

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 some kind of property or real estate in Cook County, Illinois which each columns show specific properties of what materials, location, sale, and other descriptions.

The granularity of this dataset is high and shows individual real estate property of 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.

The data above could be applied to many things. One example I believe this was collected for is to see property/real estate assessment purposes. Maybe one of the reasons could be to evaluate the price of some property on what land, location, materials, size, etc. to determine the property tax or some other kind of tax. With these data, real estate companies could predict what kind of houses are trending, what people would buy the houses, and how to increase the prices of the house itself and to determine if they should increase tax prices as well.

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.

1. What is the correlation of the size of the building and its property tax? To determine this I would create a box plot of ‘Land Square Feet’ and ‘Sale Price’
2. How does age of the house relate to its sale price? To determine this I would create a scatter plot between ‘Sale Price’ and ‘Age Decade’.

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.

How does the correlation of annual income relate to the Property's Value? I would use a scatter plot between 'Annual Income' and 'Sale Price'. This would help me determine if rich people would buy higher priced property.

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.

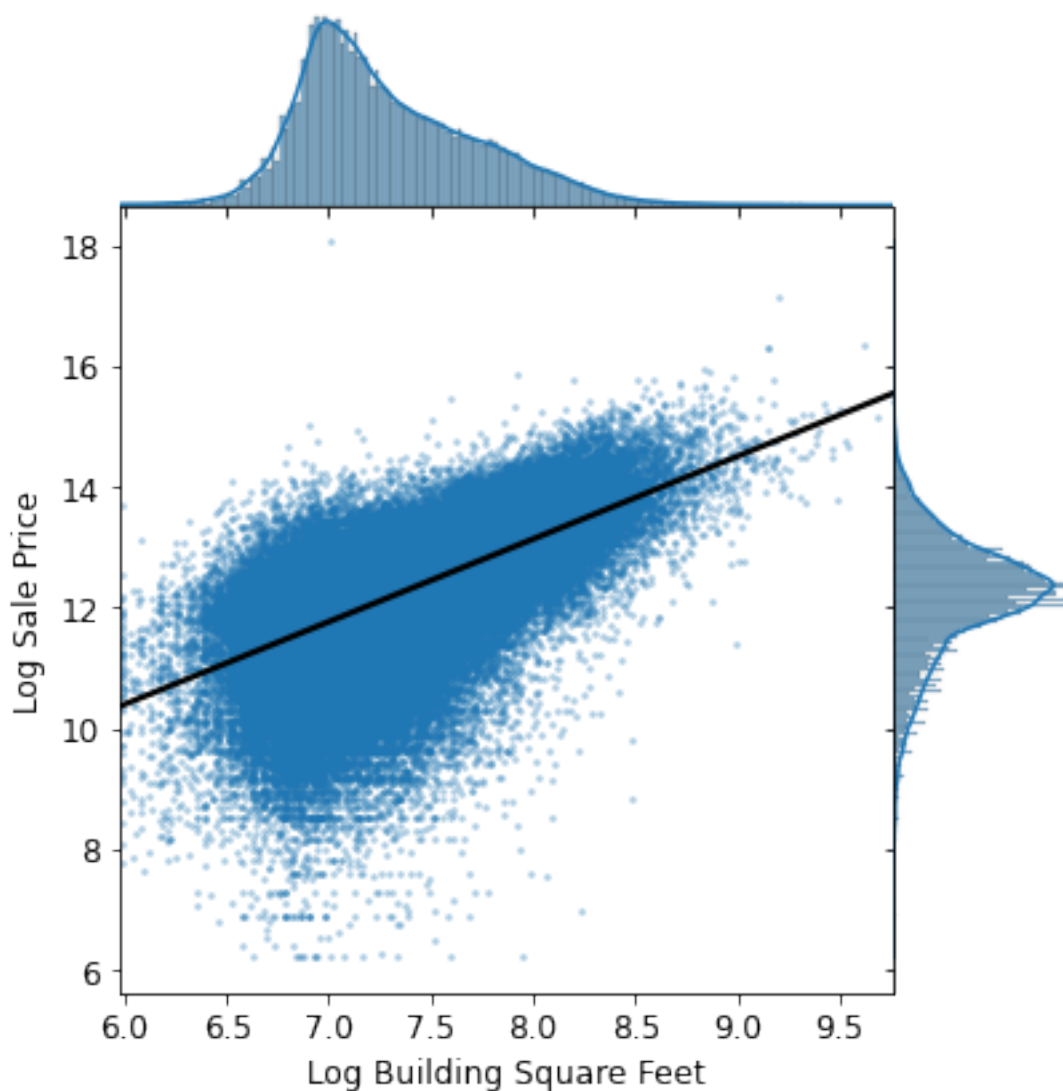
We see that one issue is that the graph is very right skewed. One way to overcome this is have a log transformation to the 'Sale Price', which may make this easier to visualize the graph as it will be closer to a symmetric bell curve.

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 does look like a good candidate as one of the features for the model. This is because we see positive correlation and linear relationship shows that building square feet might be a good predictor for sale prices.

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 [28]: # sns.set(style="whitegrid")

sns.boxplot(data=training_data, x='Bedrooms', y='Log Sale Price')
plt.title('Bedrooms vs Log Sale Price')
plt.xlabel('Number of Bedrooms')
plt.ylabel('Log Sale Price')
plt.show()
```

