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.

Homeowners would want to know because they'd love to see a high price—more value in their home means a bigger payoff if they sell, but at the same time, the homeowners who doesn't want to sell would like to see a lower housing price, so that they could pay less property tax. Real estate investors, on the flip side, would probably hope for a lower price so they can buy cheap and sell more to make money later, but the agent(non-investors) would like to see higher housing price since they earn commission based off of the percentage of the house value. Besides, the local government might root for a higher price to rake in more property taxes that goes to the county's budget.

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 find Scenario C the most unfair because its regressive nature systematically disadvantages lower-income homeowners, who are often less equipped to handle an increased tax burden, thus widening economic inequality. Scenario A is also unfair but affects individuals rather than a systematic group, making it less severe in its broader impact. Scenario D, while still unfair due to its systematic bias, aligns more with progressive taxation principles, which some might argue is fairer since it places a heavier burden on those with greater financial capacity. Scenario B is the least unfair from the individual homeowner's perspective, as they benefit from a lower tax bill, though it may have broader implications for tax revenue fairness if widespread.

In summary, Scenarios A, C, and D strike me as unfair due to their impact on tax equity, with Scenario C being the most unfair due to its regressive effects, while Scenario B is the least unfair to the individual homeowner but may pose fairness issues at a systemic level.

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 Chicago Tribune's investigative series revealed that Cook County's previous property tax system was deeply flawed, with assessments that were both racially biased and inaccurate, placing a heavier burden on lower-income, non-white homeowners while giving wealthier, mostly white homeowners a break. This regressive setup consistently overvalued cheaper homes and undervalued pricier ones, leading to a clear racial disparity in tax responsibilities. The root causes included unfair assessment methods that favored certain groups, a lack of openness in how valuations were determined, insufficient oversight, and the use of outdated statistical models running on an old mainframe system from the early 1990s. On top of that, a broken appeals process made things worse—owners of high-value properties were more likely to appeal and win reductions, underscoring the pressing need for changes to create a fairer, more equitable tax system for everyone in Cook County.

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 placed a disproportionate tax burden on non-white property owners because the regressive assessment practices—overvaluing inexpensive properties and undervaluing expensive ones—aligned with existing racial and socioeconomic divides in the region. Non-white homeowners, who were more likely to live in lower-income neighborhoods with less expensive properties, faced higher effective tax rates due to these overvaluations, while wealthier, predominantly white homeowners in more affluent areas benefited from undervaluations and paid less in taxes. Additionally, the appeals process favored those with resources to challenge assessments, a group that often included white property owners, further exacerbating the racial disparity in tax burdens. This systemic bias effectively meant non-white communities shouldered a heavier tax load, deepening economic inequality along racial lines.

