0.1 Question 0

0.1.1 Question 0a

"How much is a house worth?" Who might be interested in an answer to this question? Please list at least three different parties (people or organizations) and state whether each one has an interest in seeing the value be high or low.

People who are looking to buy a house in Cook County woould be interested in this question and they would want the value to be low. current Cook County house owners looking to sell their house in Cook County would want to know if the price is high so they can make more money. The county economic department would also want to know the price of a house because higher house prices would mean that there is more economic activity in the county and more people want to move into the county and do business there.

0.1.2 Question 0b

Which of the following scenarios strike you as unfair and why? You can choose more than one. There is no single right answer but you must explain your reasoning.

A. A homeowner whose home is assessed at a higher price than it would sell for.
B. A homeowner whose home is assessed at a lower price than it would sell for.
C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive properties.
D. An assessment process that systematically undervalues inexpensive properties and overvalues expensive properties.

A - is unfair to future home buyers because they are forced to pay more than what they should be B - is unfair because home owners are losing money because their asset is being undervalued and if the assessed price is lower than market value then that means that C - is unfair because this can lead to those living in historically poorer communities to be subjecte to higher property taxes while the historically richer communities get away with paying less property tax.

0.1.3 Question 0d

What were the central problems with the earlier property tax system in Cook County as reported by the Chicago Tribune ? And what were the primary causes of these problems? (Note: in addition to reading the paragraph above you will need to watch the lecture to answer this question)

The main problems with the earlier property tax system in Cook County was that race and gender played heavy roles in determining the level of taxation in various areas of Cook County. Even though the assessor's office offered appeals, they were not accessible to everyone because people who were finanially more well off could more easily afford lawyers and the other necessary resouruces to make appeals. Redlining was also a key problem which was driven by race, because neighborhoods that had people of races that were considered to be more "risky" which basically meant non-white would end up being barred from receiving federally-backed mortgages.

0.1.4 Question 0e

In addition to being regressive, why did the property tax system in Cook County place a disproportionate tax burden on non-white property owners?

The property tax system in Cook County placed a disproportionately high tax burden on non-white property owners because non-white property owners were discriminated against when it came ti financial resources and the valuation of homes were also divided along racial lines, overvaluing lower value properties so they would have to pay more income tax.

0.2 Question 2

Without running any calculation or code, complete the following statement by filling in the blank with one of the comparators below:

$$\geq$$
$$\leq$$
$$=$$

Suppose we quantify the loss on our linear models using MSE (Mean Squared Error). Consider the training loss of the 1st model and the training loss of the 2nd model. We are guaranteed that:

Training Loss of the 1st Model_____Training Loss of the 2nd Model

0.3 Question 6

Let's compare the actual parameters (θ_0 and θ_1) from both of our models. As a quick reminder,

for the 1st model,

$$\text{Log Sale Price} = \theta_0 + \theta_1 \cdot (\text{Bedrooms})$$

for the 2nd model,

$$\text{Log Sale Price} = \theta_0 + \theta_1 \cdot (\text{Bedrooms}) + \theta_2 \cdot (\text{Log Building Square Feet})$$

Run the following cell and compare the values of θ_1 from both models. Why does θ_1 change from positive to negative when we introduce an additional feature in our 2nd model?

Theta1 changes from positive to negative when we introduce an additional feature in our 2nd model because we are adding an additional parameter through theta2 which does not have as strong of a correlation with the other feature that theta1 measures which causes a more negative correlation to compensate.

