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 appears to record the variables for one house in the county. Therefore, the granularity is one house.

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.

This housing survey could've been collected by the Census Bureau for determining the price of a house (perhaps for auctions), calculating tax rates, and analyzing living qualities or other demographic-related rights for different communities.

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.

Property Class is an example of a demographic-related variable since it gives information about how the house is used and even what type of people are inhabiting it. For example, values beginning with 2 or 3 means that the house is for residential or multi-family use while 7 means for commercial use. This gives information about what type of demographic we can expect to see in this house. For example, 205 is for “Two-or-more story residence, over 62 years of age up to 2,200 square feet”, which tells us its residents are elderly people.

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 like to understand the distribution of the ages of the house. I would do so by creating a histogram of the column “Age Decade” using plt.hist() and calculating the summary statistics for using Series.describe().

I would also like to investigate what percentage of houses have a garage. I would do so by calculating the number of houses with and without a garage using the “Garage Indicator” column and Series.value_counts() and dividing the former by the sum of the values.

1.2 Question 2

1.2.1 Part 1

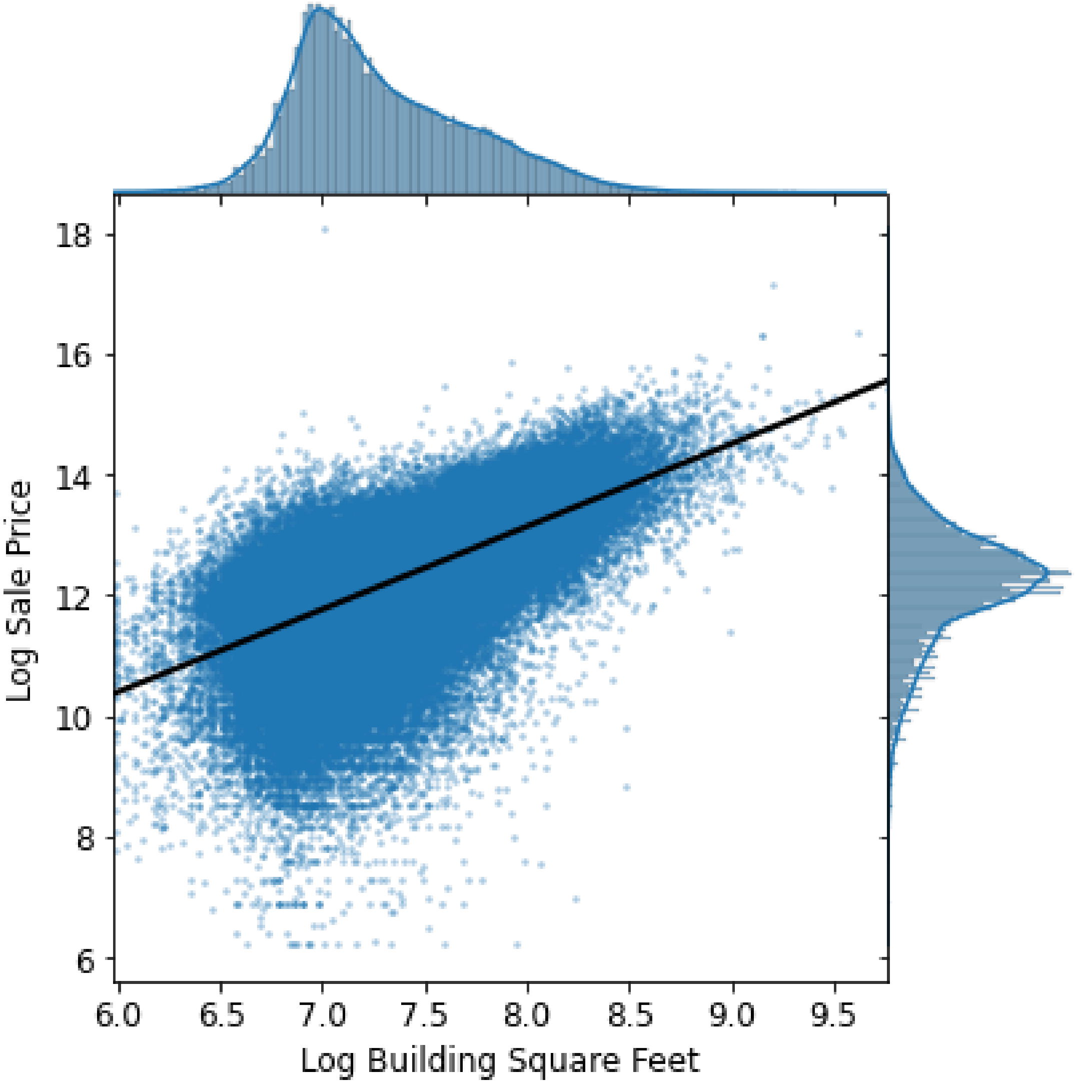
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The issue with this visualization is that the majority of the data (i.e. the part we're are probably most interested in) only takes up a small portion of the plot. The reason for this is because the distribution is right skewed, which means that there exists very few houses that are very expensive, while the majority of the houses are much less expensive in comparison. One way to fix the visualization is to remove outliers and plot the remaining data while including the outliers in the footnotes/comment.