0.2 Question 4a

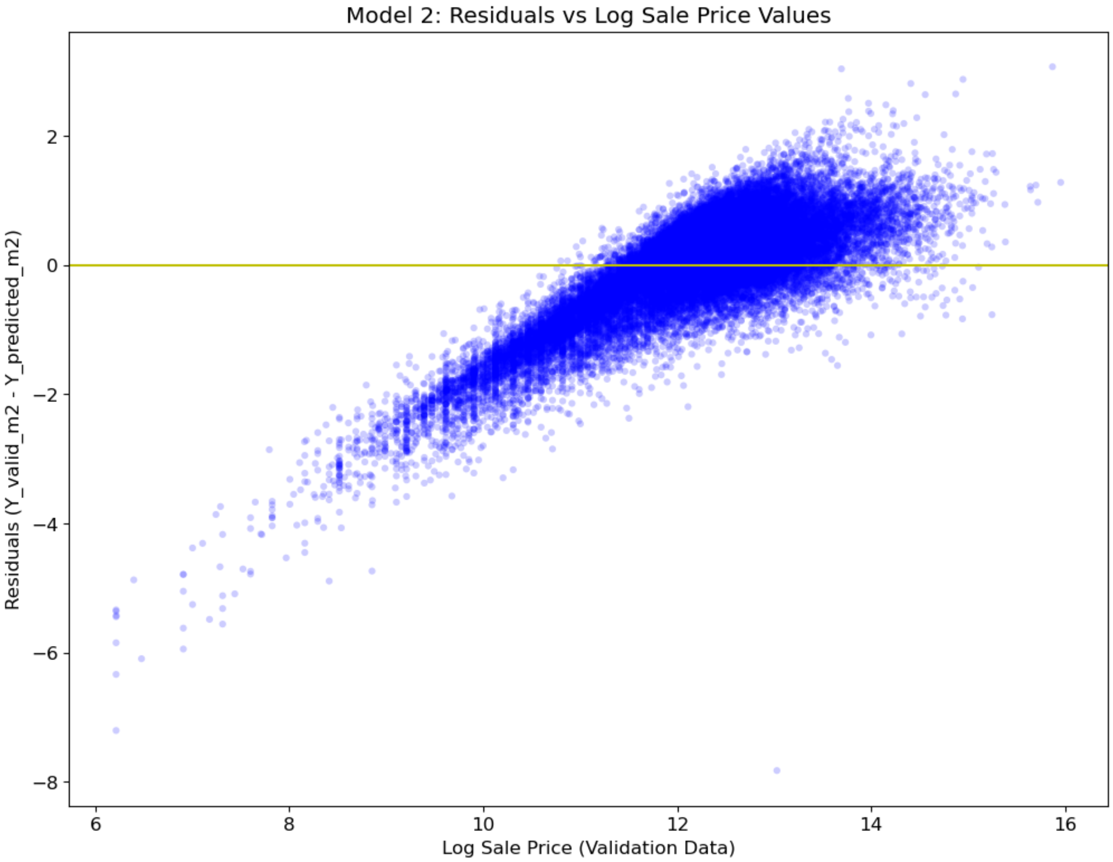
We can assess a model's performance and quality of fit with a plot of the residuals $(y - \hat{y})$ versus the observed outcomes (y).

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 [141]: residuals_m2 = Y_valid_m2 - Y_predicted_m2
          plt.scatter(Y_valid_m2, residuals_m2, alpha=0.2, s=20, c='b', edgecolors='none')
          plt.axhline(color = 'y')
          plt.xlabel('Log Sale Price (Validation Data)')
          plt.ylabel('Residuals (Y_valid_m2 - Y_predicted_m2)')
          plt.title('Model 2: Residuals vs Log Sale Price Values')
```

```
Out[141]: Text(0.5, 1.0, 'Model 2: Residuals vs Log Sale Price Values')
```


Model 2: Residuals vs Log Sale Price Values
Residuals (Y_valid_m2 - Y_predicted_m2)
2
0
−2
−4
−6
−8
6
8
10
12
14
16
Log Sale Price (Validation Data)

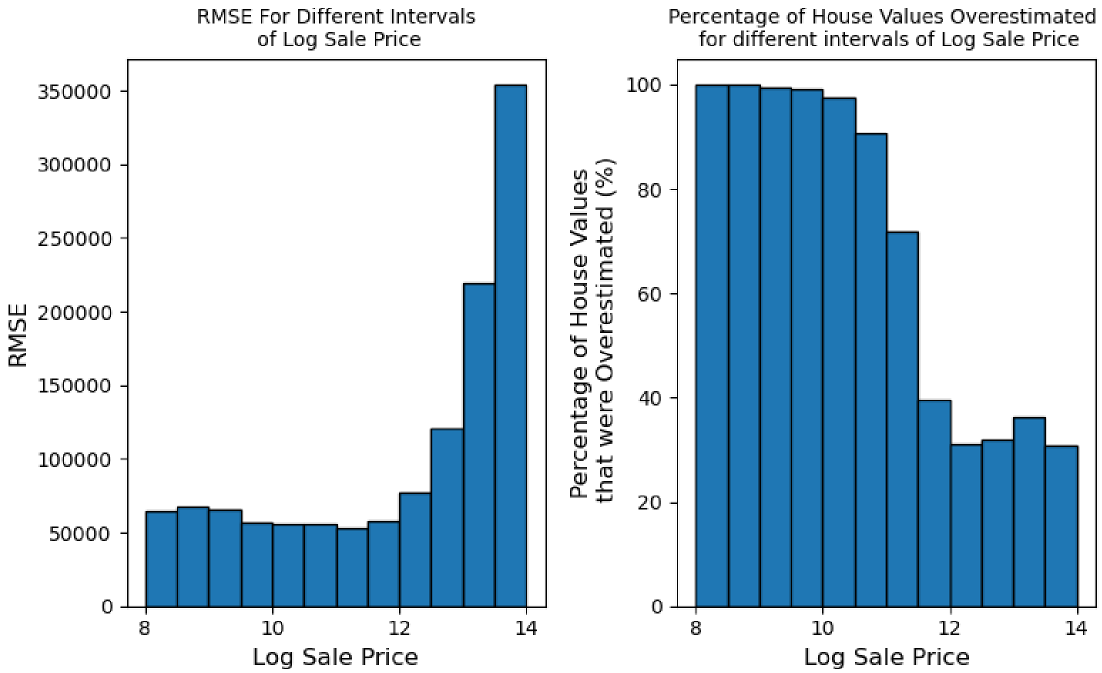
0.2.1 Question 6c

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

