
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.

0.2 Question 1b

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

You should watch Lecture 15 before attempting this question.

This data was collected to infer the sales price based on various property features. The data can be represented whether the price of the house is fair or not, whether they have a valid price to sell and whether the tax amount is also fair within the community.

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.

I would create a scatter plot of Sales Price to visualize the correlation between price and the number of rooms and bathrooms. Since the number of rooms and bathrooms plays a crucial role in determining a house's price, analyzing this relationship will provide insights into how prices are influenced by these factors.

Additionally, I would like to visualize the distribution of sales across various addresses using neighborhood codes. Similar to the number of rooms, the neighborhood information is important, as understanding the characteristics of the surrounding area helps create a clearer picture of the market. I believe analyzing this data would also yield interesting insights.

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 would like to examine the relationship between the "Age" of the household owner and the neighborhood, specifically how house sales have changed based on the owner's age.

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.

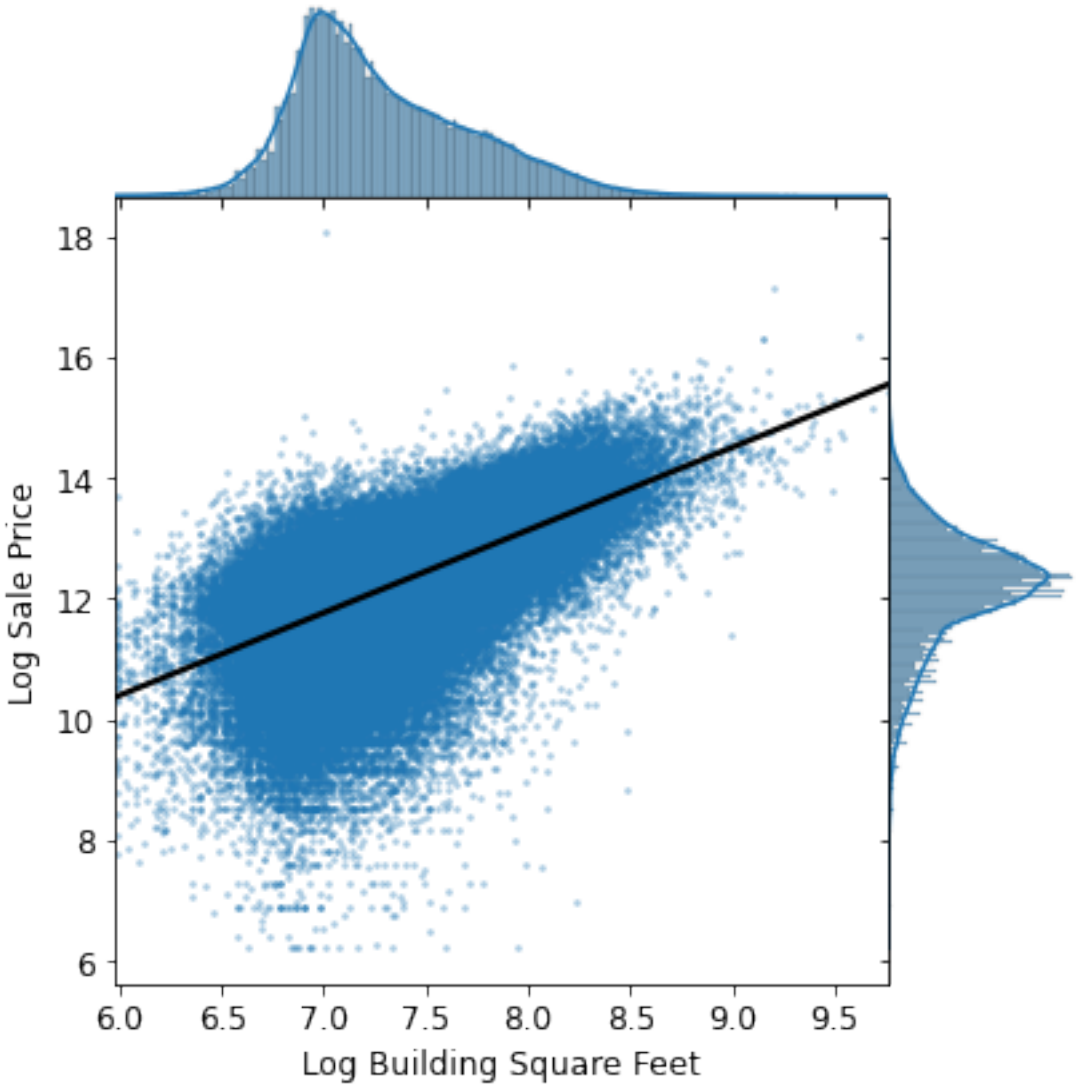
Within predicting the Sales Price, I would like to collect data, such as "Garage Size", "Building Square Feet", 'Sale Year','Sale Quarter', 'Sale Half-Year', 'Sale Quarter of Year','Sale Month of Year', 'Sale Half of Year', 'Most Recent Sale', this information are important to predict the Sales Price of houses that size, month, and the size of the garage is important deciding the price.

