1 Part 1: Contextualizing the Data

Let's try to understand the background of our dataset before diving into a full-scale analysis.

1.1 Question 1

1.1.1 Part 1

Based on the columns present in this data set and the values that they take, what do you think each row represents? That is, what is the granularity of this data set?

each row is a property that can be sold in the cook county, with corresponding information.

1.1.2 Part 2

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.

data could be collected for dealers with purposes of making a sale or record all available information. Most likely by the property dealers/agency to document relevent information of all properties on the territory for sale or in the scope of potentially buying them.

1.1.3 Part 3

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

variables like census tracking and use are demographic. One of which is information of occupant of the terrotory and the other one is whether the house is shared by one family or multiple.

1.1.4 Part 4

Craft at least two questions about housing in Cook County that can be answered with this data set 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 data sets you would need to answer that question.

I would create a plot of house sale price and neighborhood to see if there is a potential pattern of generating profits in certain neighborhood. I would plot the sale price of houses to single family vs. that to the multi-family and capture the difference with a price model.

1.2 Question 2

1.2.1 Part 1

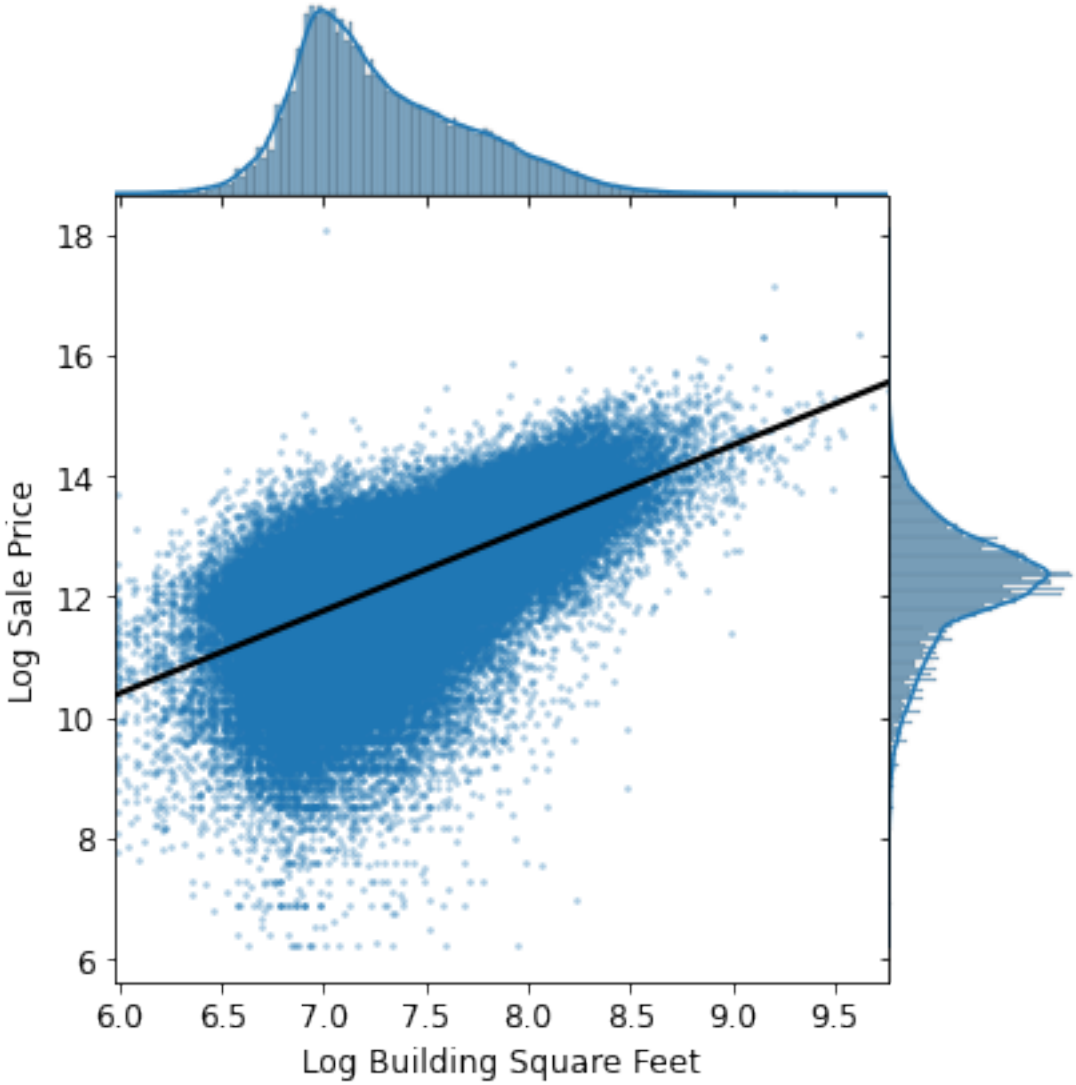
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The scale of the visualization is not helping understanding the data due to the outliers. It is too spread out due to the extreme outliers. We can solve this with regularization like taking the log.

