0.1 Question 0

In what ways were the CCAO's property assessments discriminatory in late 2017?

The CCAO's property assessments were racially and ethnically discriminatory. High priced homes were undervalued while lower-priced homes were overvalued. This meant that wealthy homes which were high priced were owned by wealthy homeowners, who paid less in property taxes, usually white, while non-white homeowners who were working class had to pay more property taxes. The properties were fundamentally valued by race,class, and not a transperent way to value homes.

1 Part 1: Contextualizing the Data

Let's try to understand the background of our dataset before diving into a full-scale analysis.

Prior to the open data initiative and even prior to the assessment modeling initiative, Cook County's assessor office received much of their data for assessments from their relationships with "local elected officials, community leaders, real estate professionals and other citizens knowledgeable about real estate in the area." Because CCAO field inspectors cannot enter homes to gather data, this information must be gathered through either curbside observations or real estate records.

You can read more about data collection in the CCAO's Residential Data Integrity Preliminary Report.

1.1 Question 1

1.1.1 Part 1

Take a moment to assess the granularity of this dataset. What does each row represent?

Each row represents the features of the house. The data is gathered to assess the value of a house, where data collected yields different household values.

1.1.2 Part 2

Name a feature that isn't listed in this dataset but may be useful for predicting sales values. What insights could this feature provide? How might it increase or decrease a home's sales value?

Two features that I don't see on our dataset that could be added are crime rates and school. Crime rates would be helpful in providing us information on whether a neighborhood is safe or not, and thus will determine disirability for families looking to determine values of homes. Homes in safer neighborhoods will be valued greater than homes with large crimes.Similiar to crime, the quality of a school will provide families information or whether they want to raise their children in high education, safe schools, and quality education, which will further success and money for individuals. Thus a home's place relative to a school will value the home's value greater than schools with poor education systems.

1.1.3 Part 3

Let's take a look at the `Site Desirability` column. What do the column's values represent? Does the codebook provide sufficient guidelines as to how a property's `Site Desirability` is determined? Why or why not?

Site Desirability is a numerical value that values the site based on location. The guidebook states that this column is not appriorate for the purpose of property validation. The guidebook does not provide any guidelines on how site desirability is depicted, and instead encodes a value of 2.0 on whether or not the property is desirable. It seems heavily based on individual interpretation, and it seems to be highly generous given that the majority if not all of the dataset was given a 2.0, indicating site desirability was ranked very lax.

1.1.4 Part 4

Beyond a home's internal characteristics (such as number of rooms, bathrooms, etc.), describe a factor that might influence whether a home is desirable, and elaborate why. Think from the perspective of a real estate agent - i.e. what would an agent market to potential buyers and why?

Hint: Consider writing about characteristics related to a home's location and its proximity to other places.

As stated above, we can include factors such as schools or crime rates. Schools with higher education will push families to value that home greater because quality education promotes better success. Crime rates will determine a homes value, where homes located in high crime areas will be valued less, since a family would feel unsafe and thus wouldn't consider buying the home, while cities with less crime are more desirable to families. We could also include a boolean characteristics that determines whether or not the home is in a gated community, since gated communities are safer they will yield greater home value. Local amenities are also a less common but still desirable characteristics, where families expect there to be grocery stores, resturants and transporation within 5-10 minutes. Families are interested in making use of their time, having those amenities close to a family can potentially entice buyers.

