## 0.1 Question 1: Human Context and Ethics

In this part of the project, we will explore the human context of our housing dataset. **You should watch Lecture 15 before attempting this question.**

---

### 0.1.1 Question 1a

"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 housing price be low or high.**

1. The Cook's County Assessor's Office wants to understand the price of a house, but ideally should not care for how high or low the price is; instead, if the Assessor's Office is just and uncorrupt, they only care about the true accuracy of the house price in order to make fair assessments for the purpose of determing property tax.

2. A homeowner who is less well-off and perhaps lives in South Chicago, where property values are historically lower, would wish to see their housing price *assessed* as low. This is because a lower property value assessment would mean that their property tax would also decrease, allowing them more disposable income and also critically allowing them to keep their homes at the same/lesser budget as when they purchased it.

3. A homeowner who is more well-off and lives in a more affluent neighborhood in Chicago would wish to see their property value also *assessed* as lower, once again for property tax reasons. They could pay less property tax on their property that is perhaps undervalued in assessment.

4. Conversely, homeowners in both lower income as well as affluent neighborhoods would wish for their "house worth" to be seen as high as possible when *selling* their homes and listing them on the market. Obviously, they wish to sell their home for the highest amount possible, and therefore wish for their home to be seen as high value by potential buyers. The critical thing here is context, as a homeowner who wishes not to sell wants their home to be undervalued for property tax reasons, but a homeowner who wishes to sell their home wants their property to be valued as high as possible to profit as much as possible.

5. Corporations who own their buildings and likely don't wish to sell anytime in the near future want their property values to be undervalued in assessment, once again for property tax reasons. Thus they can utilize their corporate tax lawyers to challenge the assessment of the CCAO and lower their assessed property value, via the appeals board, in order lower the amount of property tax they pay each year.

### 0.1.2 Question 1b

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. Would you consider some of these scenarios more (or less) fair than others? Why?

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.

Option A is unfair for a homeowner who has a low income and only owns their singular home and does not want to sell. This is because the overvaluation of their home would lead to an increase in property tax they have to pay, and this may lead to catastrophic consequences such as them no longer being able to afford their mortgage (due to the fact that the property tax has increased in ways they did not budget for when they originally purchased their home).

Conversely, in a higher income neighborhood, where perhaps the homeowner owns multiple properties, a home being assessed at a higher price would still be unfair, but *less* unfair than for the low-income scenario. This is because the increase in property tax is more likely just to be an inconvenience to this homeowner, eating into their potential profits, but not catastrophically ruining them by forcing them to forclose on their only home. Thus, the impact/severity of the consequences of option A vary based on context, making its fairness also context dependent.

The most unfair scenario would thus be option C, for the reasons mentioned above. Not only would the lower-income individuals with more inexpensive properties be unfairly overcharged in property tax, potentially ruining them, the higher-income individuals would gain tax breaks and be able to gain more profits; this would kick those who are down while pushing the rich to even higher riches, deepening the wealth disparity.

Conversely, option D could be considered less unfair, as those who are in need of tax breaks and more disposable income would get those tax breaks. Those who are wealthy and can afford to pay additional taxes without any significant consequences would do so, and the zero net sum tax collection would still be met but with the least amount of catastrophic consequences for the whole population.

### 0.1.3   Question 1d

What were the central problems with the earlier property tax system in Cook County as reported by the Chicago Tribune? What were the primary causes of these problems?

**Note:** Along with reading the paragraph above, you will need to watch Lecture 15 to answer this question.

The earlier property tax system certainly had some inaccuracies in their model for predicting property value, leading to initial inaccuracies in property values, but another critical element was the property value appeals process. The appeals process was intended to provide a "human element" to the assessement process by allowing homeowners to file for an appeal if they believed their property was improperly evaluated. In theory, this appeal process would be equally accessible to everyone, low-income and affluent homeowners alike. However, in practice, virtually all homeowners who filed appeals were affluent individuals or corporations who had the access and funds to tax attorneys. The wealthy thus disproportionately filed appeals, and were the counties where the wealthy resided had vastly decreased numbers of homes that were overvalued by the CCAO.

