0.1 Question 1a

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

Each row represents all information of a house in Cook County, Illinois. The granularity of this dataset is property.

0.2 Question 1b

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.

I think the most likely to collect this data is the government. They could use it for urban planning, economic analysis, social research, policy development and ect.. The other possiable is some big company like real estate or housing sales company. They could use it for market analysis, customer targeting, stratigy plaing and ect..

0.3 Question 1c

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

It could be Town Code, Latitude, Longitude, and Neighborhood Code. Because in Census data, there very likely have information about demographic data, and we can join thoes together. Census tract is directly connect with Census Bureau.

0.4 Question 1d

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

1 What is average Land Square Feet in each town? I would calculate the average for Land Square Feet and aggragate town code. 2 Which property class is most popular? I would make a count graph or histgraph to show the weight of each property class and tell which porperty class is belongs to majority population.

0.5 Question 2a

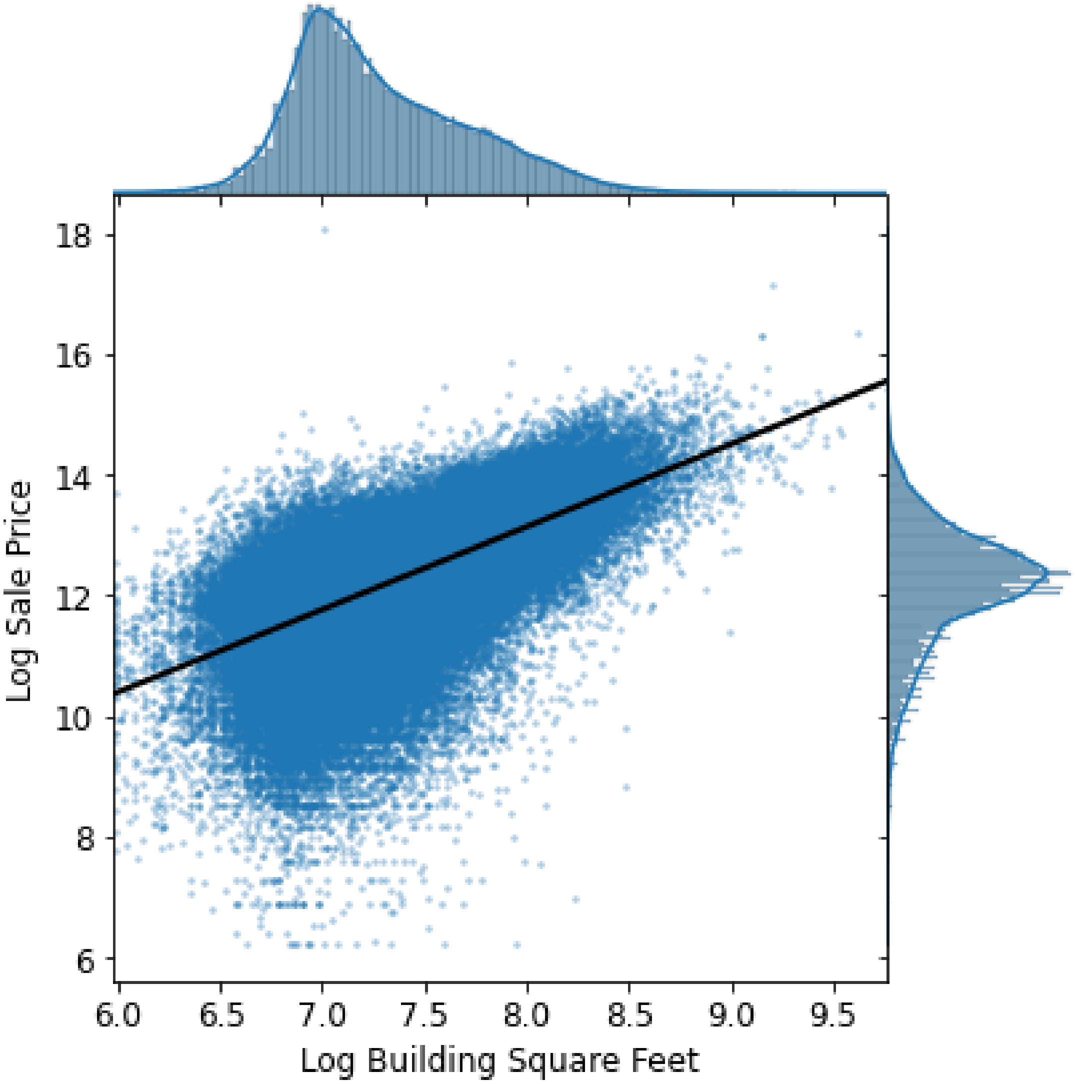
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

There is a extramly large outlier which is 7.100000e+07. This make the graph strach far away to right. We could use log alrithoms to renew unit of values or remove this outlier, in order to overcome this issue.

