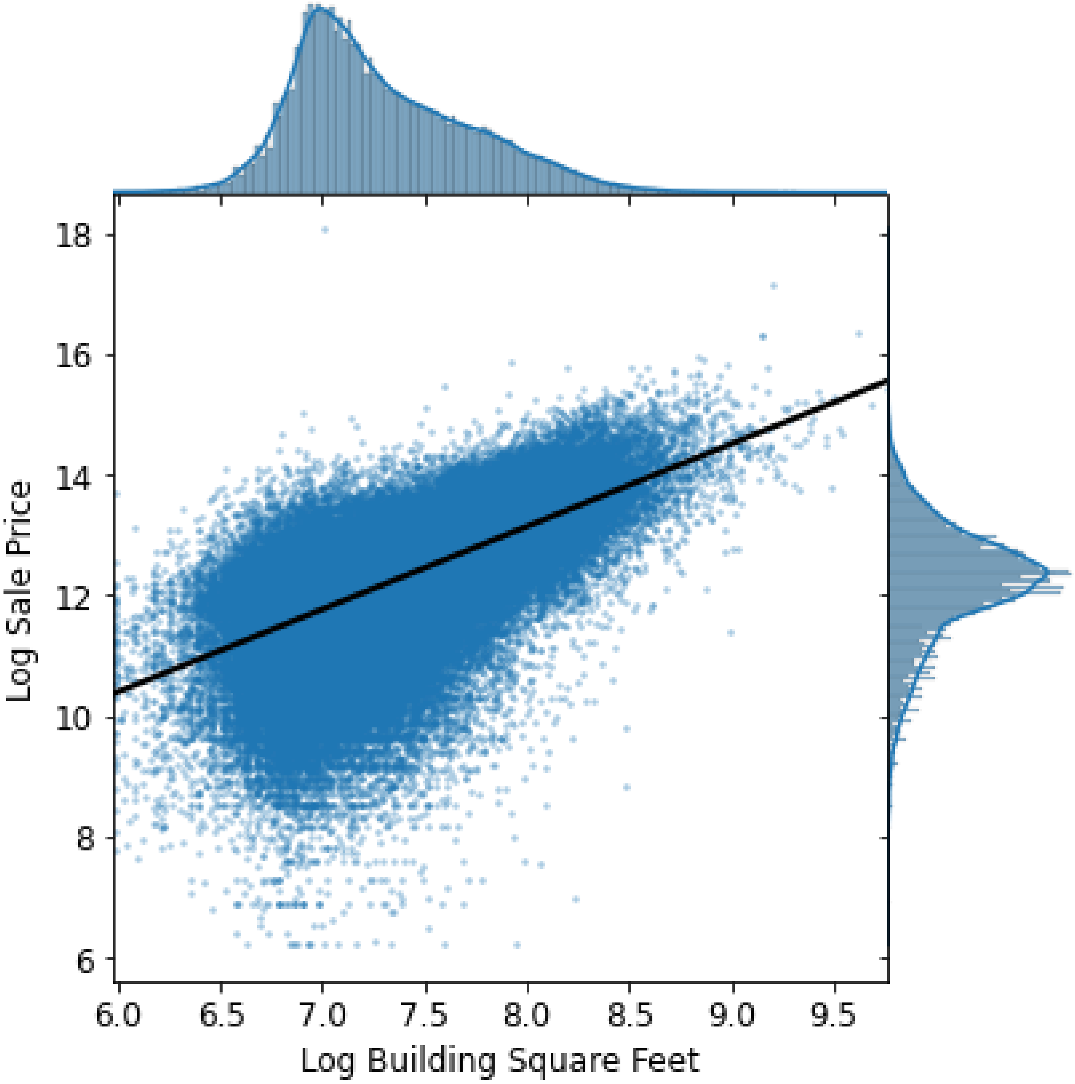
I would give feature Name "Proximity to Public Transit". it's Useful: Homes closer to public transit (bus, subway, train) are generally more desirable, leading to higher sale prices. Easy access to transit reduces commute times, increasing demand for these properties. New transit developments can drive property appreciation over time. Potential Drawbacks: In car-dependent cities, transit access may not significantly impact prices. Homes too close to transit lines may suffer from noise pollution, reducing value.