1.2 Question 2

1.2.1 Part 1

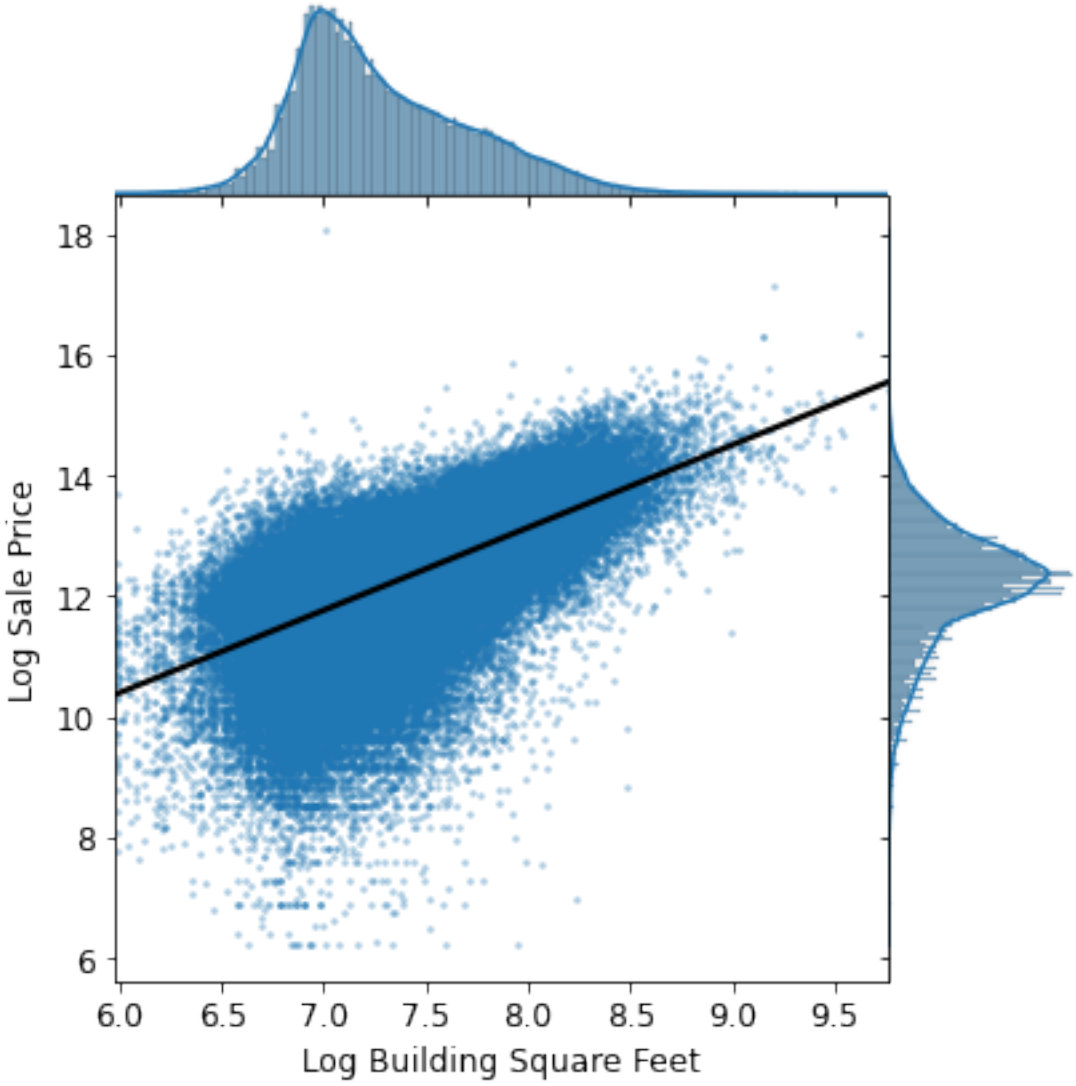
Identify one issue with the visualization above and briefly describe one way to overcome it. You may also want to try running `training_data['Sale Price'].describe()` in a different cell to see some specific summary statistics on the distribution of the target variable. Make sure to delete the cell afterwards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

One thing wrong with the data is that it is extremely small, due to the fact that the data is small compared to the size of the chart. Looking at the statistics, we can start by removing the outliers, or reduce the range of our x axis. The size of our chart will be the same, but our distribution will further align with the size of the chart.

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?



