
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 seems to represent a property with all its details. Area it covers, details of any buildings on it, where it is located, etc.

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.

A dataset like this would most likely be collected by a City/State in order to better understand the housing environment/demographic in the area and compare that trend over time in order to better understand how the City/State is growing/decreasing.

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.

How would you predict the price of a home with relation to its variables? OLS regression on the estimated price of a building with respect to number of rooms and the sqft.

How would you predict the price of a land with relation to where its located? OLS regression on the estimated price land with respect to its longitude and latitude.

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.

The average price of the property depending on the race/ethnicity of the owner. I would first groupby race/ethnicity and then np.mean the Estimate(Land) and Estimate(Building) and create a new column with the addition of the two of them called Estimate(Property). Doing this, I would be able to see the average Land, Building, and Property estimate for each race/ethnicity.

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.

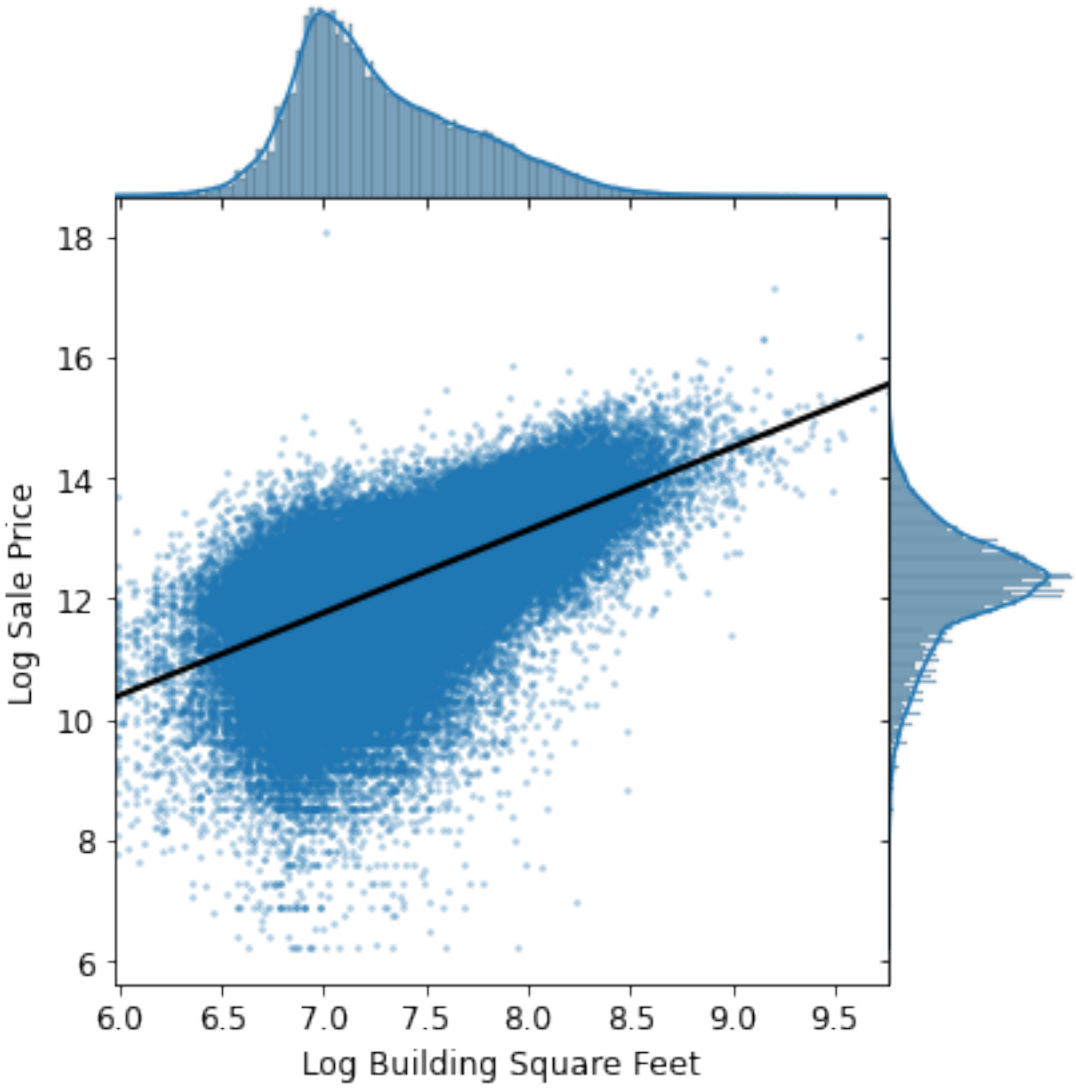
The visualization above seems to be close to useless since you can't even see the quartiles of the graph. It is being skewed far too much by the max sale price of (7.1e+07). We could either zoom in on the mean of the data or remove the outliers.

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?



Positive correlation between building square feet and sale price. This implies that square footage would be a 'good' feature to predict the sale price for our model.

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 [97]: sns.boxplot(x='Bedrooms', y='Log Sale Price', data=training_data)
         plt.title('Log Sale Price vs. Bedrooms on a property')
```

```
Out[97]: Text(0.5, 1.0, 'Log Sale Price vs. Bedrooms on a property')
```