In [22]:
```
# Parameters from 1st model
theta0_m1 = linear_model_m1.intercept_
theta1_m1 = linear_model_m1.coef_[0]

# Parameters from 2nd model
theta0_m2 = linear_model_m2.intercept_
theta1_m2, theta2_m2 = linear_model_m2.coef_

print("1st Model\n 0: {}\n 1: {}".format(theta0_m1, theta1_m1))
print("2nd Model\n 0: {}\n 1: {}\n 2: {}".format(theta0_m2, theta1_m2, theta2_m2))
```

```
1st Model
 0: 10.571725401040084
 1: 0.4969197463141442
2nd Model
 0: 1.9339633173823696
 1: -0.030647249803554506
 2: 1.4170991378689644
```

0.4 Question 7

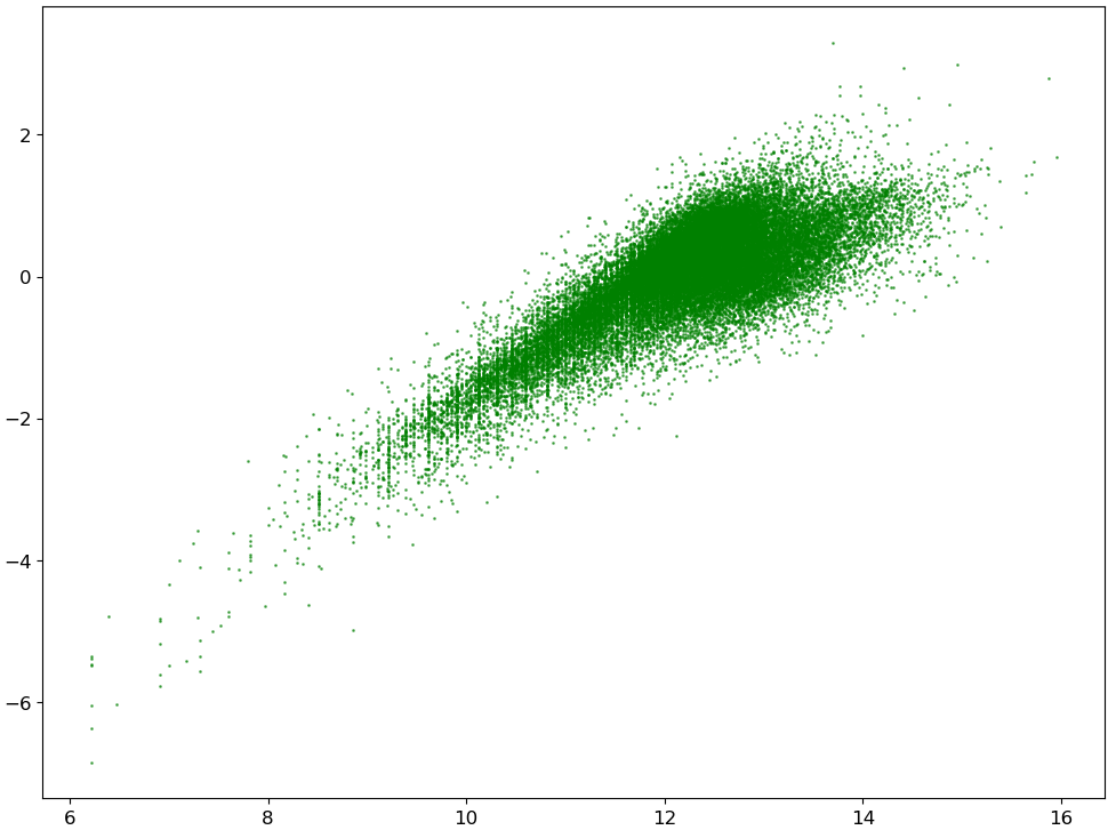
0.4.1 Question 7a

Another way of understanding the performance (and appropriateness) of a model is through a plot of the model the residuals versus the observations.

In the cell below, use `plt.scatter` to plot the residuals from predicting `Log Sale Price` using **only the 2nd model** against the original `Log Sale Price` for the **test data**. You should also ensure that the dot size and opacity in the scatter plot are set appropriately to reduce the impact of overplotting.

```
In [23]: x = y_test_m2
         y = y_test_m2 - y_predicted_m2
         plt.scatter(x, y, c = 'g', s = 1, alpha = 0.55)
```

```
Out[23]: <matplotlib.collections.PathCollection at 0x7f947bab6eb0>
```



0.5 Question 9

In building your model in question 8, what different models have you tried? What worked and what did not? Brief discuss your modeling process.

