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.

A homeowner might want the value of their house to be high because it increases their wealth and helps if they plan to sell. On the other hand, a buyer is likely hoping for a lower price so they can afford the property. The local government or the Assessor's Office, like the CCAO, may want accurate or even higher assessments, since property taxes are based on those values. But high prices can also raise concerns about fairness and affordability, especially for low-income residents. So, depending on the perspective, "how much a house is worth" can be a question of power, access, and trust — not just money.

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.

C strikes me as the most unfair, because it systematically overvalues less expensive homes and undervalues more expensive ones. This shifts the tax burden toward lower-income households, which makes the system regressive. According to the lecture, this pattern was part of why Cook County faced criticism and reform — it deepened existing racial and income inequalities.

A is also unfair, but more on an individual level — when your home is assessed too high, you're paying more than you should in taxes. But C creates a larger structural injustice.

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.

One major problem was that the Cook County tax system consistently overvalued cheaper homes and undervalued expensive ones, leading to a regressive tax burden. This meant lower-income, often non-white homeowners ended up paying more than their fair share in taxes, while wealthier, mostly white property owners paid less. The root cause was a flawed and opaque assessment system that relied on outdated methods, biased data, and a broken appeals process that favored those with more time, money, and knowledge. These issues were exposed in The Tax Divide investigation and became the basis of a lawsuit against the Assessor's Office.

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 tax system in Cook County disproportionately impacted non-white property owners because it consistently overvalued homes in majority-Black and Latinx neighborhoods. These communities were more likely to have lower-priced properties that were assessed too high, leading to inflated tax bills. At the same time, wealthier, whiter neighborhoods benefited from undervalued assessments and were more likely to appeal and win reductions. This pattern reinforced racial inequality by forcing working-class communities of color to subsidize public services while wealthier homeowners paid less than their fair share.

0.2 Question 4a

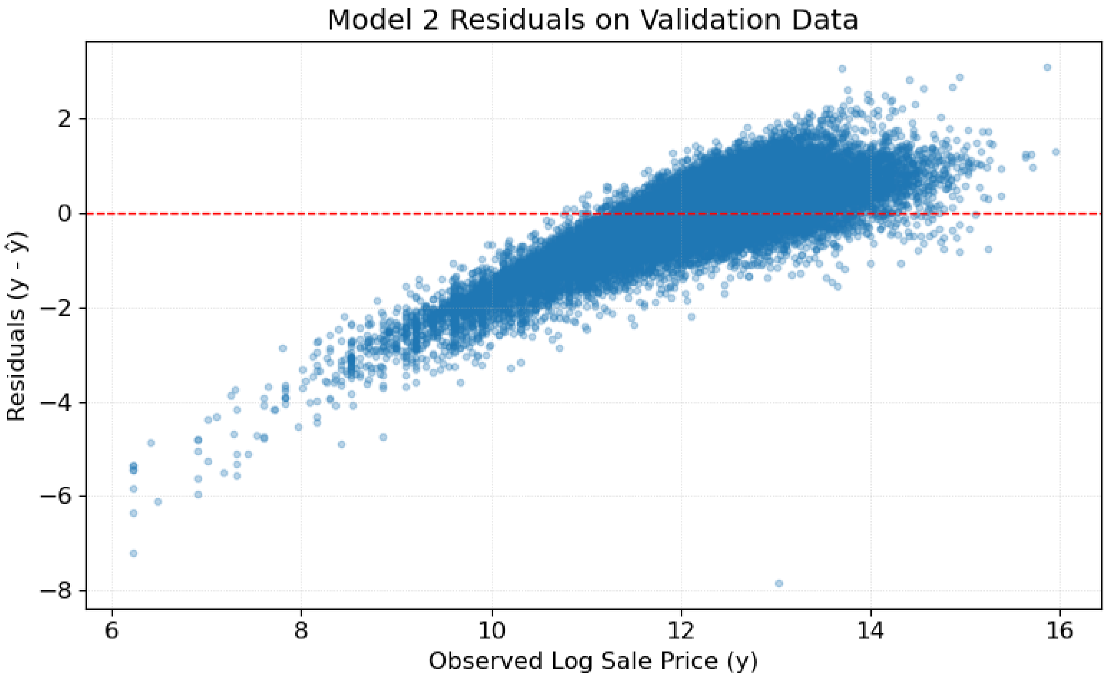
We can assess a model's performance and quality of fit with a plot of the residuals $(y - \hat{y})$ versus the observed outcomes (y).

