
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?

The granularity of this dataset (what each row represents) is each house in Chicago.

0.2 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.

Given the detailed information on how these houses were built, this data was likely collected to gain insights into the qualities of a structurally sound, well-built home. Additionally, it may help answer questions like, "What characteristics do high-selling houses share?" or "Which features are most appealing to buyers?" since we also have data about location and other relevant attributes.

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.

"What defines a luxury or high-end home in Cook County?"

To answer this, I would calculate the percentage of high-priced homes that include specific luxury features, using columns like "Sale Price," "Building Square Feet," and "Construction Quality." Then, I would filter for homes with two garages (using "Garage 1 Size" and "Garage 2 Size") and high construction quality (using the "Construction Quality" column) to estimate the typical cost of a high-end home with these features. I would likely calculate the mean and median "Sale Price" for homes that meet these criteria to provide a summary of their cost.

"How does proximity to noise or environmental factors impact home sale prices in Cook County?"

To answer this, I would investigate the impact of environmental factors on home sale prices by analyzing columns like "O'Hare Noise" and "Floodplain" alongside "Sale Price." I would create boxplots or violin plots comparing sale prices of homes that are affected by these factors (e.g., near noise from O'Hare Airport or in a floodplain) to those that are not. This would help visualize how proximity to noise and environmental risks impacts the sale prices of homes in Cook County. Additionally, I could calculate the average "Sale Price" for homes located within these zones versus homes outside of them to quantify the price difference.

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.

"Are there differences in property preferences (e.g., number of rooms, number of bathrooms, number of stories, or sale price) based on the age or income of homeowners in Cook County?"

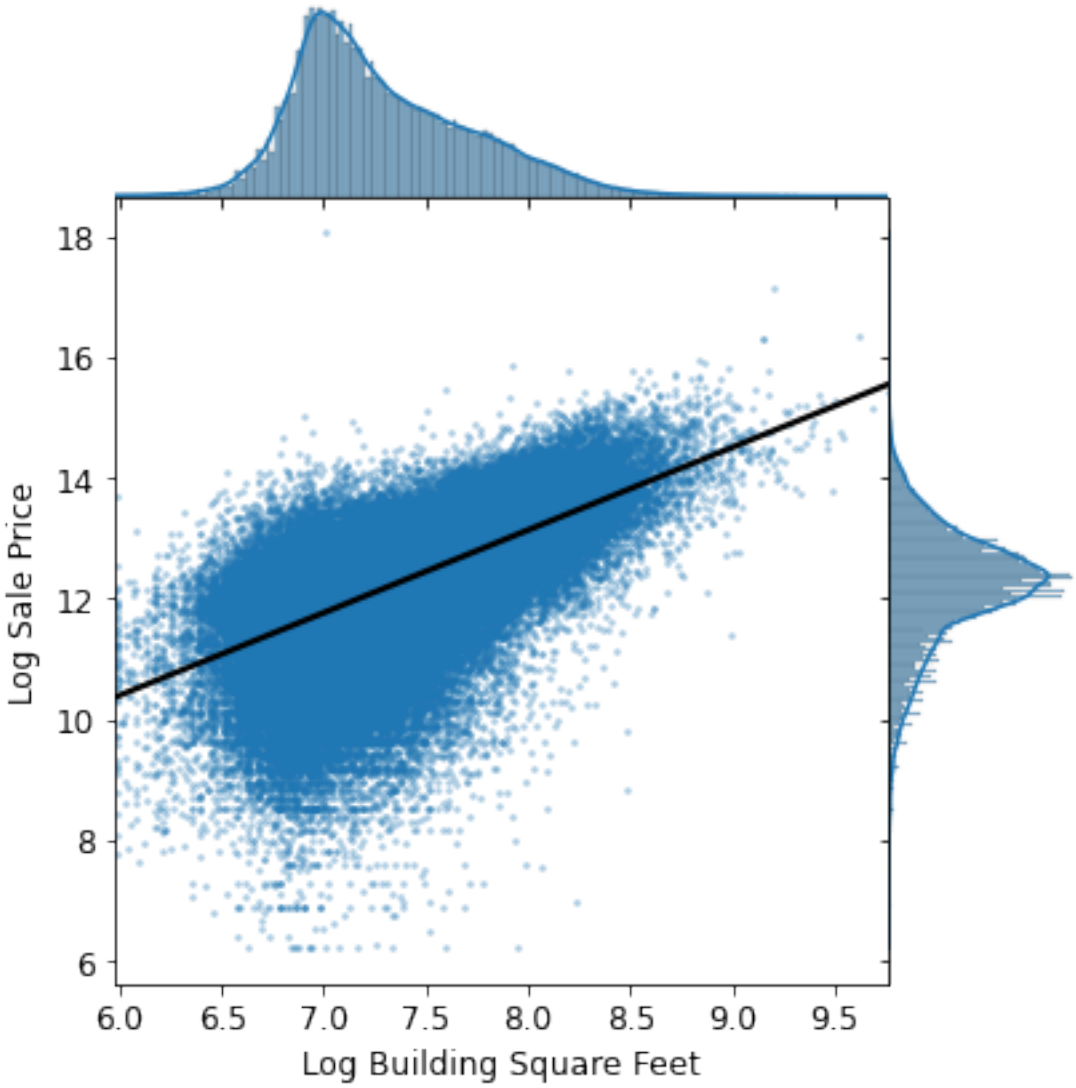
To answer this, I would use a combination of summary statistics and visualizations. Specifically, I would calculate the average "Sale Price," "Number of Rooms," "Number of Bathrooms," and "Number of Stories" for different age groups and income brackets, using the new "Age" and "Annual Income" columns, alongside the existing property data. I would then create box plots or violin plots to compare these property features across different age and income groups.

0.5 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?



This joint plot indicates that Log Building Square Feet would make a good candidate as one of the features of our model because the linear regression line fits through the center of the data point cluster. Although the residuals on both sides of the line suggest some room for improvement, the general alignment of the data with the regression line shows a reasonably good fit.

0.6 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 might risk overplotting.

```
In [309]: sns.boxplot(data = training_data, x ="Bedrooms", y ="Log Sale Price")
          plt.title('Association of Sale Price and Number of Bedrooms Increase')
```

```
Out[309]: Text(0.5, 1.0, 'Association of Sale Price and Number of Bedrooms Increase')
```