
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 unique property record with its property attributes (size, materials), sales details (sale year, sale price), and geolocation details (longitude, latitude, Neighborhood Code). This dataset contains Fine-grained data.

0.2 Question 1b

Certain variables in this dataset contain information that either directly contains demographic information (data on people) or could reveal demographic information when linked to other datasets. Identify at least one demographic-related variable and explain the nature of the demographic data it embeds.

The census tract feature can be linked with the U.S. Census Bureau Data, which can be used to obtain a wide range of demographic information about the residents of each property, including age, race, sex, income, and education levels.

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. I can use this dataset to answer the question of what is the most popular size of houses people like to choose in Cook County, I would create a histogram of 'Building Square Feet'. 2. Whether the size of the garage affects the price of houses in Cook County, I would first create a new column 'garage_total_size' of total area of garage1 and garage2 and then I will make a lineplot of 'garage_total_size' and 'Sale Price'.

0.4 Question 2a

Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards as the autograder may not work otherwise.

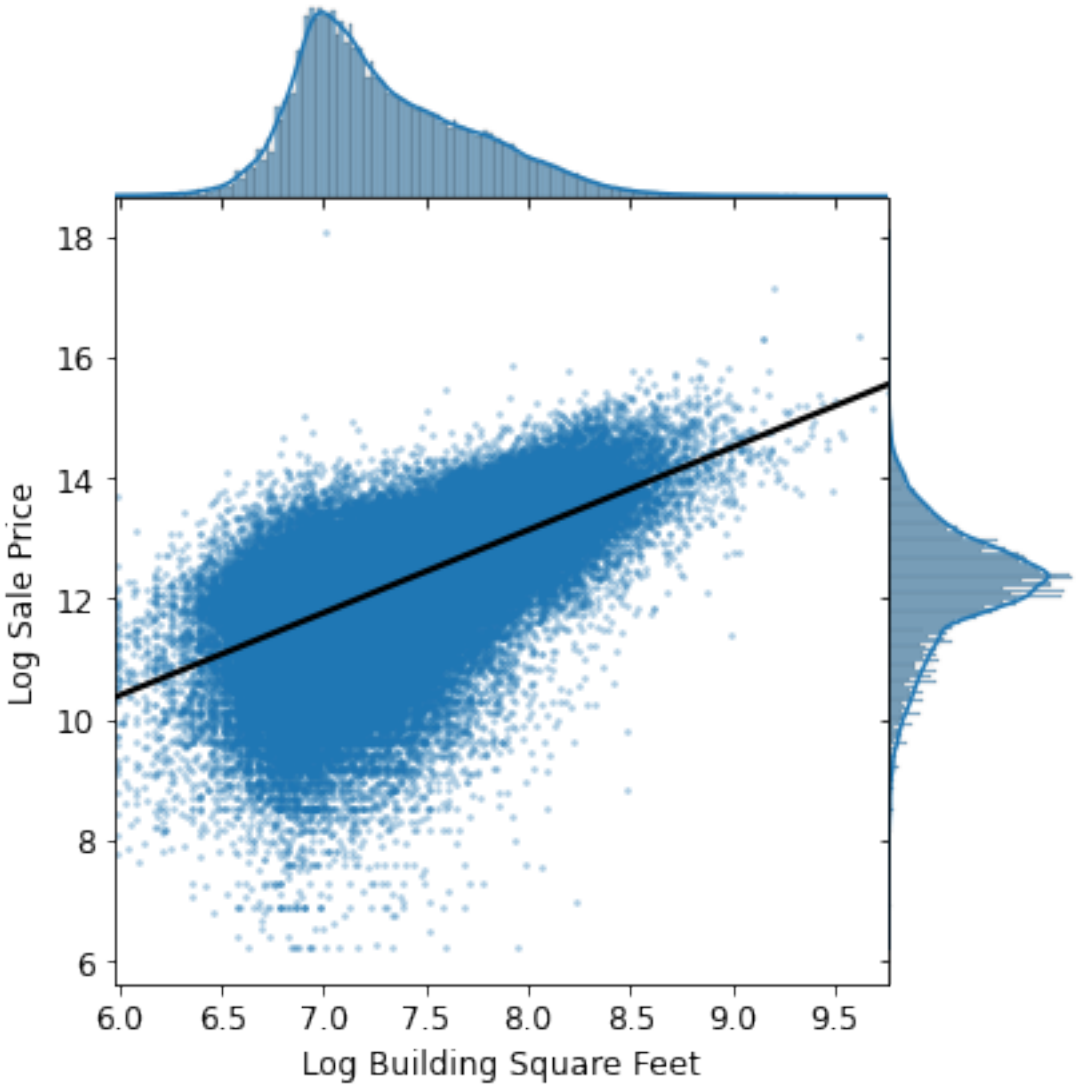
We need to scale the 'Sale Price' so we can see more details of the 'Sale Price' distribution, one way we can do is to transform our data.

0.5 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?



'Log Building Square Feet' is not a good feature candidate for our model. Looking at the graph, we can see there are many data points below the linear regression line and so many overplotting in the plot.

0.6 Question 5c

Create a visualization that clearly and succintly 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 succinct title. - It should convey the strength of the correlation between the 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 [25]: plt.figure(figsize=(10, 6))
         sns.boxplot(x='Bedrooms', y='Log Sale Price', data=training_data, palette="Set1")
         plt.title('Bedrooms vs Log Sale Price');
```