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.

Real estate agents may be interested in this question and seeing a high value to see if the property is profitable. Economists may be interested in seeing the value to be high in order to access if the housing market is in a good position, perhaps realitive to previous moments in time. Residents moving into Cook County may be interested in this question, and they may have an interest in seeing the value to be at least equal to or higher than the asking price of the home in hopes to find a valuable home for the right price.

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) seems unfair because if the home is more valuable than the selling price, the homeowners are not recieving the whole value of the property, however in the case that perhaps the housing market is down and bad for selling, the homeowners may have no choice in which it is fair.

B) also seems unfair since it is the definition of a regressive tax system. This causes lower income households to pay more property tax and wealthier households to pay less in property tax. Lower income households tend to be minority communities like Latino and Black communities, while wealthier households tend to be White communities, creating a racial divide and inequality.

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)

A regressive tax system came into place in Cook County in which poorer housholds, which were mostly comprised of Latino and Black communities, were paying a disproportionate burden of Cook County's tax cut. These Latino and Black communites with lower valued properties were paying a higher effective tax rate than White communities with more expensive properties. The CCAO's algorithm for estimating home values was not the main fault of this issue. Instead, the issue lied in the idea that wealthier families/households had a greater chance of challenging their annual tax assessments to a review board since they had greater time and resources to do so. Particularly a large recourse included hiring a tax lawyer to help appeal the tax assesment, which was a luxury made possible for wealthier households. Cook County effectively has a large tax attorney industry build around helping households lower tax assessments. Cook County has a fixed sum they collect in property taxes, so if wealthier, white homeowners are paying less in taxes, then a greater tax burden falls on lower income, minority homeowners. This issue then affects the machine learning model used to access home values. Wealthy homes have similar features that the model trains on which then effectively undervalues the price of other wealthy homes without having the owners to appeal. This systematic effect ripples out through the years and across the County.

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?

Similar to what I discussed in part 0d, Cook County collects a fixed sum of property taxes, so if white property owners are paying less taxes since they are appealing and altering the machine learning mode, that places a disproportionate tax burden on non-white property owners.

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?

Some question to brainstorm about: If you only know the # of bedroom, how much are you willing to pay if there are more bedroom? Now consider the scenario where we know the space of the house: if the two house have the same amount of space, will you be willing to pay more for the house with more number of bedroom or less number of bedroom and why? What is the relationship between space and # of bedrooms? What is the average space of each bedroom given the same fixed amount of space but different number of room? Is this negative number significant?

In the 1st model, without any additional information, it can be assumed that a greater number of bedrooms would result in a higher house sale price, thus the positive θ_1 coefficient. However, if you have additional information about the log building square feet we can think of the scenario in which the square feet of a building remains the same but more bedrooms would cause the price to decrease since the space would become more cramped and undesierable.

