## 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 part.**

---

### 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 to be high or low.**

1. Real estate agents, would like to see high housing price.
2. Buyers, would like to see low housing price.
3. Seller, would like to see high housing price.

### 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.

B. Unfair because then it means that this homeowners may enjoy lower tax obligations, while others pay higher taxes for houses with similar value.

### 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: In addition to reading the paragraph above you will need to watch the lecture to answer this question)

The wealthy were paying lower effective tax rate, placing a disproportionate burden on lower income homeowners. Some are due to the recession in 2008, where house prices plummeted and influced the pricing model. Another part was due to the appeal process where the wealthy can appeal their assessment.

### 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?

Typically most the wealthy get to appeal their housing assessment successfully, and a large of the wealthy are made up of white homeowners.

## 0.2 Question 4a

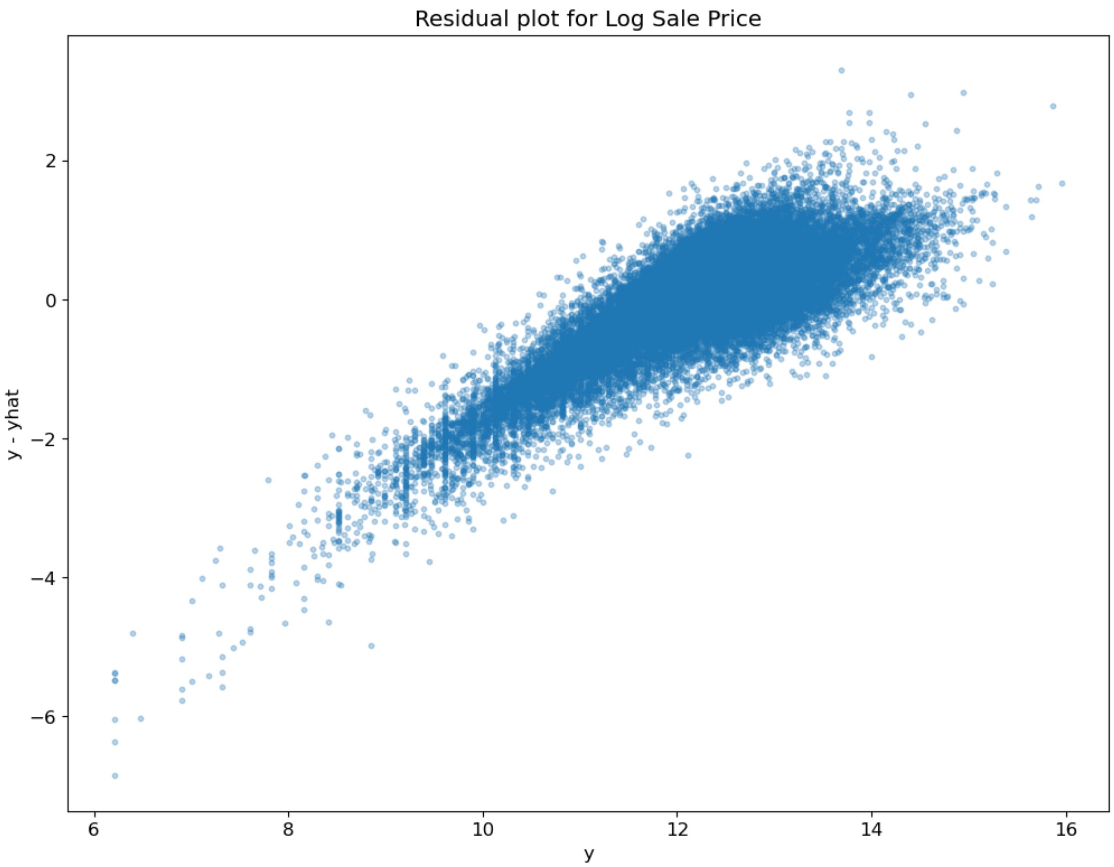
One way of understanding a model's performance (and appropriateness) is through a plot of the residuals versus the observations.

