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

Consider the following question: *"How much is a house worth?"*

Who might be interested in an answer to this question? Be sure to list at least three different parties (people or organizations) and state whether each one has an interest in seeing a low or high housing price.

Your response should be approximately 3 to 6 sentences.

I think homeowners, government, and real estate agents would interested in seeing a low or high housing prices. The homeowners interest would be that the housing price increases so that they can retain a larger return on investment. The real estate agents would interest in knowing the value of a house in order to provide an accurate estimate to their clients and prefer to see the higher prier as it can potentially increase their commission. Lastly, the government wold interested in housing property. Higher property values typically lead to higher property tax revenue.

0.1.2 Question 1b

Which of the following scenarios strikes 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.

Your response for each chosen scenario should be approximately 2 to 3 sentences.

I think A and C are the most unfair. A leaves the home owner with a higher tax cost than b. C is unfair as it places a harsher tax burden on lower income individuals and a lesser burden on higher income individuals, furthering the income disparity between classes. Furthermore, C instantiates that this is systematic process which means unlike A and B being one time occurances the tax breakdown and income marginalization with happen to a larger pool of individuals.

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?

Your response should be approximately 2 to 4 sentences.

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

The assessor's office undervalued expensive homes and overpriced low-priced residences, resulting in "racially discriminatory assessments and taxes" and a clear racial gap. These were the main issues with Cook County's previous property tax system. Effective tax rates decrease with increasing income levels. Furthermore, property owners found it challenging to comprehend the assessment process's opaqueness and how their tax bills were determined. Property owners had limited options if they thought their assessments were inaccurate, and the system lacked transparency. Political influence might affect the assessment process, and property values were frequently adjusted to favor those with connections. These issues were mostly brought on by a system that was too complex and challenging to use, a lack of control, and insufficient money. Furthermore, because politicians and other well-connected people frequently profited from the current system, their influence made system transformation challenging.

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?

Your response should be approximately 3 to 4 sentences.

The property tax system in Cook County disproportionately burdened non-white property owners by reinforcing financial hardships and systemic inequalities. Many lived in high-poverty areas, leading to higher tax delinquency and foreclosure rates. Redlining further deprived these neighborhoods of investments, lowering property values and tax revenue. Since property taxes fund schools, underfunded education systems worsened disparities, leaving non-white property owners paying more while receiving fewer benefits.

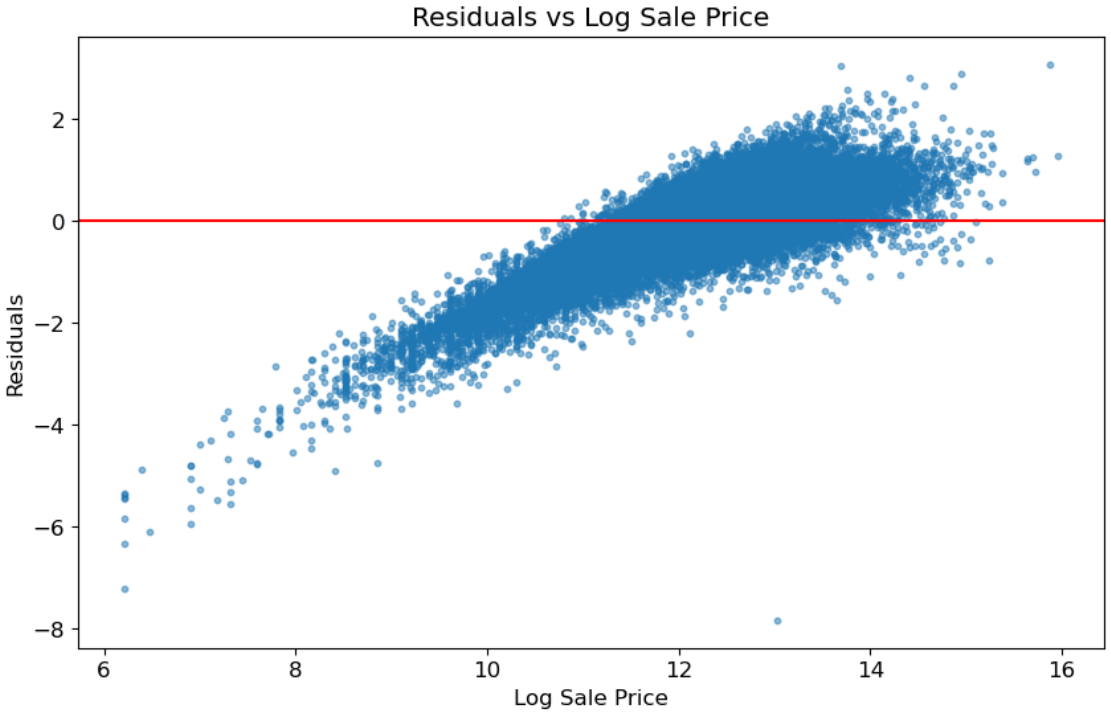
0.2 Question 4a

We can assess a model's performance and quality of fit with a plot of the residuals $(y - \hat{y})$ versus the observed outcomes (y).

In the cell below, use `plt.scatter` (documentation) to plot the **model 2** residuals of `Log Sale Price` versus the original `Log Sale Price` values. For this part, you only need to plot the residuals and outcomes for the **validation data**.