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?



Log Building Square Feet is not a suitable candidate as one of the model's features. The plot graph reveals that the data points are widely scattered, indicating a large variance. This wide dispersion suggests that Log Building Square Feet lacks a clear, consistent pattern, making it less effective as a predictive feature for the model.

0.7 Question 5c

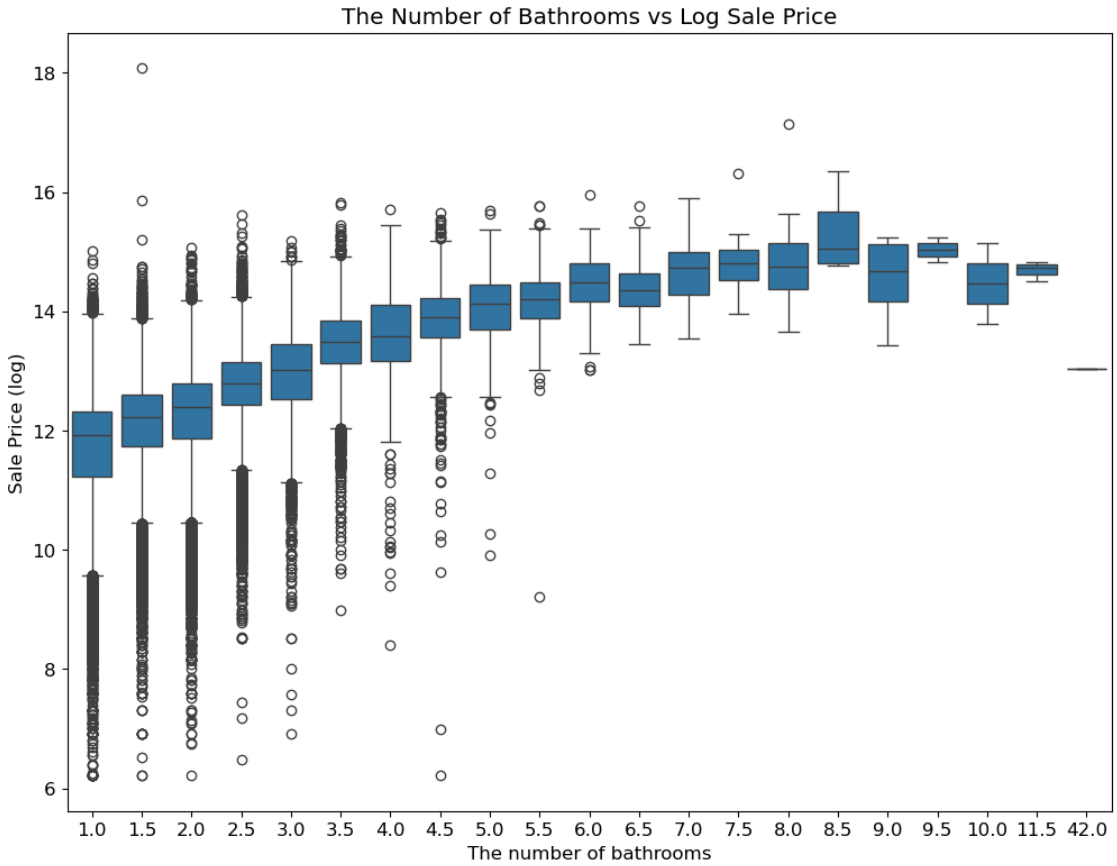
Create a visualization that clearly and succinctly shows if there exists an association between `Bathrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and a succinct title. - It should convey 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 [260]: training_data["Description"].iloc[1]
```

```
Out[260]: 'This property, sold on 02/18/2016, is a one-story houeshold located at 11415 S PRAIRIE AVE.I
```

```
In [261]: sns.boxplot(data = training_data,
                      x = "Bathrooms",
                      y = "Log Sale Price")
          plt.xlabel("The number of bathrooms")
          plt.ylabel("Sale Price (log)")
          plt.title("The Number of Bathrooms vs Log Sale Price");
```


The Number of Bathrooms vs Log Sale Price
Sale Price (log)
The number of bathrooms
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