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. someone interested in buying a house – interested in low housing price
2. someone interested in selling a house – interested in high housing price
3. real estate brokers – interested in 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.

I think that all of these scenarios are unfair: A. this is unfair to the homeowner because they could end up paying higher property taxes because the value of the house is higher than it actually is B. this could be beneficial to the homeowner because they might pay lower property taxes, but it could be unfair to other homeowners in the area paying higher property taxes for a similar property. C. this scenario is unfair because it disproportionally affects homeowners with inexpensive properties by making them pay higher taxes, which would increase the inequality gap. D. this scenario benefits homeowners of inexpensive properties but is a disadvantage to those with more expensive properties who will have to pay higher taxes than what their property is actually worth.

i think that scenarios A and C are less fair than B and D, but all of them have disproportionate effects.

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 main issue with the property tax system in Cook County is that it was found that there was systematic bias and corruption in the system that made the tax system highly regressive. This meant that a higher tax burden disproportionately fell on low income black homeowners. This was a result of low income homes being overvalued and wealthy homes were being undervalued. Additionally, the system allowed wealthy homeowners to more easily adjust and lower their property value by challenging the assessments in front of a review board. wealthy homeowners had greater access to things like time off, tax lawyers (who can find loopholes), and connections to members of the tax review board that could accomplish this, while lower-income homeowners did not.

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?

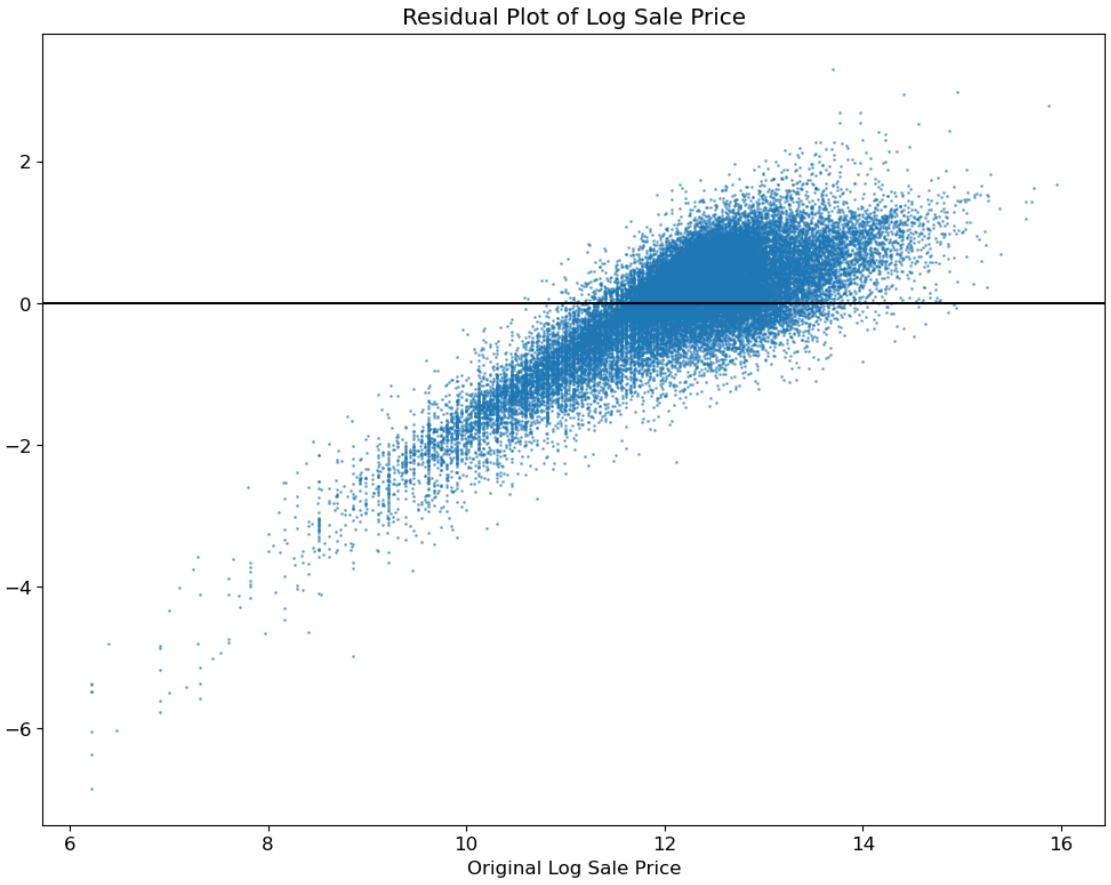
The property tax system disproportionately burdened non-white property owners due to Chicago's history of racial inequality. Chicago is one of the country's most segregated cities, and thus has evidence of redlining which made it very difficult for African Americans to get mortgages to buy a house. The fact that specific neighborhoods were redlined was factored into house valuation procedures, which overassessed non-white homeowners' properties because they were deemed "too risky"

