
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 different property in Cook County. (Each record has its own pin).

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 possibly collected to analyze amenities of different properties. Researchers may be interested in analyzing which features are prioritized more among residents, which features are more popular to have than others, and if there is an association between seemingly unrelated features.

0.3 Question 1c

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.

Garage size could indicate the size of groups living in each property. For example, a property with a size of 0 or 1 may house only 1 resident whereas a property with a greater size may house a couple or family. This would be linked with the use feature, which says if the residents are single-family or multi-family.

0.4 Question 1d

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.

What is the relationship between aviation and vehicle traffic for a given property? I would create an overlaid scatter plot of Site Desirability on Road Proximity with two different distributions based on the value of the O'Hare Noise feature. What is the relationship between the size of the family (Use) and the size of the garage(s) (Garage 1 Size, Garage 2 Size)? I would calculate the expected value of garage size for a single and multi-family resident group.

0.5 Question 2a

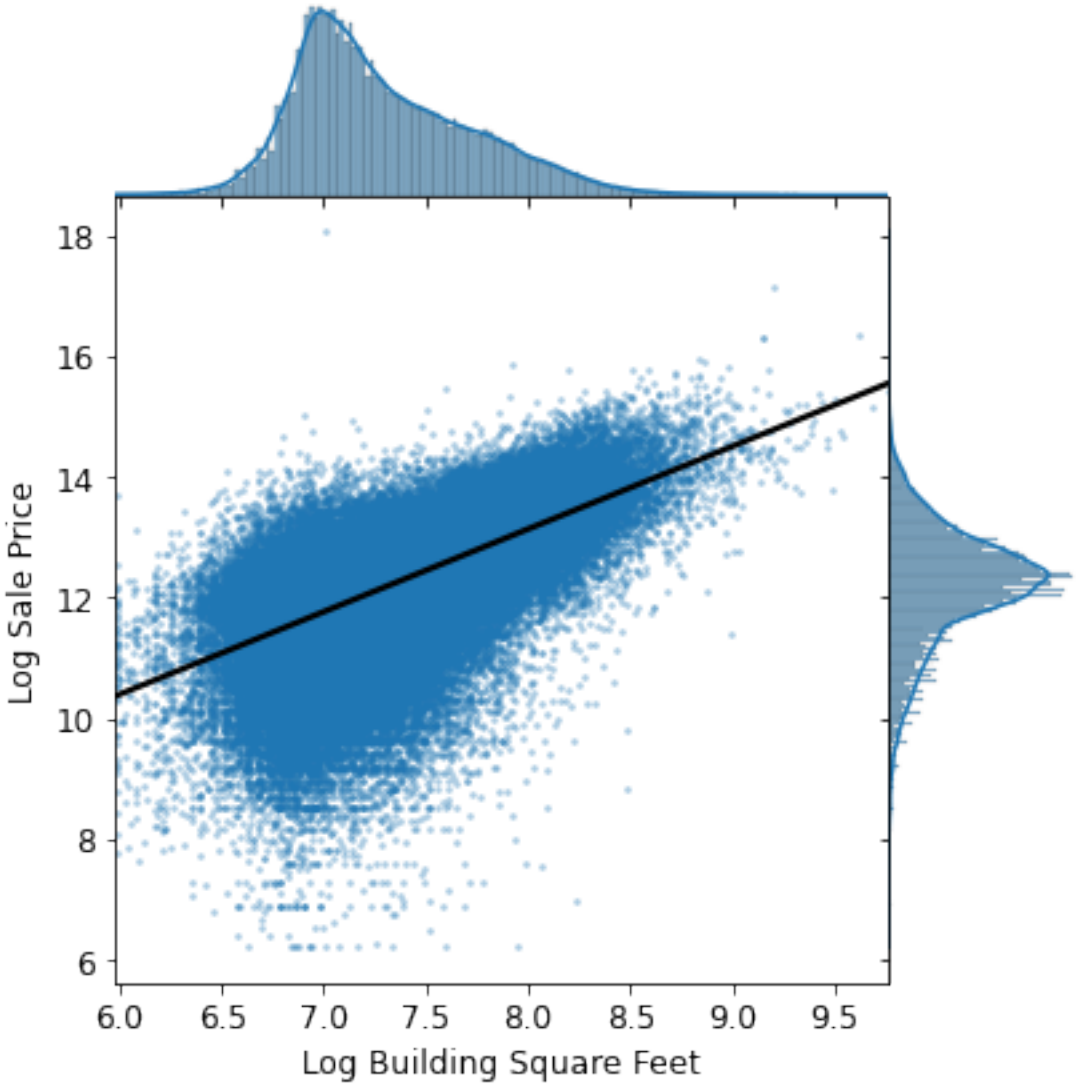
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The x-axis scale is not optimal for visualizating the distribution. Since the limits were decided to be the minimum and maximum, it is including the outliers way to the right of the rest of the data, causing it to be heavily skewed. In order to fix this issue, we can log the x-values so that they are closer together and the data is less crushed.