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 [61]: # Calculate residuals for Model 2 (validation set)
         residuals_m2 = Y_valid_m2 - Y_predicted_m2

         # Create the residual scatter plot
         plt.figure(figsize=(8, 5))
         plt.scatter(
             Y_valid_m2,              # x-axis: observed Log Sale Price
             residuals_m2,            # y-axis: residuals
             s=10,                    # small dot size
             alpha=0.3                # semi-transparent dots to reduce overplotting
         )
         plt.axhline(y=0, color='red', linestyle='--', linewidth=1)  # reference line at 0
         plt.xlabel("Observed Log Sale Price (y)")
         plt.ylabel("Residuals (y - ŷ)")
         plt.title("Model 2 Residuals on Validation Data")
         plt.grid(True, linestyle=':', linewidth=0.5, alpha=0.6)
         plt.tight_layout()
         plt.show()
```


Model 2 Residuals on Validation Data
Residuals (y - ŷ)
2
0
−2
−4
−6
−8
6
8
10
12
14
16
Observed Log Sale Price (y)

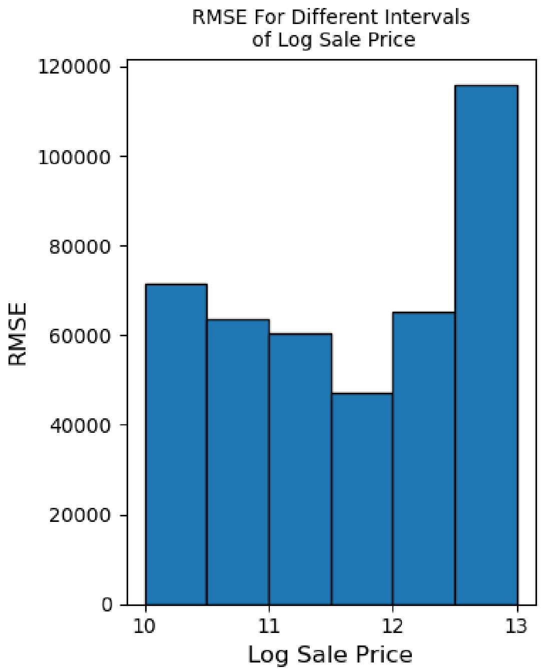
0.2.1 Question 6c

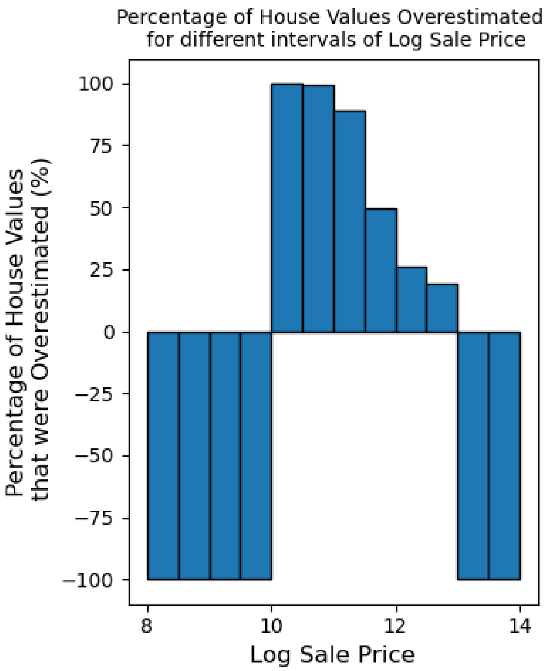
Using the functions above, we can generate visualizations of how the RMSE and proportion of overestimated houses vary for different intervals:

```
In [42]: # 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 expensiv
D. An assessment process that systematically undervalues inexpensive properties and overvalues expensiv
```

Your response should be approximately X to Y sentences.

Between the two, the overestimation plot is more helpful for figuring out whether the assessments are progressive or regressive. RMSE tells us how off the predictions are, but it doesn't show if the model is consistently over- or underestimating prices. The overestimation plot clearly shows that—especially whether cheaper houses are being overvalued more often than expensive ones.

Looking at the plot, there's a higher percentage of overestimated values for cheaper homes. That suggests the model tends to overvalue less expensive properties and undervalue more expensive ones. So this pattern matches Scenario C from Q1b: the one where inexpensive homes are overvalued and expensive ones are undervalued.

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 positive residual means the predicted sale price is lower than the actual sale price, so the homeowner might get under-taxed. A negative residual means the model overestimated their property's value, which could lead to higher taxes than they should fairly pay. This matters because overvaluation can burden some homeowners more than others, especially if there are patterns by neighborhood or demographic. That's why minimizing residuals is important—not just for accuracy, but for fairness too.

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 fair model for tax assessment doesn't just mean one with a low RMSE. While accuracy is important, fairness also depends on who is consistently over- or under-valued. For example, if the model has a low RMSE overall but tends to overvalue cheaper homes and undervalue expensive ones, it might lead to regressive outcomes where lower-income homeowners pay more than they should.

So, fairness is not just about minimizing the average error—it's about making sure those errors aren't biased toward certain types of homes or neighborhoods. A truly fair model would aim for both low error and balanced residuals across different value ranges, so that no group is unfairly taxed due to systematic over- or underestimation.