0.6 Question 3b

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?



I thinkLog Building Square feet would make a good candidate for one of the features for our model because based on scatter plot, we can see that it has the linear relationship between Log Building Square Feet and log sale Price. Moreover, it also has a positive slope.

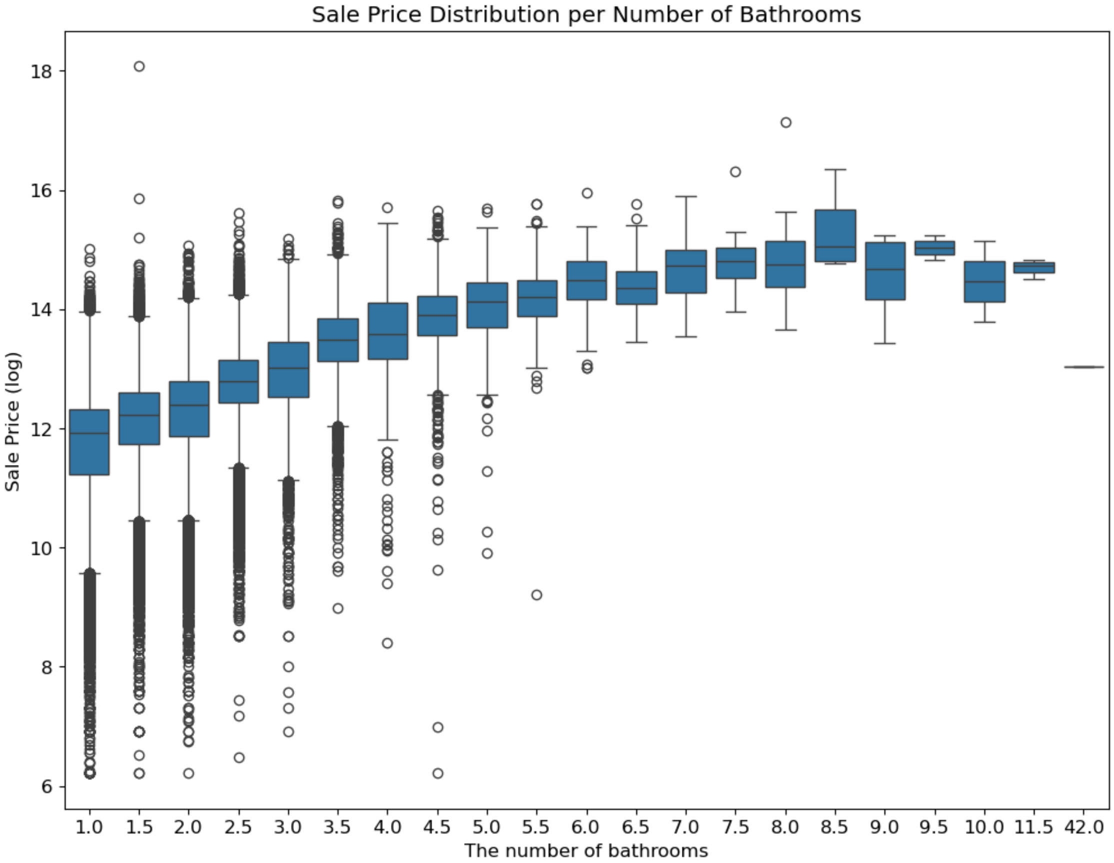
0.7 Question 5c

Create a visualization that clearly and succinctly shows if there exists an association between Bathrooms 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 Bathrooms.

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 [166]: sns.boxplot(data=training_data, y='Log Sale Price', x='Bathrooms')
          plt.xlabel("The number of bathrooms")
          plt.ylabel("Sale Price (log)")
          plt.title("Sale Price Distribution per Number of Bathrooms")
```

```
Out[166]: Text(0.5, 1.0, 'Sale Price Distribution per Number of Bathrooms')
```


Sale Price Distribution per Number of Bathrooms
Sale Price (log)
18
16
14
12
10
8
6
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
7.0
7.5
8.0
8.5
9.0
9.5
10.0
11.5
42.0
The number of bathrooms