
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 a single house at a single location in cook county. Every house in the dataset has an associated sale price, so it must have been sold at least once. That means that each row represents a house which has previously been 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.

The data was collected by the Cook County assesors office for the purpose of determining taxable property value on the basis of predicted sale price.

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.

QUESTION 1: If I were an investor, I might want to know how the neighborhood a building is in affects its "site desirability" rating. **TOOL 1:** I would create side-by-side box plots where each category coresponds to a unique value in the "Neighborhood Code" column and each box represents the distribution of values in the 'Site Desirability' column for each category. **QUESTION 2:** If I were a developer, I might want to know which consruction features have the largest influence on the "Construction Quality" rating. **TOOL 2:** I could group the dataset by the values in each construction variable column (E.g. 'Wall Material', 'Roof Material', etc) and find the average value of the "Construction Quality" rating for each group. Using the codebook to translate the materail codes, I could then determine the materials that were associated with the higest construction quality ratings.

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.

Question: How does local racial heterogeneity effect property value? **Tool:** Using the 'Longitude' and 'Latitude' columns, you can identify the closest "k" houses to any individual house with a method similar to a KNN algoritim. You can then gather the information about the race of the owners of the k nearest houses to generate a "racial heterogeneity score" for each house in the dataset. Then you could plot the sale price of each house against this new variable to investigate how the racial hertero/homogeneity of the local area might influence the sale prices of each house.

0.5 Question 2a

Using the plots and the descriptive statistics from `initial_data['Sale Price'].describe()` above, identify one issue with the visualization above and briefly describe one way to overcome it.

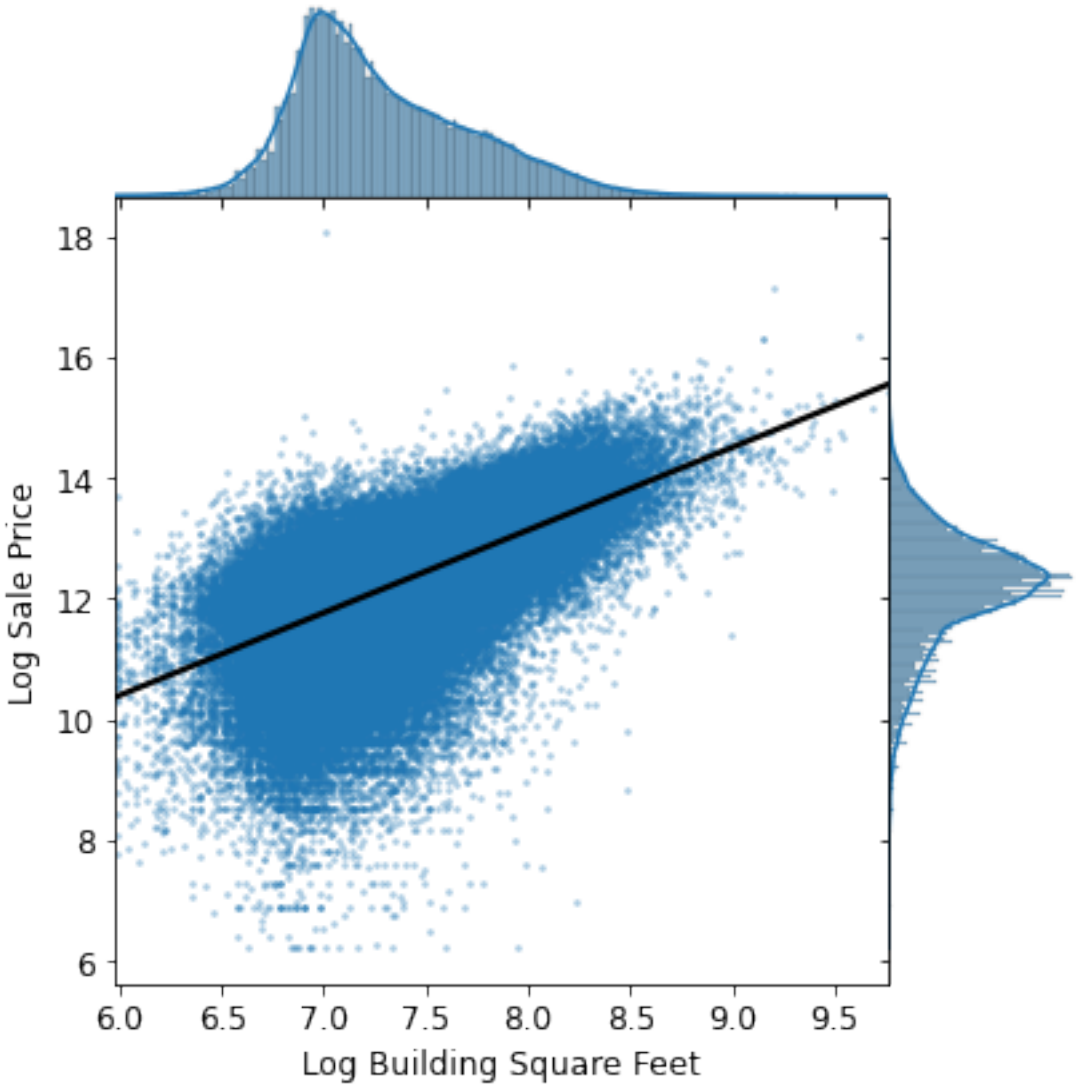
There is at least one huge outlier at a value of 7.100000e+07 which is distorting the x axis and making it impossible to see the trends for the bulk of the data. This can be resolved by changing the range displayed on the x-axis to focus on the distribution of the values closer to the mean.

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, log sqft would be a good candidate for a variable in our model. This is because, as indicated by the figure, it is moderately correlated with log sale price, which means that log sqft can function as a partial predictor of sale price.

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 [119]: sns.stripplot(data=training_data, x='Bedrooms', y='Log Sale Price', jitter=0.3, size=2, alpha=
          m, b = np.polyfit(training_data['Bedrooms'], training_data['Log Sale Price'], 1)
          plt.plot(training_data['Bedrooms'], m*training_data['Bedrooms'] + b, color='red', label="Regr
          plt.title("Log Sale Price by Number of Bedrooms (x values jittered)")
          plt.legend();
```