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 each one property, which contains the information about some details about this specific property. For example, the class, land area, wall material, etc.

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 data could be collected by the data analysts working for real estate company in order to decide some details about their prospective property project. They might need to use the data to decide the best location, lot size, basement, wall material for their next property project.

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 demographic variable in our data is the "Census Tract" which means the the size of the neighborhood (number of people) where the property is located. It is the most relevant variable to demographic data (data on people). The nature of the demographic data it embeds is the number of people living in that neighborhood which this property is located, and it's quantitatively discrete.

Another demographic variable will be "Modeling Group" because it contains the information about the people living in this property. The nature of this demographic data will be the structure of the people living in this property (whether single family or multiple), and it's categorical nomial.

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 create a scatter point of column "Sale Price" and column "Land Square Feet" to see verify the sale price of the property is linear related to the land square feet.

I also would like to calculate the mean sale price of property with different wall materials using the information in 'Wall Material' to see the how much the wall material influence the sale price overally.

1.2 Question 2

1.2.1 Part 1

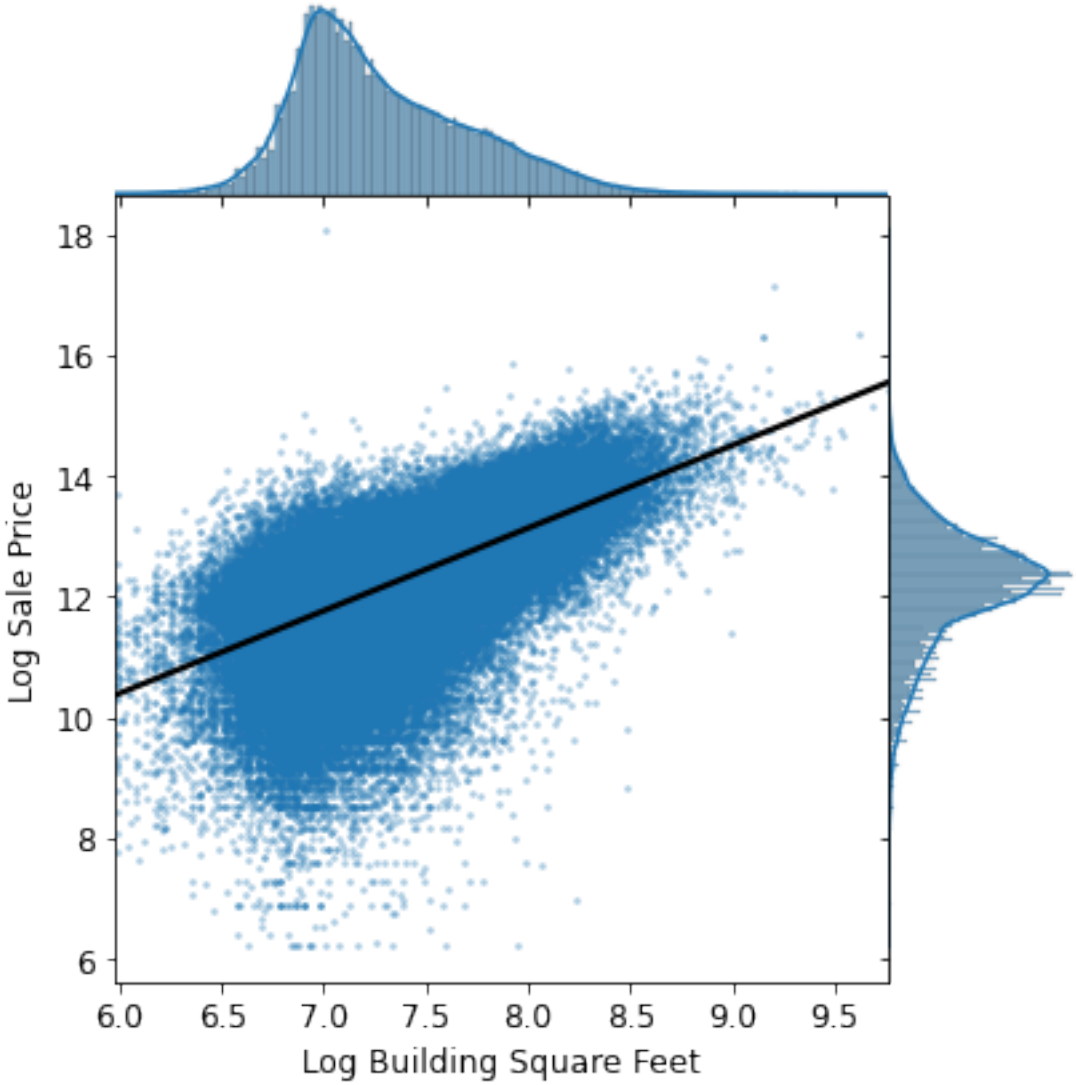
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

Issue: The box plot is way too condensed: it's hard to see the distribution where the major points lie. This happens due to the reason that some extreme outliers making the plot scale very long, as we can see the max value is about 200 times than the mean value. One wat to overcome it, we could remove the outliers to focus more on the major points. Or we could do some log transformation to make the spread to be smaller.

