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?

Rows represent different features that need to be considered for purchasing a 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 data could be collected by some real estate agents to investigate the relationship between features of proporty and the price of it.

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.

Town and Neighborhood: People may consider the local ethic issue when considering buying house.

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.

How is the proproty class affecting sale price? I would like to create a bar plot of porproty class and mean of sale price;

Is there a relationahip between land square and sale price? I would like to create a line plot of land square feet and sale price.

1.2 Question 2

1.2.1 Part 1

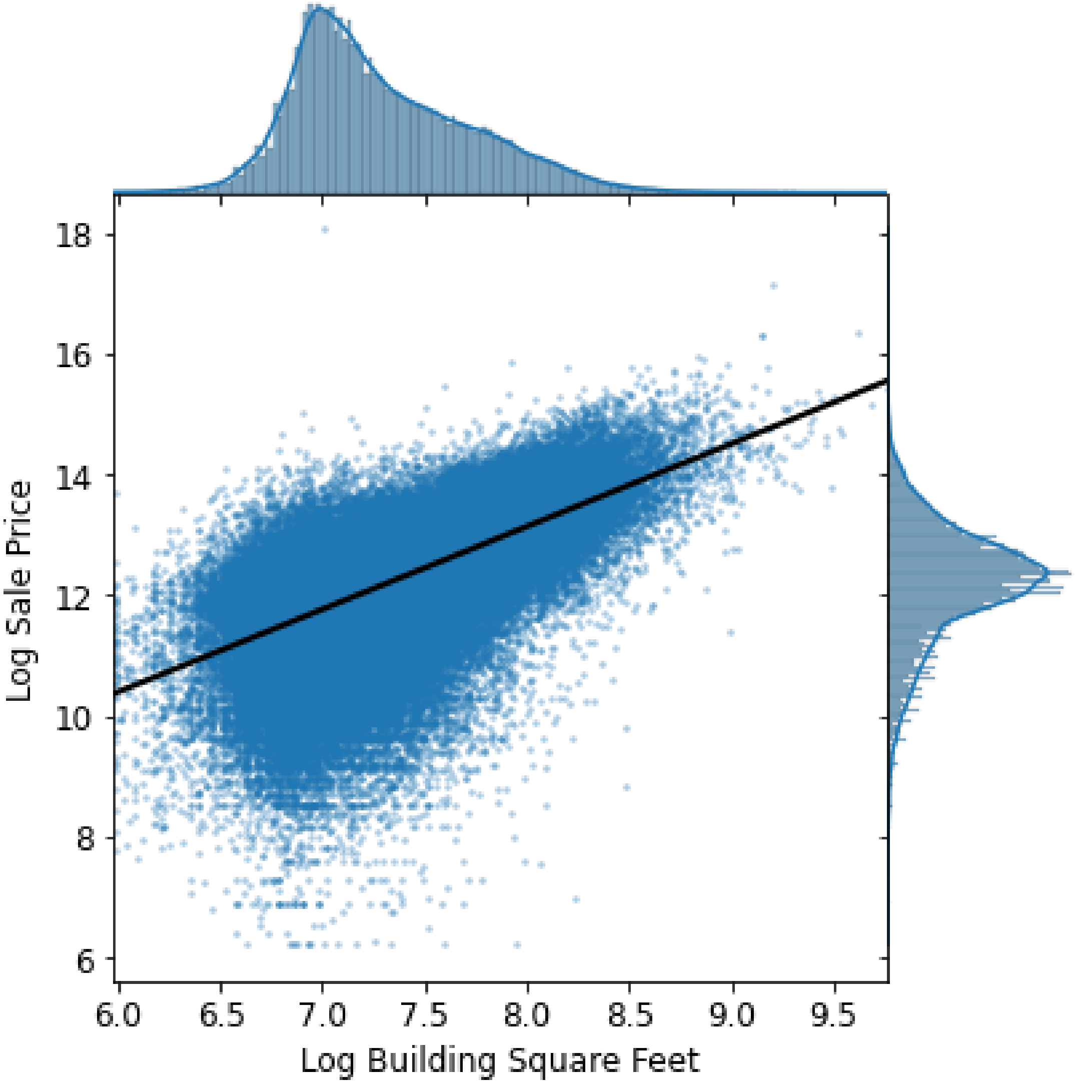
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

The sale price includes some unresonable data like price = 1, which would make the real distribution unclear. I think one possible way to deal with it is to do some data cleaning like cutoff price below $10000

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?