Note: We are looking for a single correct answer. Explain what you did in question 8 and you will get point.

When building my model, I worked through the different columns in the data and plotted them against the Log Sale Price using a scatter plot to see if there was a correlation. Then I removed the outliers for the Log Sale Price since that most of the features had some sort of correlation with the Log Sale Price. From there I tranformed the columns of the features that had some sort of relationship with Log Sale Price and selected those columns when running RMSE to get my predictions.

0.6 Question 10

When evaluating your model, we used root mean squared error. In the context of estimating the value of houses, what does error mean for an individual homeowner? How does it affect them in terms of property taxes?

In this context, when estimating the value of houses, error represents the difference between the actual price of the house at which it was sold and the predicted value given by the county office. In terms of property taxes, if the actual price is greater than the predicted value it means that the error is positive and the house was being undervalued whereas if the error is negative it means that the house is being overvalued. Undervaluation and overevaluation is bad because it can impose unfair tax rates on the homeowners since they are not being calculated based on market price.

In the case of the Cook County Assessor's Office, Chief Data Officer Rob Ross states that fair property tax rates are contingent on whether property values are assessed accurately - that they're valued at what they're worth, relative to properties with similar characteristics. This implies that having a more accurate model results in fairer assessments. The goal of the property assessment process for the CCAO, then, is to be as accurate as possible.

When the use of algorithms and statistical modeling has real-world consequences, we often refer to the idea of fairness as a measurement of how socially responsible our work is. But fairness is incredibly multifaceted: Is a fair model one that minimizes loss - one that generates accurate results? Is it one that utilizes "unbiased" data? Or is fairness a broader goal that takes historical contexts into account?

These approaches to fairness are not mutually exclusive. If we look beyond error functions and technical measures of accuracy, we'd not only consider *individual* cases of fairness, but also what fairness - and justice - means to marginalized communities on a broader scale. We'd ask: What does it mean when homes in predominantly Black and Hispanic communities in Cook County are consistently overvalued, resulting in proportionally higher property taxes? When the white neighborhoods in Cook County are consistently undervalued, resulting in proportionally lower property taxes?

Having "accurate" predictions doesn't necessarily address larger historical trends and inequities, and fairness in property assessments in taxes works beyond the CCAO's valuation model. Disassociating accurate predictions from a fair system is vital to approaching justice at multiple levels. Take Evanston, IL - a suburb in Cook County - as an example of housing equity beyond just improving a property valuation model: Their City Council members recently approved reparations for African American residents.

0.7 Question 11

In your own words, describe how you would define fairness in property assessments and taxes.

I think that fairness in the context of property assessments and taxes is difficult to achieve because there are multiple ways to interpret what should be considered "fair". For example, if a property, based on its objective features like square footage and health of building is really high but is in a historically underprivileged community, pricing it really high would be unfair towards the people living there. I think that fairness when defining property assessments and taxes should take into account both the physical properties of the house but also consider the economic status of the neighborhood in which the property is located.

0.8 Question 12

Take a look at the Residential Automated Valuation Model files under the Models subgroup in the CCAO's GitLab. Without directly looking at any code, do you feel that the documentation sufficiently explains how the residential valuation model works? Which part(s) of the documentation might be difficult for nontechnical audiences to understand?

I think throughout the documentation I found that there were a lot of acronyms (like SHAP, AWS, RMSE) that are not part of a non technical audience's immediate vocabulary. There were also technical words being thrown around like testing and training data which aren't obvious to nontechnical users to understand. So overall, it would make it a lot more digestable for nontechnical audiences if the documentation explained in further detail the key terms that are needed in order to make science of the modeling that takes place.