
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 property.

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.

The data might be collected by the country's government to estimate property value.

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.

1. Is there a correlation between the number of central heatings and the sale price of various properties? I would create a box plot of 'Central Heating' and 'Sale Price' to visualize the correlation.
2. What is the maximum sale price for each unique neighborhood code in year 2019? I will use data filtering to select only the rows of 'Sale Year' = 2019 and group the dataframe by 'Neighborhood Code' and calculate the maximum sale price for each group.

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.

How does the annual income of property owners relate to property sale price in Cook County? I would create a scatterplot of 'annual income' and 'Sale Price' to visualize the correlation between the two variables.

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.

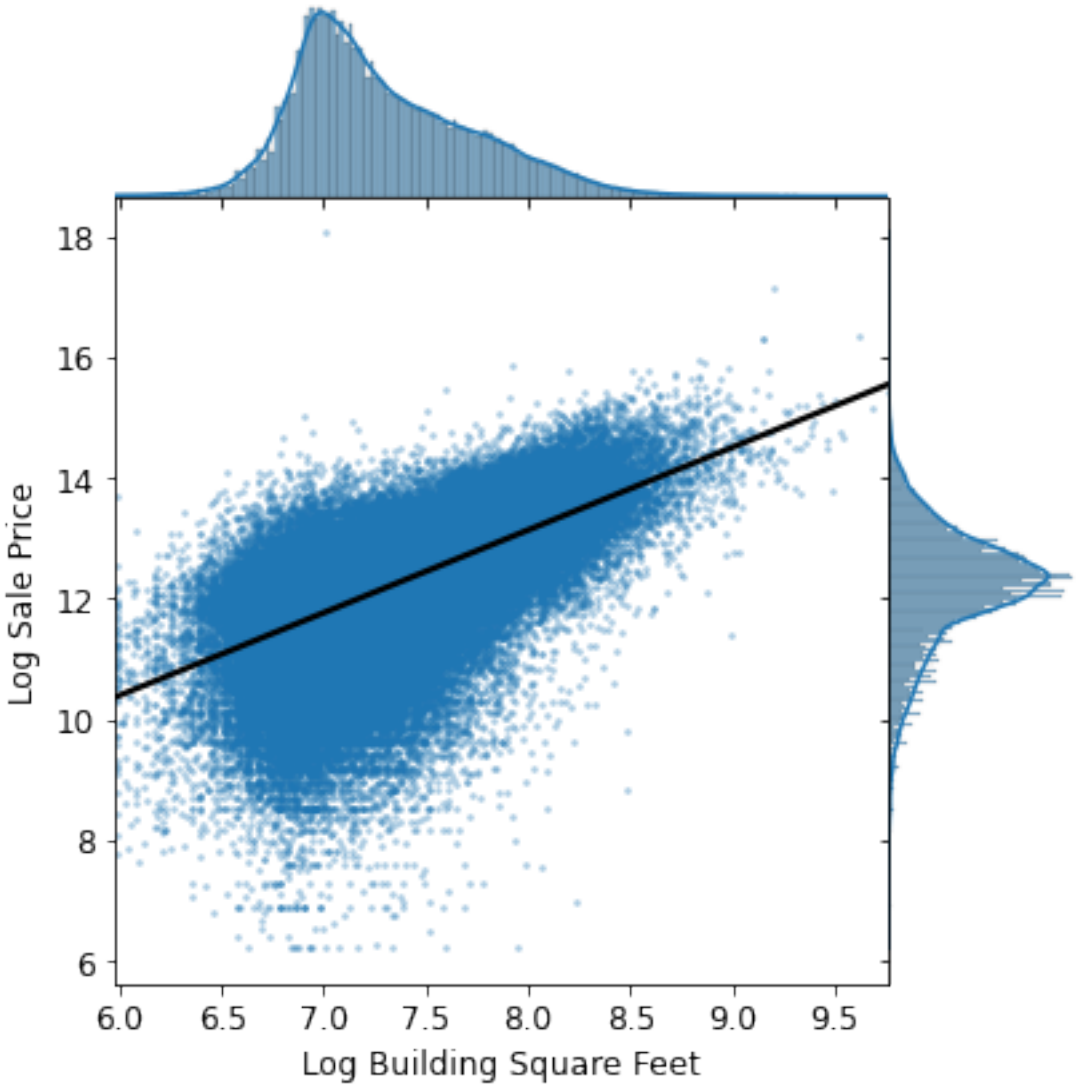
The outliers of the dataset causes the x-axis interval to not fit most of the data, we can remove the outliers to solve the problem.

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?



There's a positive correlation between log building square feet and log sale price as the datapoints are plotted along the linear regression line, therefore log building square feet should be consdiered a feature for the model. However, there are still many outliers to clean up to improve the fit(especially for properties with square feet 6.5-7.5).

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 [102]: sns.boxplot(data = training_data, x = 'Bedrooms', y = 'Log Sale Price', showfliers = False)
          plt.title('Correlation between log Sale Price & number of Bedrooms')
```

```
Out[102]: Text(0.5, 1.0, 'Correlation between log Sale Price & number of Bedrooms')
```