
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 corresponds to different features of a unique property in Cook County, Illinois

---

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.

In [ ]:

This data could have been collected by realters looking to use the various specifications and features for different properties in order to estimate their potential sale price. This would allow them to use a reliable metric on which to base their starting property prices and sell houses to customers. This information is also useful for housing regulation agencies and property managements to get a well described chart in order to map properties in a specific region.

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 with the column's neighborhood code and sale price to identify if there are any relationships between the area in which the property is located and its corresponding sale price. I could also calculate the mean Land Square feet for properties grouped by their neighborhood code maybe to see if there is any pattern between the neighborhood properties are located in and their land size.

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.

We could use a histogram to compare the proportion of properties(by number) owned by different ethnicities in the community

---

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.

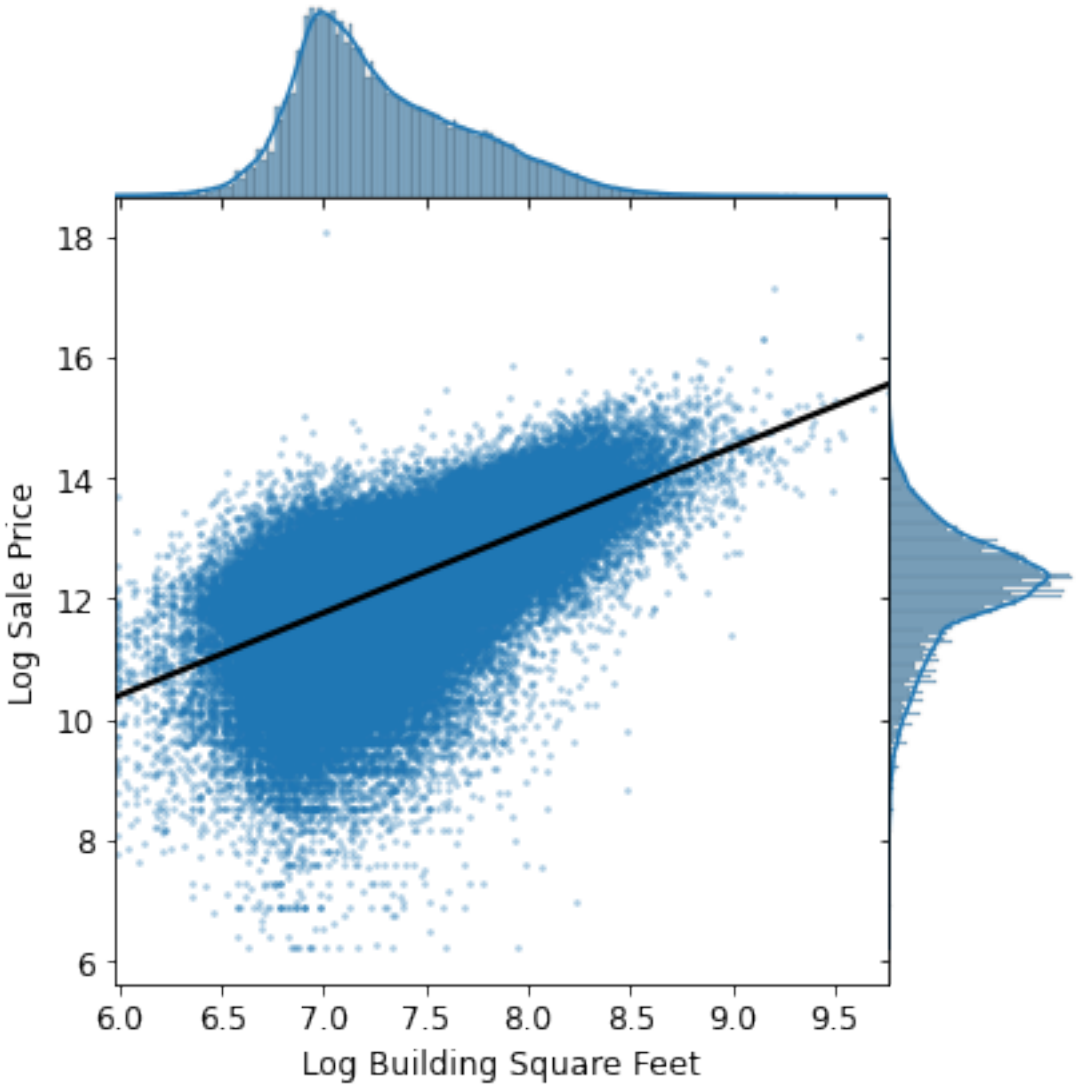
The scale of the x-axis is too big, and so since most of the sale prices are focused around 10 million they are all bunched to the left of the axis. To overcome this the scale of the axis could be reduced, such that the x-axis could start at 0 and end at 0.5 X 10^7, for example.

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 is a good candidate. Since we are trying to estimate the Sale Price for the properties based on the feature, we can see from the plot that there is somewhat of a linear relationship between building square feet and sale price. Therefore, building square feet is a property that we must consider when estimating the sale price, making it a good candidate to be one of the features

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 [122]: plt.figure(figsize=(10,6))
          sns.violinplot(data=training_data, x="Bedrooms", y="Log Sale Price")
          plt.xlabel('Number of Bedrooms')
          plt.ylabel('Log Sale Price')
          plt.title('Association between Bedrooms and Log Sale Price')
          plt.show()
```