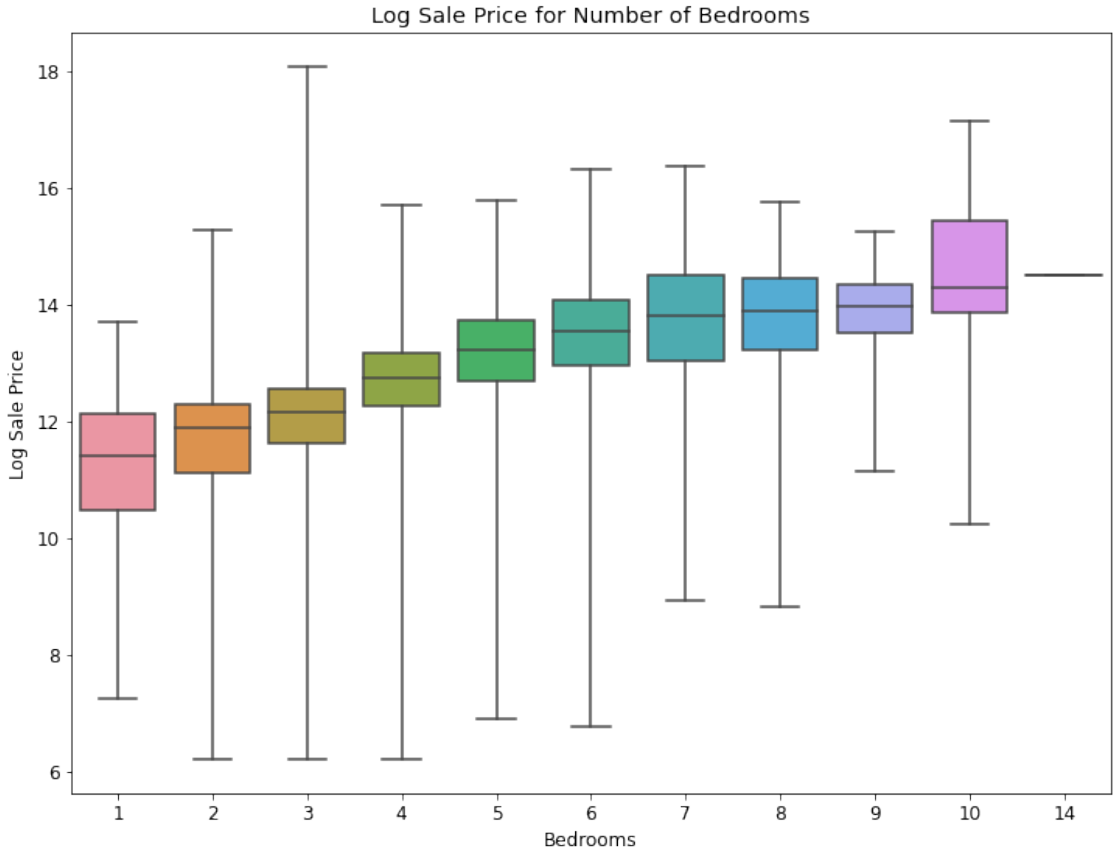
There is most certainly a relationship here. We can see good correlation between our two variables. The positive slope of the line indicates the correlation is positive, and because y seems to increase with increases in x Log Building Square Feet would make a good feature for our model.

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 [23]:
```
#Source https://stats.stackexchange.com/questions/60856/what-is-the-best-way-to-visualize-rela
#Boxplot best for discrete x, continous y
#Order by bedroom
#Whis is to avoid overplotting, whis clearly shows values
sns.boxplot(x='Bedrooms',y='Log Sale Price',data=training_data,order=["1","2","3","4","5","6",
plt.title("Log Sale Price for 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?

There is variation between Log Sale price and neighborhoods based on our above boxplot we can see from the interquartile range there is little variability with the residuals being mostly in the range of 12-14. When looking at the whiskers however, we determine the remaining residuals are larger, and looking at the plot we see a large amount of data at the ends of the whiskers. For example, looking at neighborhood code 10, we see at the bottom of the whisker values range from 5-11. Looking at the below plot we can see the data isnt uniformly distributed. The probability of each neighborhood code is not equally likely and we can see that houses in neighborhood code 30 make up a count of 8000 while houses in neighborhood code 71 only count for 2430. Thats almost a 4x difference in the amount of houses between both neighborhoods as an example. Certain neighborhoods contribute a greater portion in our training data than others, which is a problem.