In the cell below, use `plt.scatter` to plot the residuals from predicting `Log Sale Price` using **only the 2nd 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]: #2nd model, validation data, residuals
         residual =  Y_valid_m2 - Y_predicted_m2

         plt.scatter(y = residual, x = Y_valid_m2, alpha = 0.3, s = 10
                    )
         plt.title("Residual plot for Log Sale Price")
         plt.xlabel("y")
         plt.ylabel("y - yhat")
```

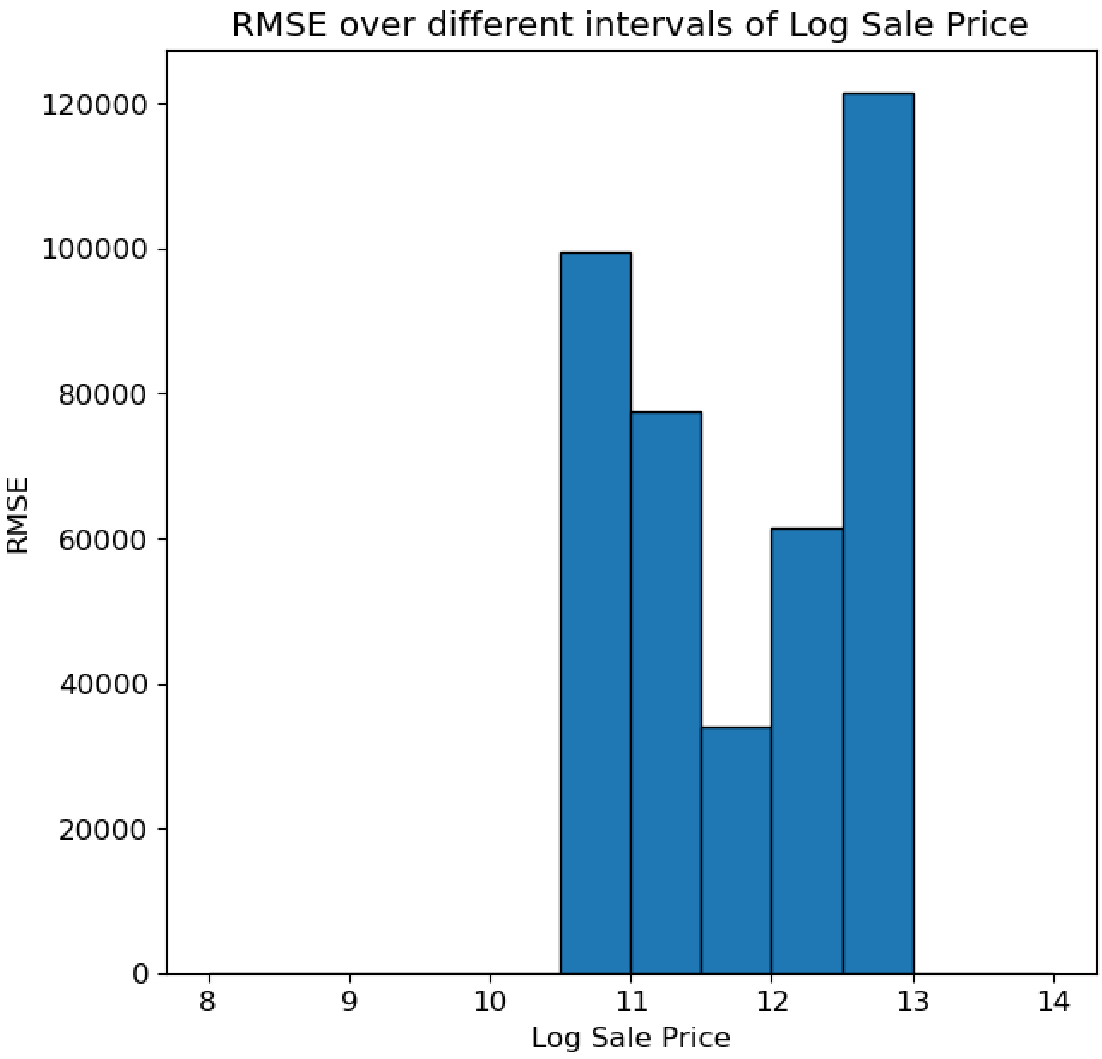
```
Out[24]: Text(0, 0.5, 'y - yhat')
```


Residual plot for Log Sale Price
y - yhat
2
0
−2
−4
−6
y
6
8
10
12
14
16

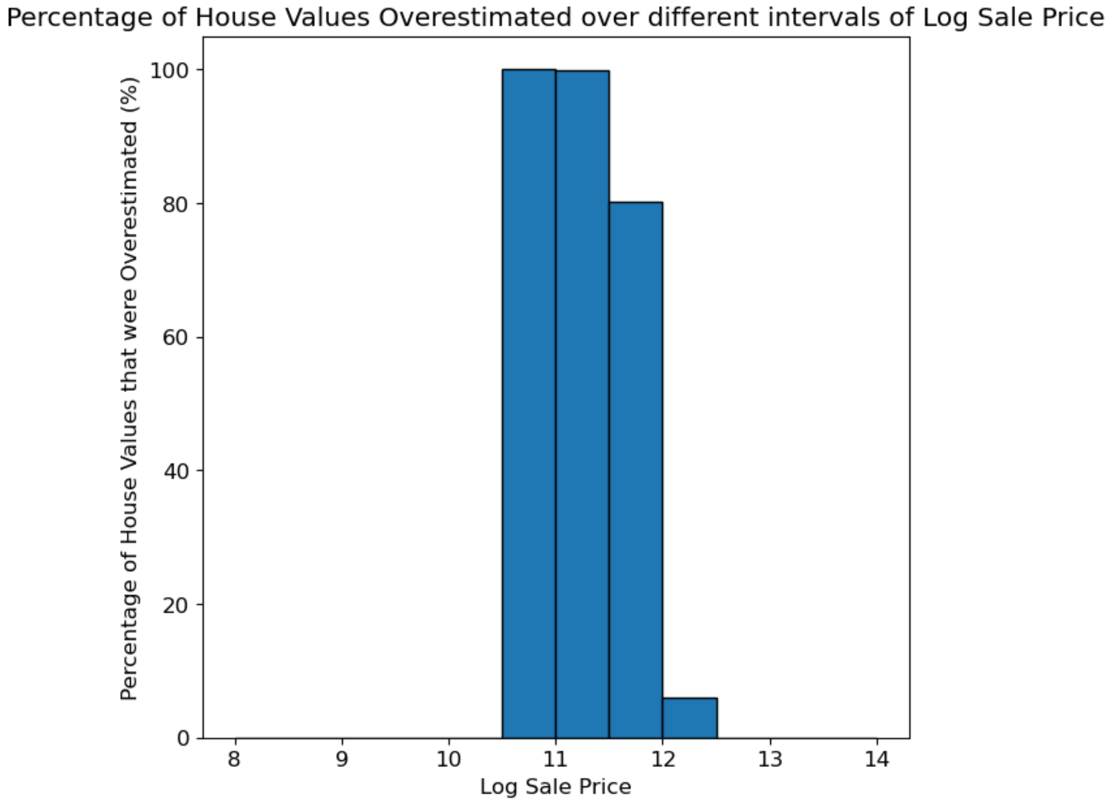
### 0.2.1 Question 6c

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.

