
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?

Houses sold in Cook County

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.

Data might be collected to analyze housing market or broader economy. Could be useful for real estate agents.

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.

- Do bigger houses sell for higer prices? I would create a scatterplot of sale price and building square feet.
- Do houses with different property classes sell for different prices? I would create a boxplot of sale price and property class.

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.

Does gender impact sales price? To examine the relationship between sales price and Land Square Feet and gender, create a linear regression with gender, Land Square Feet and Age as the independent variable. We will get two lines when the gender coefficient is 0 and 1.

0.5 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.

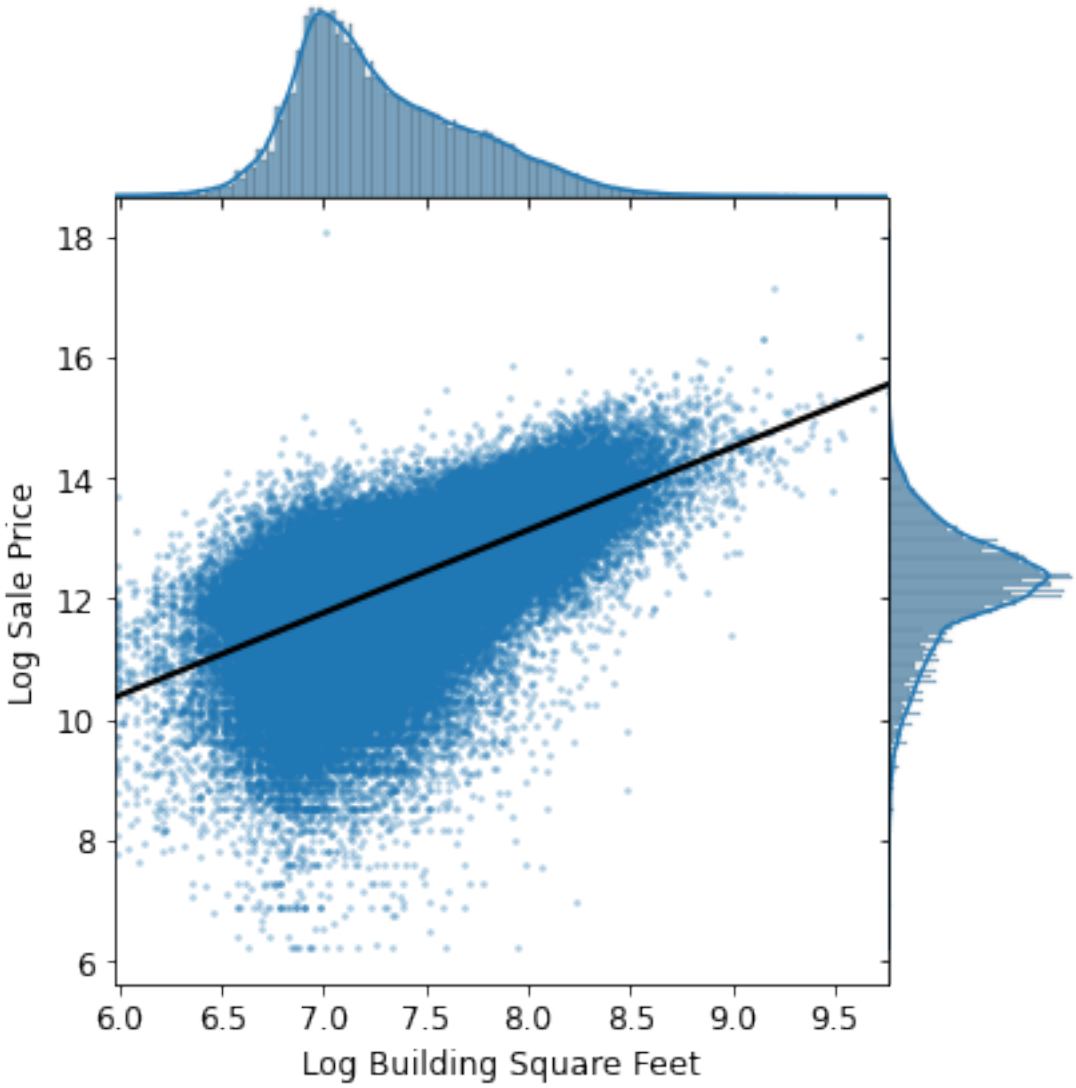
It is heavily skewed due to the outliers- very expensive housing. Maybe perform log transformation t0 transform skewed distribution to a normal distribution.

0.6 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?



Yes. It displays a clear moderate linear relationship between the two independent variables. Linearized relationship can help us easier to understand the underlying relationship between the variables.

0.7 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 might risk overplotting.

```
In [192]: sns.boxplot(data=training_data, x='Bedrooms', y='Log Sale Price')
          plt.title('Boxplot of Log Sale Price distributions of Number of Bedrooms')
```

```
Out[192]: Text(0.5, 1.0, 'Boxplot of Log Sale Price distributions of Number of Bedrooms')
```