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 represents a singualar property, there are columns for when the property was sold and how old it is showing that it is talking about individual transactions that took place to buy the property.

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.

The data was most likely collected by some housing company as there are many details about the layout of the house. This data is useful for sellers so they can inform buyers on the details of the house including things that can increase its value such as the square footage, number of fireplaces, and site desireability.

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.

One piece of demographic information can be seen in the deed number column which is the proof that is used to show that a person owns some property. Linking the deed number to a data base of ownership can easily reveal more information about the owner of the property.

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.

One question that may be interesting would be the exact affect that square footage has on the sale price of the house. To do this a scatter plot with the axes being the sale price and the other one being square footage would help to identify some patterns. Another question that would be worth looking into is how house sale price has changed over time and compare it to the national average. This would be a line plot that looks at the change in sale price as a percentage so that it can be compared to national percent changes in price and this would give us a good idea if houses here were rising at a higher or lower rate than the rest of the county.

1.2 Question 2

1.2.1 Part 1

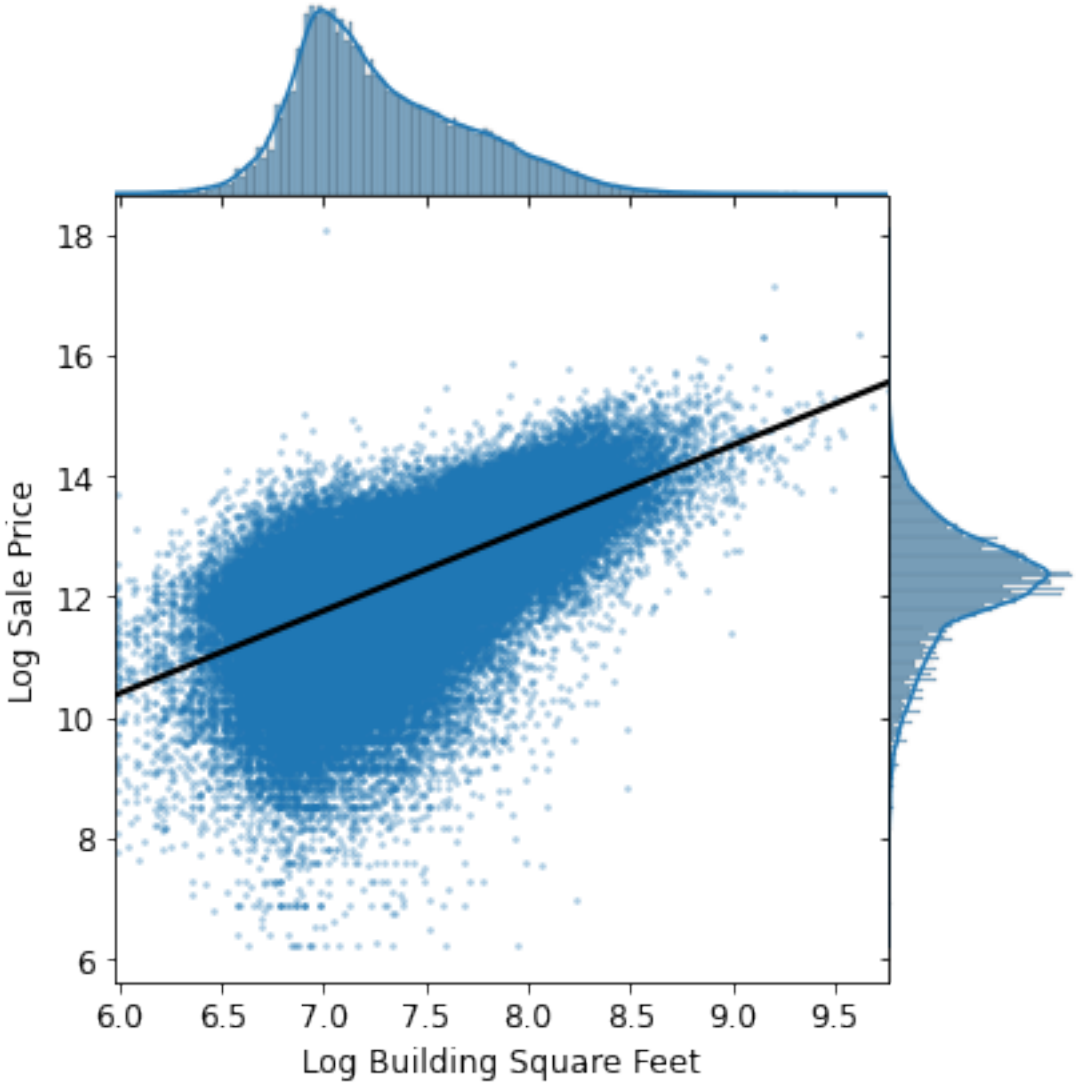
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The largest problem with the visualization is that all the data is so scrunched up that it makes it hard to read. One fix to make it at least readable would be to make the x-axis much smaller so that the visual is larger and its easier to find any points on interest on the visualizations.

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?



Because the joint plot of the two has a deccently close density it is safe to say that there is some correlation

between the two variables. While not perfect it would be fine to use this as one of our features since they are generally clusted togeather.

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 [ ]:
```
#sns.kdeplot(x=training_data['Bedrooms'], y=training_data['Log Sale Price'])
#sns.jointplot(data = training_data,x='Bedrooms', y='Log Sale Price')
#plt.scatter(x=training_data['Bedrooms'], y=training_data['Log Sale Price'])
sns.boxplot(x=training_data['Bedrooms'], y=training_data['Log Sale Price'])
plt.title('Box Plot Comparison of Bedrooms vs Log Sale Price');
```

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?

There seems to be a relationship is that if you have a lot of houses in your neighborhood than you will be close to the median of the log sale price.