```
In [116]: # 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 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()
```



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.

The right plot would be more helpful and the right plot aligns with scenario C. It directly reveals the model's bias in over- or underestimating property values across price ranges, which affects tax fairness. Despite that the left plot also shows that there's a bigger error as the sale price increases, however, it doens't tell us whether this error is over estimate or underestimate errors, whereas the plot on the right shows us that information, that the lower the sale price, the higher the percentage that it was overvalued, and vice versa. This plot on the right indicates that inexpensive properties are often overvalued (higher overestimation percentage), while expensive properties are less likely to be overvalued (lower overestimation percentage), aligning more closely with Scenario C, where inexpensive properties are overvalued and expensive ones undervalued, potentially leading to regressive taxation.

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.

For an individual homeowner, the residual represents the difference between the true sale price of their house and the model's predicted sale price, which is used as the assessed value for property taxes. A positive residual (true value higher than predicted) means the model underestimates the house's value, leading to a lower tax bill, as the homeowner pays taxes on a value less than the true value, which would be a benefit to them since they're paying less tax than they should be. A negative residual (true value lower than predicted) means the model overestimates the house's value, resulting in a higher tax bill, as the homeowner pays taxes on a value greater than the true value, which can be a financial disadvantage because now they pay more tax than they should be.

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.

In the context of property tax assessments, a model's predictions can be considered "fair" if they minimize systematic bias across different property value ranges, ensuring that no group of homeowners—whether they own inexpensive or expensive properties—is disproportionately over- or undervalued, which would lead to inequitable tax burdens, and also further highlight the discrepancies between income brackets. Fairness in this sense, means that the model's errors (residuals) are not consistently skewed in a way that disadvantages certain groups, such as overvaluing inexpensive properties (leading to regressive taxation) or undervaluing expensive ones (leading to progressive taxation). Reflecting on Questions 6 and 7a, my model showed a tendency to overaluate inexpensive houses and less likely to overvalued expensive properties, as seen in the percentage of overestimated values plot. This suggests a lack of fairness because the low and mid-range properties are systematically overtaxed compared to others, creating an uneven tax burden that doesn't align with true property values.

Considering the relationship between RMSE, accuracy, and fairness, a low RMSE indicates that the model's predictions are, on average, close to the true values, which is a measure of accuracy. However, a low RMSE doesn't guarantee fairness. For example, a model could have a low RMSE but still systematically overvalue inexpensive properties and undervalue expensive ones, as seen in the left plot of 'RSME vs log sale price', leading to regressive taxation that disproportionately burdens lower-income homeowners. In my model, the RMSE varied significantly across intervals, specifically peaking at log sale price > 12, suggesting larger errors for higher-priced homes, but the percentage overestimation plot revealed the direction of those errors, which is critical for fairness. A model with a low RMSE isn't necessarily fair if its errors are biased in a way that creates inequitable tax outcomes, highlighting that fairness requires not just accuracy (low RMSE) but also an equitable distribution of errors across all property value ranges.