In [188]:
```
# 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

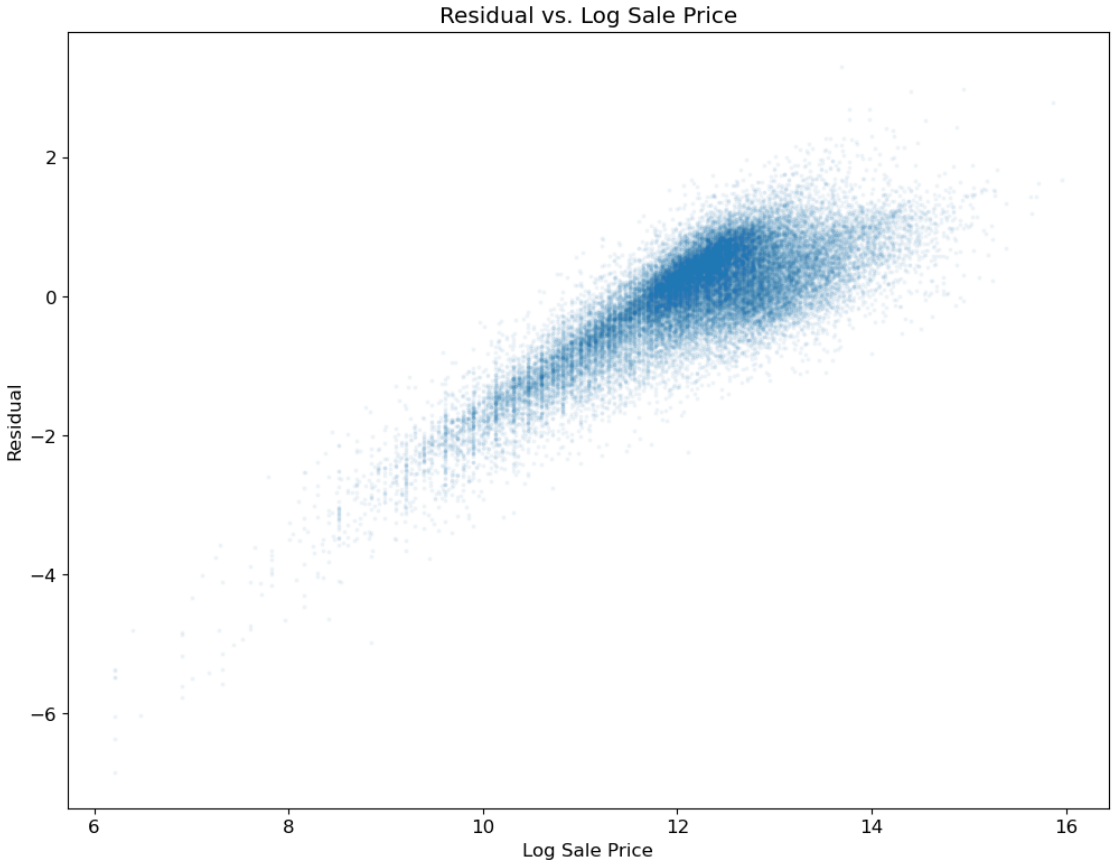
0.4.1 Question 7a

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 [189]: plt.scatter(y_test_m2, y_test_m2 - y_predicted_m2, s=3, alpha=0.05)
          plt.title('Residual vs. Log Sale Price')
          plt.xlabel('Log Sale Price')
          plt.ylabel('Residual')
```

```
Out[189]: Text(0, 0.5, 'Residual')
```



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.

In building the model for question 8, I would plot various graphs of sale price versus different columns I intuitively thought would have the greatest relationship with sale price with the help of codebook.txt. I would graph scatter plots and box plots to analyze the replationship and used box plots to visualize outliers in certain variables. I created a function, find_outlier(), to find the lower and upper bound for the column name passed in, following the dataframe it is pulling from. To create the pipeline, I refered to boxplots to visualize which columns had significant outliers to determine which columns to apply the remove_outliers function to. With those columns, I then utilized my find_outlier() function to determine the lower and upper bound parameters. I utlized scatterplot of sale price versus different columns to visualize which columns needed to be logged transformed. I also added a number of bedroom column with the add_total_bedrooms function inside the pipeline. I toggled with different combinations of columns that resulted in the best training RMSE. I also tried one hot encoding the roof material column, but found it only increased my training RMSE as well as the test RMSE, so I negated OHE of the roof material column. I created 2 pipelines, one for training and one for testing which were identical besides removing the remove_outlier functions for the test pipeline as well as any functions/columns dealing with the 'Sale Price' column.

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 in this model is the difference between the actual sale price (y) of the home and the CCAO's prediction($\hat{y}$); (y - $\hat{y}$). If the RMSE for an individual's home is positive that means the CCAO's prediction is lower than the actual sale price, so the house is being undervalued, in which the homeowners are paying less property taxes. Contrastly, if the RMSE is negative that means the CCAO's prediction is higher than the actual sale value, so the house is being overvalued, in which the homeowners are paying more 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.

Fairness would ideally be that the CCAO's prediction value of a home is equal to the actual sale value resulting in an RMSE of 0 and defined property taxes across all wealth levels. The CCAO's problem was that their model was built upon the wealthy's appeals to the board, so fairness would be to evaluate each house individually and adjusting the model so that it doesn't depend on other homes similar to it that have appealed to lower their tax.

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?

N0 I don't think the documentation is sufficient to explain how residential evaluation models work especially for nontechnical audiences. It was hard for someone like me with experience working around GitLab to even navigate the folders and files. The link names in Home.md are technical makin it confusing where to even start or where to look. The entire model is complex and I would suspect it would take hours for a nontechnical audience to even begin to understand the GitLab.