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 exists a positive correlation (correlation coefficient > 0) between `Log Sale Price` and `Log Building`

`Square Feet` as we can see a clear upward trend in the graph. Hence, the `Log Sale Price` could be included as a good candidate as one of the feature for our model.

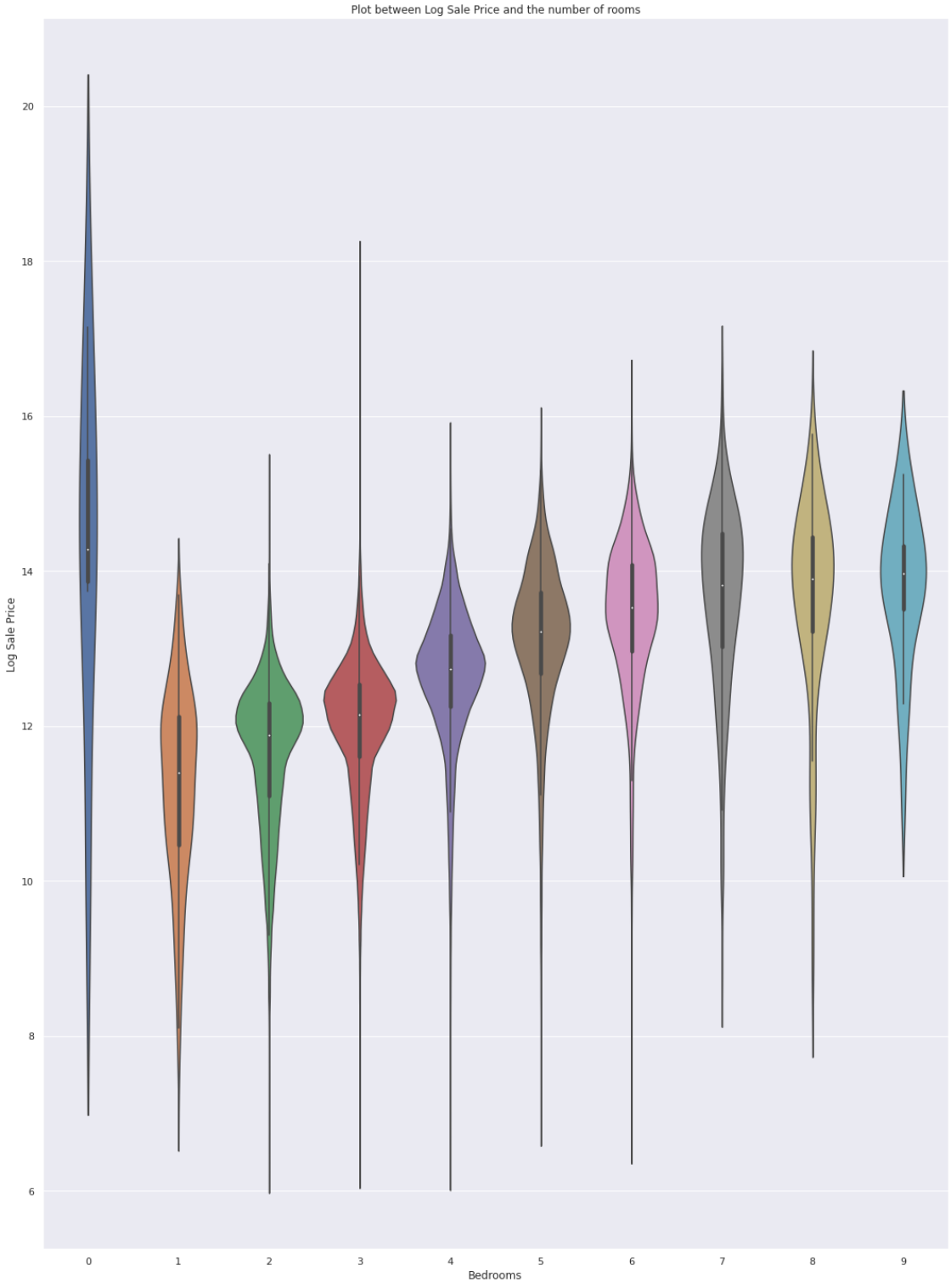
1.2.3 Part 3

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.set(rc={"figure.figsize":(16, 18)})
         sns.catplot(x='Bedrooms', y='Log Sale Price', data=training_data, height=20, aspect=.75, order=
         plt.title("Plot between Log Sale Price and the number of rooms")
```

```
Out[26]: Text(0.5, 1.0, 'Plot between Log Sale Price and the number of rooms')
```


Plot between Log Sale Price and the number of rooms
Log Sale Price
Bedrooms
20
18
16
14
12
10
8
6
0
1
2
3
4
5
6
7
8
9

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?

Log Sale Price is not related to their neighborhoods. As we can see from the violin graph, different neighborhoods has different price distribution(their means, median, and stds are all different). I don't find specific patterns among different neighborhoods