0.6 Question 3c

As shown 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?



Since there is a clear positive association between log building square feet and log sale price, we can say that

Log Building Square Feet would make a good candidate as one of the features of our model. However, outliers are definitely visible in the scatterplot.

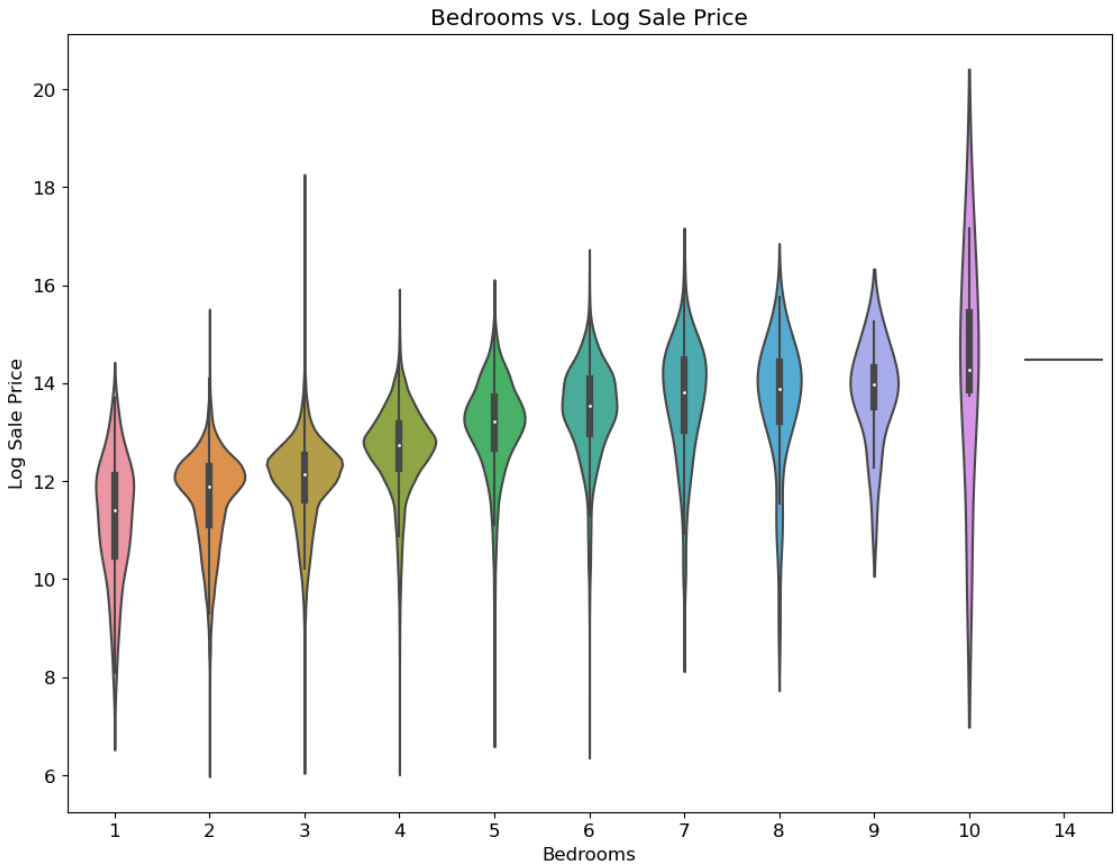
0.7 Question 5c

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.

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 [24]: sns.violinplot(data=training_data, x='Bedrooms', y='Log Sale Price')
         plt.title('Bedrooms vs. Log Sale Price')
```

```
Out[24]: Text(0.5, 1.0, 'Bedrooms vs. Log Sale Price')
```



0.8 Question 6c

It looks a lot better now than before, right? Based on the plot above, what can be said about the relationship between the houses' `Log Sale Price` and their neighborhoods? Is there a relationship?

There appears to be no correlation between the neighborhood code and the log sale price because there is no positive or negative pattern. All values are roughly centered around 12.