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 might be interested in seeing how much a house is worth would be a homeowner, a person who wants to buy a house, and a local government. 1. A homeowner who wants to sell their house would be interested in seeing a higher house worth as they would want to sell their property in the market at a higher price. 2. A person who wants to buy a house would want to see a lower value of property as they would want to buy their house at the most affordable value. 3. A local government tax assessors would want to be interested in the worth of a house as they would need it to determine property tax. While local government gain a part of their funding from property tax, they might want to see higher house price so that they can earn higher property tax. However, I don't think they have any interest in seeing a high or low value as they would want to assess it fairly and unbiased.

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.

I think part C. if an assessment overvalues inexpensive properties and undervalues expensive properties, it would be unfair if a tax assessor fails to value the properties justly as it can put tax burden on low income households as they might have to pay higher tax more than they can afford while lessening the burden for high income families who owns expensive properties. However, each scenario above can be unfair depending on the context. For instance, part A. would also be unfair to those who are buying the home if the home is assessed at a higher price than it should be selling for as they would be paying at higher price than they should be.

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 property tax system in Cook Country reported by the journalist of Chicago Tribune has been described to come from the long systemic bias and corruption as properties owned by low income households are overvalued and high income are undervalued which resulted in a regressive tax system. The ability to appeal property tax decision in Chicago creates inequity and places a disproportionate tax burden on non-white property. While it offers homeowners who feel like they have been taxed unfairly, the option to appeal the assessment decision and ask for a re-evaluation and lower property value. This might not be an option to all household as appealing the decision requires hiring an attorney which might be not a viable option to low income households as they can't afford one which creates the regressive tax system where high income homeowners can get a lower property tax. Morover, this system creates an unjust and unfairness to population who are at a disadvantaged as not everyone have equal access to the appeal board for instance people who are working 9-5 and aren't able to spare their time to go to the court, disabled people, etc.

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?

In addition to the the inequity of appealing the assessment decision as not everyone has the means to do so. The chicago's tax is also fixed, a zero sum, if a person is paying more than someone will have to pay less and thus if the white rich neighborhood's properties are undervalue they would pay less tax, while the non-white property would have to pay more which put a disproportionate tax burden on the non-white homeowners.

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

$$\geq$$

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?

The theta1 change from positive to negative after adding another feature is because it now has to take into account the weigh from the coefficient of Log Building Square Feet which might have a more significance in one unit increase to the Log Sale Price. It just signifies that the an increase in Log Builiding Square Feet would lead to a higher increase in Log Sale Price, but when an increase of number of bedroom with a fixed number in log building square feet, it would lead to a decrease log sale price.

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

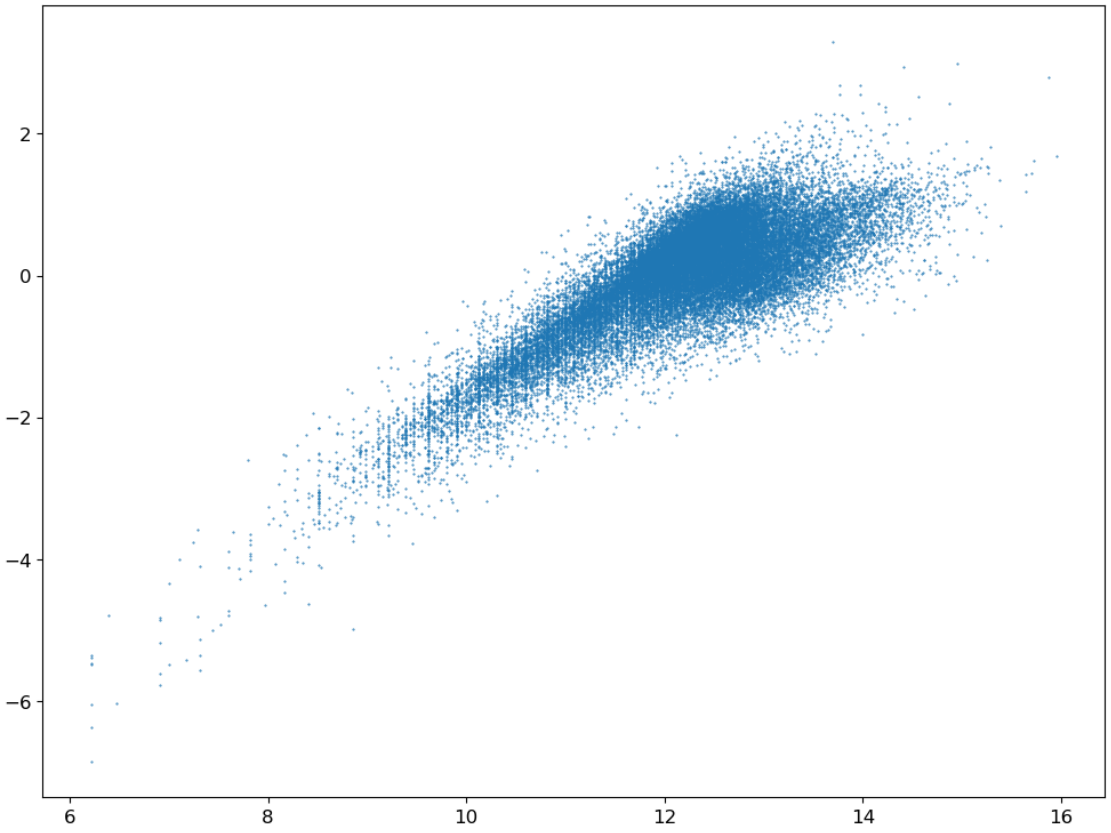
0.4.1 Question 7a

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]: plt.scatter(y_test_m2,y_test_m2-y_predicted_m2, s = 0.2)
```

```
Out[23]: <matplotlib.collections.PathCollection at 0x7f7fa1217820>
```



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.

I started with a simple model from above and from there I added more features as I go, but I also realized that removing outliers has a significant impact to the model's performance, figuring out the right boundary for removing outliers is really key in the project. However, I didn't find using the IQR method help me. Another thing that I found to be very useful was finding correlation between the variables and the sale price which helped me visualize what features have high correlation and which are the one I can exclude from the model. I also found that squaring the variables is extremely not useful, know how to normalize and linearize the model is very important.

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?

The rmse is the average difference between the predicted log of sale price from the actual value. The higher the error the higher the deviation of the predicted sale price from their actual value. Higher error puts them at risk of undervalue or overvalue of property which might lead to higher or lower than their actual property taxes.

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 a fair property assessments and taxes is an unbiased, transparent and accountable system of assessment that don't put another party at an unfair disadvantage. It should also not be determined on subjective attributes such as the homeowners' skin of color or occupation, but rather objectively based on the property's condition, age or land value.

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?

The documentation have a sufficient explanation on how the residential valuation model works for data scientist, however as they are written in technical terms that might not cater or be a sufficient explanation to audience who are not from data science background. Most of the part of the documentation are not accessible to public audience and might be difficult for nontechnical audiences to understand would be the model selection as not everyone would know what a training model is or what hyperparameter selection part are or what are train test validation model, thus these create a lack of transparency and accessibility in the documentation.