This plot shows a positive correlation between ‘Log Sale Price’ and ‘Log Building Square Feet’. Although

‘Log Building Square Feet’ may not be able to perfectly predict ‘Log Sale Price’, the clear correlation makes it a good candidate as one of the features for our model.

1.2.3 Part 3

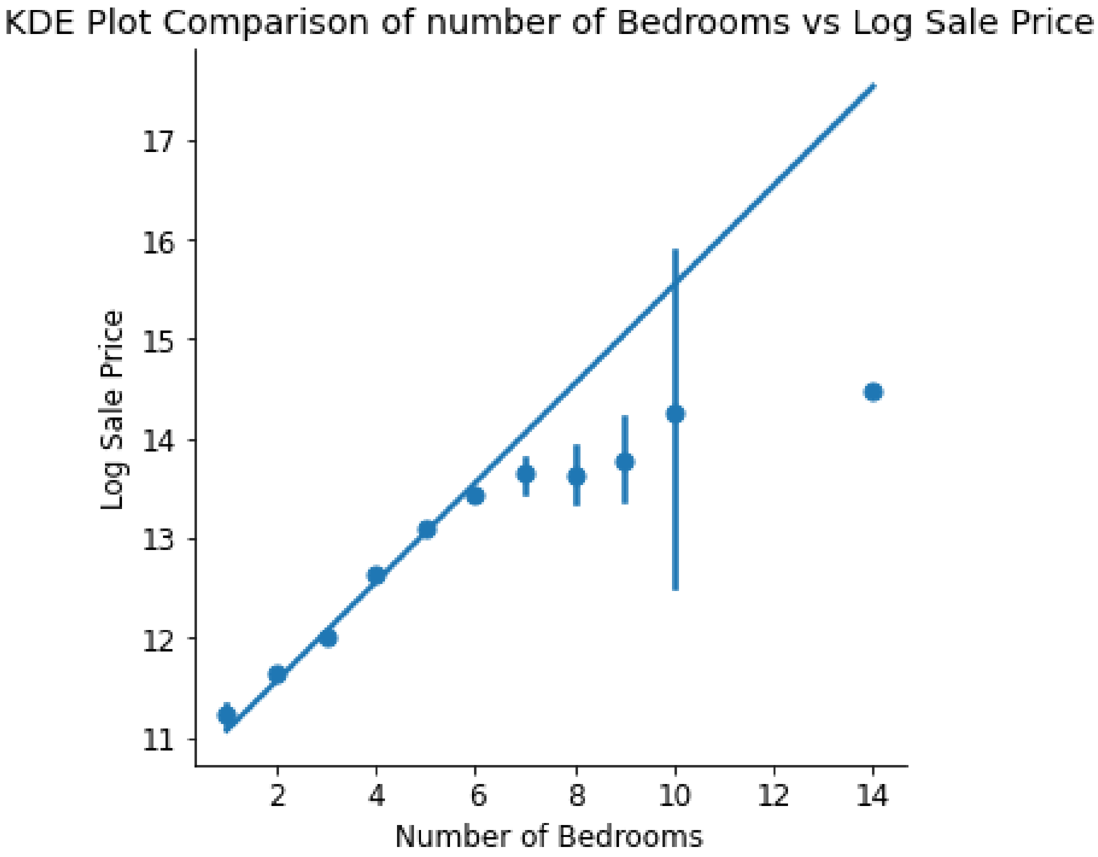
Create a visualization that clearly and succintly shows if there exists an association between `Bedrooms` and `Log Sale Price`. A good visualization should satisfy the following requirements: - It should avoid overplotting. - It should have clearly labeled axes and succinct title. - It should convey 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 [24]: plt.figure(figsize=(12,8))
         # sns.kdeplot(x = training_data['Bedrooms'], y = training_data['Log Sale Price'])
         sns.lmplot(x = 'Bedrooms', y = 'Log Sale Price', data = training_data, x_estimator=np.mean)
         plt.title('KDE Plot Comparison of number of Bedrooms vs Log Sale Price')
         plt.xlabel('Number of Bedrooms')
         plt.ylabel('Log Sale Price')
```

```
Out[24]: Text(29.71, 0.5, 'Log Sale Price')
```

```
<Figure size 864x576 with 0 Axes>
```


KDE Plot Comparison of number of Bedrooms vs Log Sale Price
Log Sale Price
17
16
15
14
13
12
11
2
4
6
8
10
12
14
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?

Generally differernt neighborhoods have about the same range of middle 50% distribution. So it is reasonable to use Neighborhood as a parameter in regression and prediction. But for Neighborhood 120, the range is outstandingly large, whcih may cause serious errors.