Conversely, virtually no homeowners in Southern Chicago and other less affluent neighborhoods filed appeals for their properties, which correlated with a much higher number of homes being overvalued by the CCAO in those neighborhoods compared to the affluent neighborhoods. This lead to increased regressive taxation as less-affluent homeowners didn't file appeals and consistently had overvalued properties, while affluent homeowners and corporations continued to file appeals and had less overvalued properties. Thus, the effective tax rate (the percentage of the home's value that the individual is paying in property tax) was increased for those in poorer neighborhoods while simultaneously being decreased for richer homeowners.

### 0.1.4   Question 1e

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

The property tax system of Cook County was a "zero sum game" policy, meaning that the decrease in effective tax that was seen by wealthier homeowners was taken up by an increase in effective tax on the working class homeowners. Say that CCAO evaluates that all homes in Cook County are worth 100 billion, then each year 1 billion dollars of tax needs to be collected. The proportion of that 1 billion is split amongst all the homeowners of Cook's County. Thus, if a wealthier homeowner is paying less tax and contributing less to that 1 billion sum due to undervalued property, the working class homeowner must pick up the slack and pay more out to cover that loss in property tax. So, the decreased effective tax of the wealthier homeowners is compensated by an increased effective tax rate of non-white property owners, placing a disproportionate tax burden on the non-white homeowners as they are not only being paid less but also taxed a higher percentage of their lower wage.

## 0.2 Question 4a

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

In the cell below, use `plt.scatter` (documentation) to plot the residuals from predicting `Log Sale Price` using **only the second model** against the original `Log Sale Price` for the **validation data**. With such a large dataset, it is difficult to avoid overplotting entirely. You should also **ensure that the dot size and opacity in the scatter plot are set appropriately** to reduce the impact of overplotting as much as possible.

```
In [24]: Y_predicted_m2.shape
```

```
Out[24]: (33864,)
```

```
In [25]: Y_valid_m2.shape
```

```
Out[25]: (33864,)
```

```
In [26]: rmse(Y_predicted_m2, Y_valid_m2)
```

```
Out[26]: 0.8113963052434996
```

```
In [27]: plt.scatter(Y_valid_m2, Y_valid_m2 - Y_predicted_m2, alpha=0.3, s=8)
         plt.xlabel('True Log Sale Price \n Y')
         plt.ylabel('Residuals of Log Sale Price \n Y - $\hat{Y}$')
         plt.title('Log(Sale Price) vs Residuals of Log(Sale Price) for Validation Set')
         plt.axhline(y = 0, color='r', linestyle='--');
```

Log(Sale Price) vs Residuals of Log(Sale Price) for Validation Set
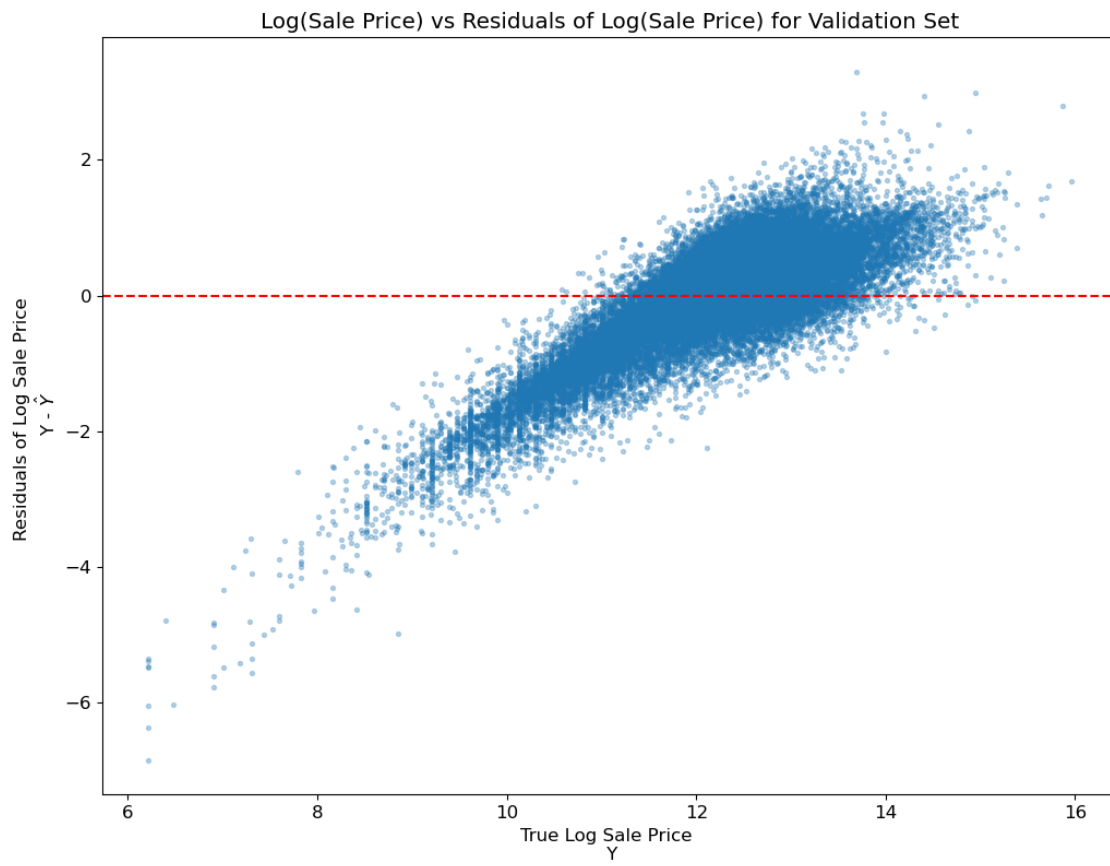
### 0.2.1 Question 6c

