## 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.**

Home buyers in Cook County have an interest in seeing the housing price be low, home sellers and developers have an interest in seeing the housing prices be high, and property tax advocacy organizations may want housing prices to be low for low-income homeowners and high for wealthier homeowners to support tax fairness.

### 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.

Each scenario reflects a form of unfairness, but some are more unfair than others. Scenario A is unfair because a homeowner whose property is overvalued for assessment purposes will end up paying more in property taxes than the home is actually worth on the market. This discrepancy can be particularly burdensome if the tax rate is now unaffordable based on the assessed value rather than the realistic market price. Scenario C, however, seems the most concerning because it systematically disadvantages lower-income homeowners by overvaluing inexpensive properties and undervaluing expensive ones. This essentially shifts a greater tax burden onto those with less financial flexibility, while wealthier property owners benefit from paying less tax. This discrepancy can contribute to a widening wealth gap by allowing wealthier individuals to retain and grow their assets, while those with lower incomes face increasing financial strain.

---

### 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.

"An Unfair Burden: Cook County failed to value homes accurately for years. The result: a property tax system that harmed the poor and helped the rich"

The central problem with the earlier property tax system in Cook County, as reported by the Chicago Tribune, was a systematic bias that disproportionately impacted Black and Hispanic communities. The assessment process often overvalued lower-priced properties, leading to higher taxes for working-class homeowners, while undervaluing higher-priced properties, benefiting wealthier homeowners. Primary causes included outdated assessment practices, the long-term effects of the 2008-2009 recession and housing crisis, and the intersection of social and technical biases that contributed to a regressive tax system.

---

#### 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?

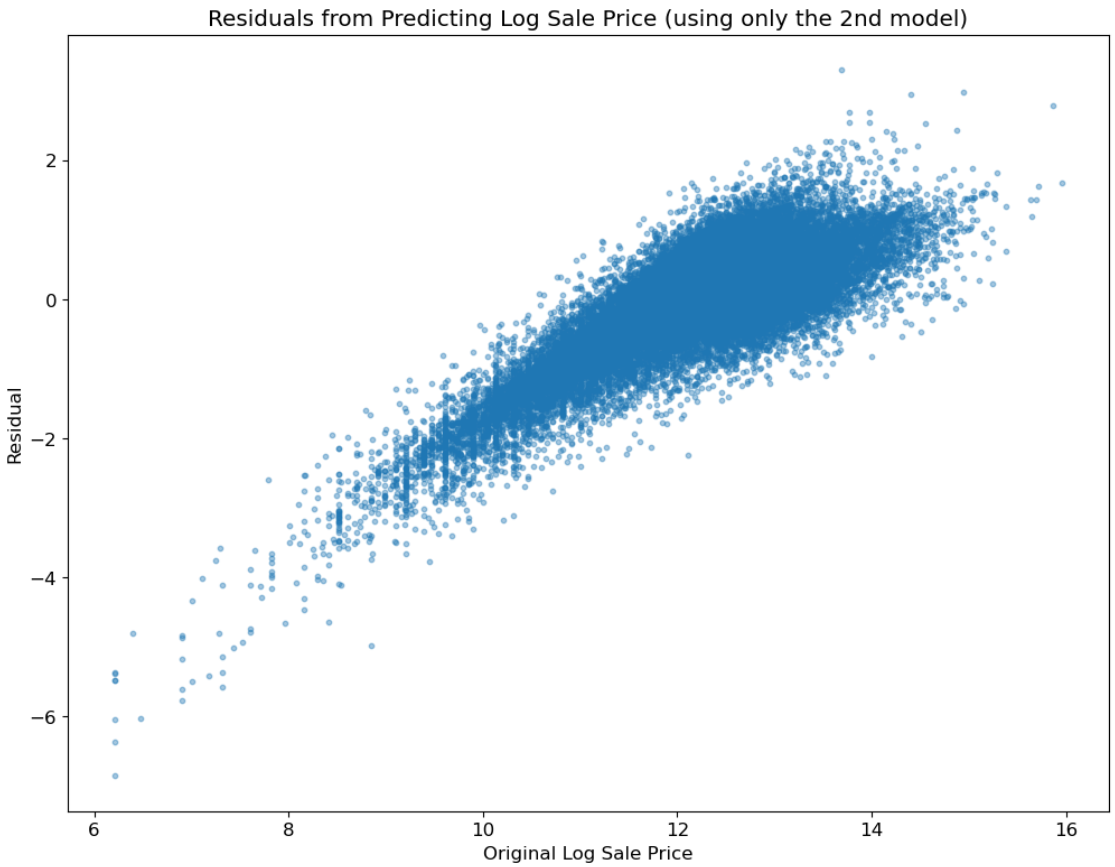
The tax system in Cook County disproportionately impacted non-white property owners by making it easier for wealthier individuals to successfully challenge their assessments before the review board, while lower-income residents faced more obstacles. This disparity existed because wealthier individuals often have more access to resources like tax attorneys, higher education levels to navigate the appeal process, and the flexibility to take time off of work, resources that lower-income homeowners typically do not have or have limited access to.