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?



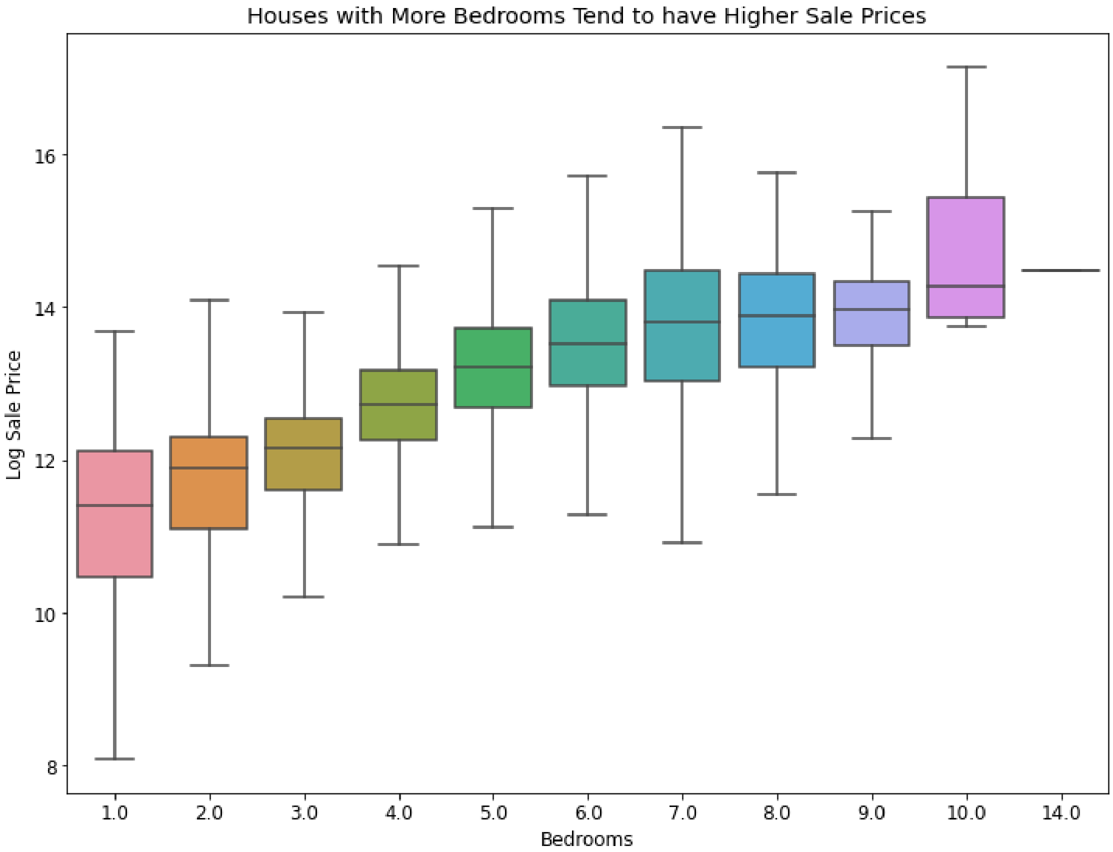
It appears that after taking the log of both variables, there is a linear correlation between the two. As the log(building square feet) increases, log(sale price) also tend to increase, and vice versa. However, we might want to consider using log(building square feet) as a feature for our model. It would be a good selection if we were predicting log(sale price). However, if we are trying to predict sale price alone, it is not guaranteed that their relationship is linear and the feature may not be good for a simple multivariate linear model.

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 [26]: sns.boxplot(x = training_data["Bedrooms"], y = training_data["Log Sale Price"], showfliers = Fa
         plt.title("Houses with More Bedrooms Tend to have Higher Sale Prices");
```



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?

Since neighborhood code is a qualitative nominal variable, there is no linear correlation between it and Log Sale Price. The median log prices for the top 20 neighborhoods appear to be quite similar, all being centered around 12. Their IQRs vary little to moderately as certain neighborhood have a large IQR (e.g. 120) and some small (e.g. 380). The same can be said about the min, max, and outliers of the neighborhoods as they all vary moderately with no apparent correlation witht he neighbothood code per se. However, log prices of different neighborhoods all seem to be right skewed.