Now that you've defined these functions, let's put them to use and generate some interesting visualizations of how the RMSE and proportion of overestimated houses vary for different intervals.

```python
In [73]: # RMSE plot
         plt.figure(figsize = (8,5))
         plt.subplot(1, 2, 1)
         rmses = []
         for i in np.arange(8, 14, 0.5):
             rmses.append(rmse_interval(preds_df, i, i + 0.5))
         plt.bar(x = np.arange(8.25, 14.25, 0.5), height = rmses, edgecolor = 'black', width = 0.5)
         plt.title('RMSE Over Different Intervals\n of Log Sale Price', fontsize = 10)
         plt.xlabel('Log Sale Price')
         plt.yticks(fontsize = 10)
         plt.xticks(fontsize = 10)
         plt.ylabel('RMSE')

         # Overestimation plot
         plt.subplot(1, 2, 2)
         props = []
         for i in np.arange(8, 14, 0.5):
             props.append(prop_overest_interval(preds_df, i, i + 0.5) * 100)
         plt.bar(x = np.arange(8.25, 14.25, 0.5), height = props, edgecolor = 'black', width = 0.5)
         plt.title('Percentage of House Values Overestimated \nover different intervals of Log Sale Pri
         plt.xlabel('Log Sale Price')
         plt.yticks(fontsize = 10)
         plt.xticks(fontsize = 10)
         plt.ylabel('Percentage of House Values\n that were Overestimated (%)')

         plt.tight_layout()
         plt.show()
```

RMSE Over Different Intervals of Log Sale Price



Percentage of House Values Overestimated over different intervals of Log Sale Price
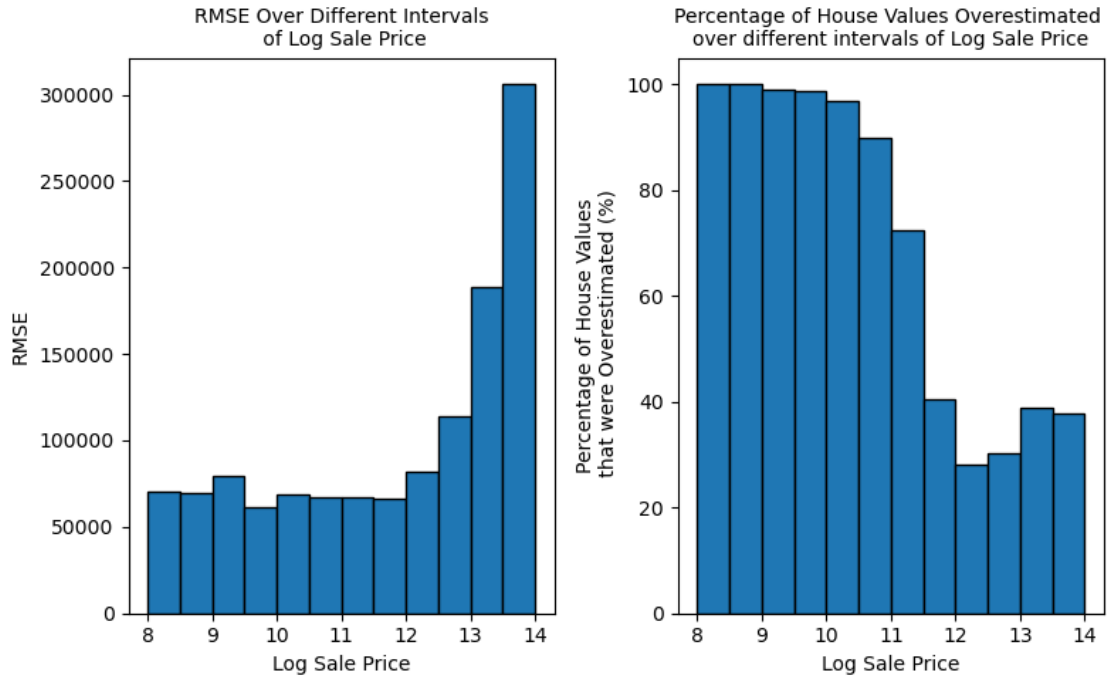
Explicitly referencing **ONE** of the plots above (using `props` and `rmses`), explain whether the assessments your model predicts more closely aligns with scenario C or scenario D that we discussed back in `q1b`. Which of the two plots would be more useful in ascertaining whether the assessments tended to result in progressive or regressive taxation? Provide a brief explanation to support your choice of plot. For your reference, the scenarios are also shown below:

```
C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive
D. An assessment process that systematically undervalues inexpensive properties and overvalues expensive
```