## 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]: plt.scatter(x=Y_valid_m2, y=(Y_valid_m2 - Y_predicted_m2), s=10, alpha=0.4)
         plt.xlabel('Original Log Sale Price')
         plt.ylabel('Residual')
         plt.title('Residuals from Predicting Log Sale Price (using only the 2nd model)');
```



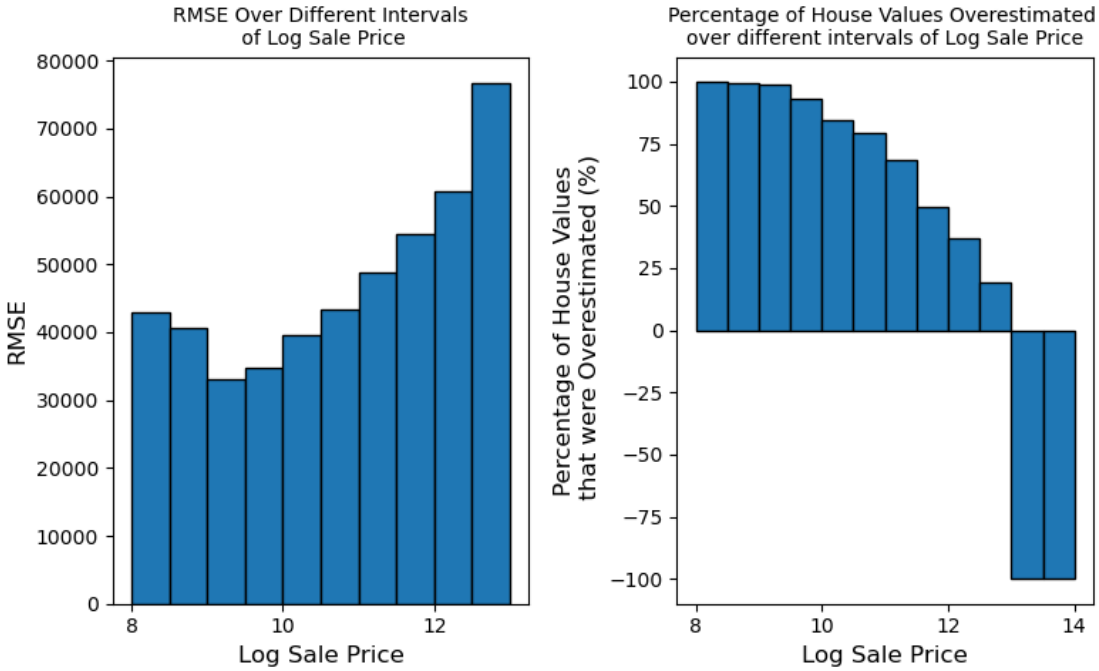
### 0.2.1 Question 6c

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 [41]:
```
# 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()
```




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
```

The overestimation plot (right plot) aligns more closely with scenario C, where the model systematically overvalues inexpensive properties and undervalues expensive ones. This plot is more useful in assessing progressive or regressive taxation, as it shows that lower-priced properties tend to be overestimated, leading to a regressive tax effect where less expensive properties face a disproportionally higher tax burden.

## 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.

The residual represents the difference between the actual and predicted sale prices for a homeowner's property. A positive residual (underestimation) means lower property taxes than the true value might suggest, benefiting the homeowner, while a negative residual (overestimation) means higher taxes, placing an extra financial burden on the homeowner.

## 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?

A model's predictions of property values are "fair" when errors are evenly distributed across all property types and values, so no group is systematically over- or under-assessed. While a low RMSE indicates overall accuracy, it doesn't guarantee fairness, as the model might still consistently overestimate lower-priced properties and underestimate higher-priced ones, leading to regressive tax effects. Fairness requires both low average error (RMSE) and balanced errors across different property segments, ensuring equitable tax assessments regardless of a property's market value.