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 [77]: residuals_m2 = Y_valid_m2 - Y_predicted_m2
         plt.scatter(Y_valid_m2, residuals_m2, s = 1, alpha = 0.5)
         plt.title('Residual Plot of Log Sale Price')
         plt.xlabel('Original Log Sale Price')
         plt.axhline(y=0, color = 'black');
```



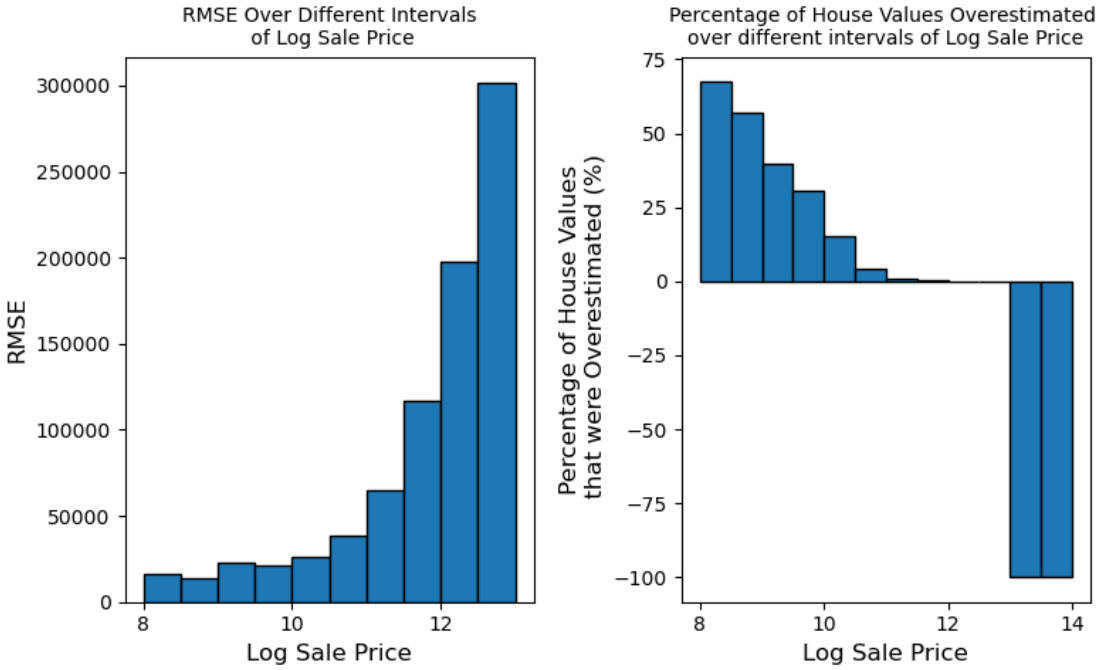
0.2.1 Question 6c

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 [104]:
```
# 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 Pr
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
```

I think that my model predictions more closely align with scenario C. As you can see in the second plot, homes with lower sale prices are overestimated most of the time whereas homes of higher value are underestimated 100% of the time. This implies a regressive model, which disproportionately affects people of lower income, making them pay higher property taxes than their property is worth, while people of high income pay lower taxes than they should. The second plot is more useful because it shows over and underestimation of property value, while the first one does not tell us anything about the estimates of the properties.

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.

In the context of house valuation, the residual is the difference between the actual market value of the house and the predicted value of the house. In terms of property taxes, a positive residual indicates that the predicted value of a house is less than its actual value, which would result in a lower property tax for the homeowner. A negative residual indicates that the predicted value of the house is larger than the actual value of the house, which would result in the homeowner paying a higher property tax than they 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?

RMSE is a good predictor for accuracy, but the fairness of the model's predictions depends on more factors than just the RMSE. In order to be fair, we must ensure that the model doesnt systematically over or undervalue certain properties. Ensuring fair predictions for property taxes requires consideration of various other factors. The model should be able to minimize bias, be transparent enough to be easily interpreted, and avoid increasing socioeconomic gaps through discrimanatory predictions.