1.2.2 Part 3

As shown below, we created a joint plot 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, does there exist a correlation between `Log Sale Price` and `Log Building Square Feet`? Would `Log Building Square Feet` make a good candidate as one of the features for our model?



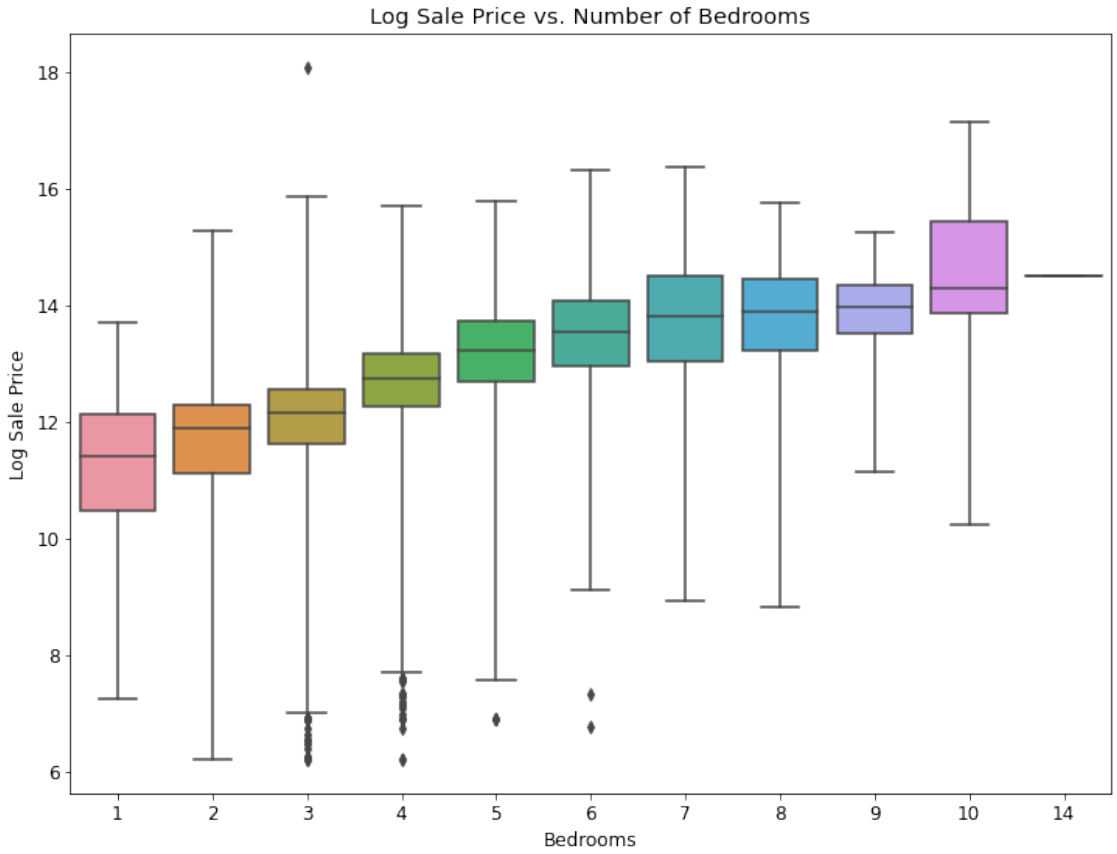
There looks like to be a strong correlation between 2 variables, so yes, it is a good candidate.

1.2.3 Part 3

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 [34]: sns.boxplot(x='Bedrooms', y='Log Sale Price', data=training_data, whis=5)
         plt.title('Log Sale Price vs. Number of Bedrooms');
```



1.2.4 Part 3

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?

most of them display a similar price pattern as compared to their neighbors, but it is noticable that 12 and 120 is a bit higher/lower than the average log price and needs to be noted. also it is clear that some of the neighborhoods are much more popular than the other ones like 30, which takes up most of the top 20 data.