- 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. However, with such a large dataset, it is difficult to avoid overplotting entirely.

```
In [62]: plt.figure(figsize = (10, 6))
         plt.scatter(Y_valid_m2, Y_valid_m2 - Y_predicted_m2, alpha = 0.5, s = 10)
         plt.axhline(y=0, color = 'r', linestyle = '-')
         plt.xlabel("Log Sale Price")
         plt.ylabel('Residuals')
         plt.title('Residuals vs Log Sale Price')
         plt.show()
```



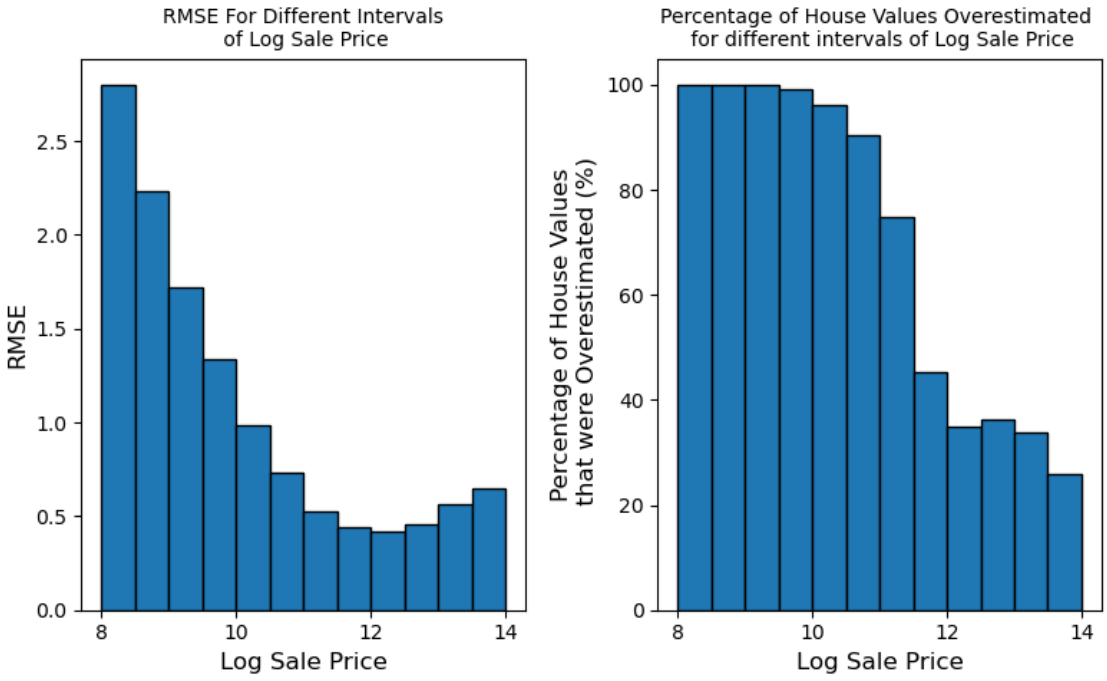
0.2.1 Question 6c

Using the functions above, we can generate visualizations of how the RMSE and proportion of overestimated houses vary for different intervals:

In [99]:
```
# 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 For 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 \n for 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()
```



Which of the two plots above 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.

Then, explain whether your chosen plot aligns more closely aligns with scenario C or scenario D from `q1b`:

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

Your response should be approximately X to Y sentences.

The second plot, "Percentage of House Values Overestimated for Different Intervals of Log Sale Price," would be more useful for ascertaining whether assessments led to progressive or regressive taxation. This is because it directly shows the bias in assessments (overestimation or underestimation) across different property value intervals, which is key to identifying systematic patterns in taxation fairness.

The plot aligns more closely with scenario C, where inexpensive properties are overvalued (nearly 100% overestimation on the left side) and expensive properties are undervalued (lower overestimation percentages on the right). This pattern suggests a regressive taxation effect, as cheaper homes are more likely to be overtaxed compared to more expensive ones

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 a positive or negative residual affect them in terms of property taxes? Discuss the cases where the residual is positive and negative separately.

Your response should be approximate 2 to 4 sentences.

A residual represents the difference between the model's predicted log sale price and the actual log sale price for a house. A positive residual means the model underestimated the home's value, which may result in the homeowner being undertaxed. A negative residual means the model overestimated the home's value, potentially causing the homeowner to be overtaxed. In either case, inaccurate residuals can lead to unfair tax burdens on individual households.

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?

Your response should be approximate 1 to 2 paragraphs. Feel free to answer the questions in the hint to structure your answer.

A model's predictions are only "fair" if they are not just accurate on average, but also equitable across different communities. A low RMSE may suggest the model performs well overall, but it can still be unfair if it consistently overestimates property values in historically marginalized neighborhoods and underestimates them in wealthier areas. Fairness in this context must consider historical and systemic inequities — especially in places like Cook County, where property tax burdens have disproportionately fallen on Black and Hispanic homeowners due to biased assessments. Therefore, a fair model is one that both minimizes error and ensures that its residuals are not systematically skewed across race, income, or neighborhood lines.