Scenario C is obviously true here when observing the props graph, as almost 100% of extremely inexpensive properties are overvalued, while this number decreases to only 30-40% of extremely expensive properties being overvalued.

The props graph clearly shows the distribution of homes that are overvalued and their corresponding true log sale prices, which allows us to determine the regressivity or progressivity of the taxation. In this scenario, the taxation is definitely regressive, as lower-assessed properties are overvalued and thus must pay a higher proportion in effective tax.

## 0.3   Question 7: Evaluating the Model in Context

---

## 0.4   Question 7a

When evaluating your model, we used RMSE. In the context of estimating the value of houses, what does the residual mean for an individual homeowner? How does it affect them in terms of property taxes? Discuss the cases where the residual is positive and negative separately.

A negative residual means that the value of Y - Y hat is negative, or that the predicted value of the property is larger than the actual value of the property. This means that the individual with a negative residual would have to pay more in tax than they should according to the true value of their property.

Conversely, a positive residual means that the model is undervaluing the property, and thus the owner would be paying less tax on their property than they actually should.

## 0.5 Question 7b

Reflecting back on your exploration in Questions 6 and 7a, in your own words, what makes a model's predictions of property values for tax assessment purposes "fair"?

This question is open-ended and part of your answer may depend on your specific model; we are looking for thoughtfulness and engagement with the material, not correctness.

**Hint:** Some guiding questions to reflect on as you answer the question above: What is the relationship between RMSE, accuracy, and fairness as you have defined it? Is a model with a low RMSE necessarily accurate? Is a model with a low RMSE necessarily "fair"? Is there any difference between your answers to the previous two questions? And if so, why?

Looking at my model, while minimizing the RMSE would ideally lead to the most accurate model, it did not actually lead to the most "fair" model. This is because the model propogated a regressive taxation model, which is not fair for the lower income individuals compared to the higher income individuals. As discussed earlier in the project, lower income individuals may be less able to afford increases in their taxation, whereas higher income individuals are more likely to view the increase in tax as a simple inconvenience. The magnitude of the consequences for inaccuracte increases in taxation are simply greater for the lower income homeowners compared to the higher income homeowners. While we should ideally aim for a completely accurate model that doesn't overvalue or undervalue any property, the more fair model would be the one that propagates a progressive model; if we had to choose between overvaluing lower true-value properties versuse overvaluing higher true-value properties (for the purposes of taxation), it would be more fair to overvalue the higher true-value property. This is because the higher income homeowner would likely have more access to the appeals process, as well as not suffer as grave consequences if they were to actually have to pay the increased property tax.

Viewing the whole situation from a Utalitarian ethical framework, the most fair model is the one that leads to the greatest net benefit in society. In this case, having a progressive model would be more fair and ethical because it would allow for a greater number of lower-income individuals to benefit immensely, while only slightly hindering a much smaller number of extremely wealthy individuals