```
In [50]: # Run the cell below to generate the plot; no further action is needed
         rmses = []
         for i in np.arange(8, 14, 0.5):
             rmses.append(rmse_interval(X, Y, i, i + 0.5))
         plt.figure(figsize = (7, 7))
         plt.bar(x = np.arange(8.25, 14.25, 0.5), height = rmses, edgecolor = 'black', width = 0.5)
         plt.title('RMSE over different intervals of Log Sale Price')
         plt.xlabel('Log Sale Price')
         plt.ylabel('RMSE');
```



```
In [51]: # Run the cell below to generate the plot; no further action is needed
         props = []
         for i in np.arange(8, 14, 0.5):
             props.append(prop_overest_interval(X, Y, i, i + 0.5) * 100)
         plt.figure(figsize = (7, 7))
         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 over different intervals of Log Sale Price
         plt.xlabel('Log Sale Price')
         plt.ylabel('Percentage of House Values that were Overestimated (%)');
```



Explicitly referencing **any ONE** of the plots above (using `props` and `rmses`), explain whether the assessments your model predicts more closely align with scenario C or scenario D that we discussed back in `1b`. 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 expensiv
D. An assessment process that systematically undervalues inexpensive properties and overvalues expensiv
```

Props plots aligns more with scenario C. The proportion of overestimated values descreases significantly as log sale price increase.

## 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 residual is positive and negative separately.

Positive residual means that the model underestimated the value of the house, the homeowner would be paying less in property taxes if they are based on the estimated value. On the otherhand, negative residual means that the model has overestimated the value of the house, homeowners paying more taxes.

## 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 upon 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?

Model with low RMSE is not necessarity "fair" as we can see from props plots because the model might be "bias". It might be more "fair" if the prop has uniform distribution.