0.6 Question 3c

As shown below, we created a `jointplot` 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, would `Log Building Square Feet` make a good candidate as one of the features for our model? Why or why not?



It could be a good candidate as one of the features for our model, because it shows a strong relationship.

We should care about the outliers, becasue there are alot of them from where ‘Log Building Square Feet’ equal to 6.0-8.5.

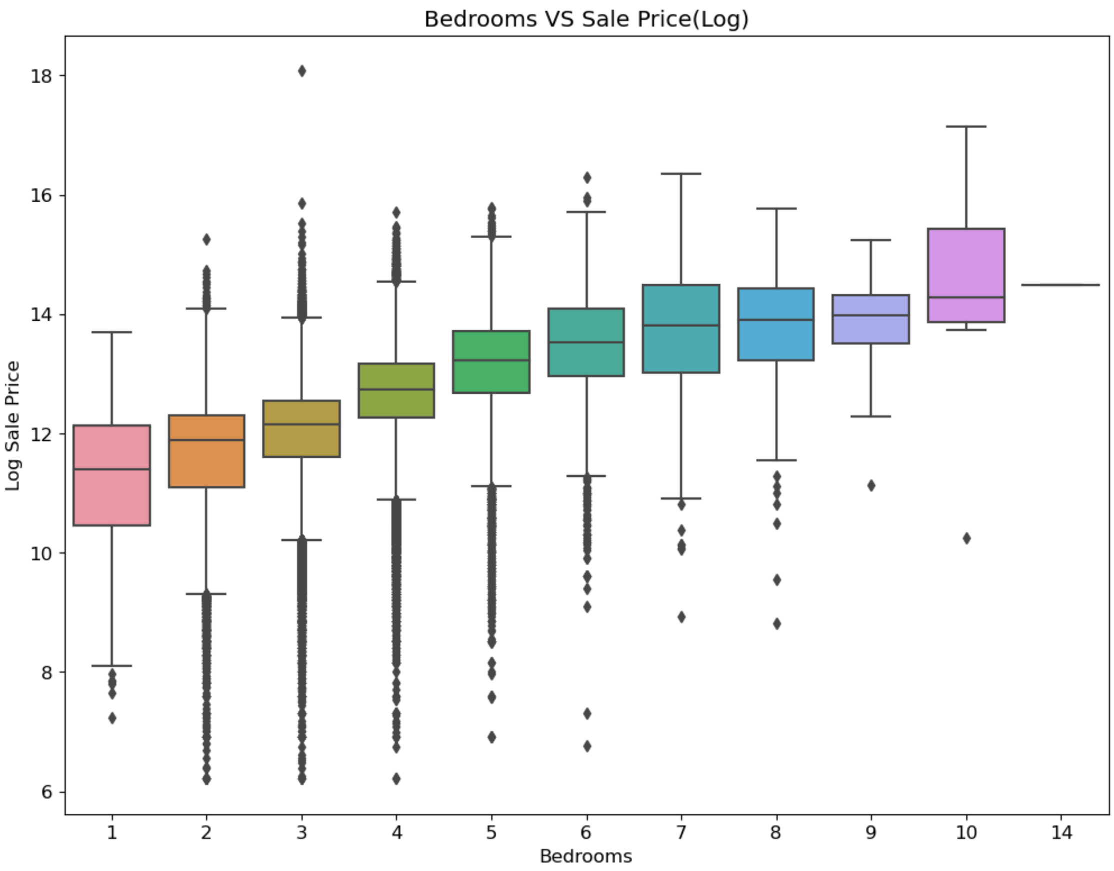
0.7 Question 5c

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 [57]: sns.boxplot(data = training_data, x = 'Bedrooms', y = 'Log Sale Price')
         plt.title('Bedrooms VS Sale Price(Log)')
```

```
Out[57]: Text(0.5, 1.0, 'Bedrooms VS Sale Price(Log)')
```



0.8 Question 6c

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? Is there a relationship?

There is no relationship between Log Sale Price and their neighborhoods, because it is hard to tell a patern in graph above.