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 purchasing a house; low so they can pay a lower price.
2. A group/agency interested in selling the house; high so they are able to make more of a profit from selling the property.
3. City planners; low because they want to see what architecture they can incorporate into the city landscape of that specific property area.

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.

Homeowners pay property taxes for how much your property is valued for.

All of these scenarios strike me as unfair, since the property taxes will be disproportionate for inexpensive property homeowners vs. expensive property homewoners. For example, a homeowner that owns an inexpensive property will have to pay more taxes if their property is overvalued, but since they own an inexpensive property, then it might be assumed they don't have the means to purchase a more expensive property and therefore pay more expensive taxes. The same holds for the converse, for expensive property homeowners that haas their properties undervalued. This is unfair to other homeowners that are paying more for their property taxes.

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 central problems with the property tax system described above for Illinois' Cook County was that their model undervalued high-price homes, while undervaluing low-price homes- which is an example of regressive taxation and can be extremeley unfair.

The primary causes of these problems are probably due to the training dataset that used to help their model estimate these property tax assessments. It was mentioned that they consider multiple factors like real estate value and construction cost, and they may have built their assessment models on flawed data, which led to incorrect evaluation of these properties.

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?

In addition to being regressive, the property tax system in Cook County placed a disproportionate tax burden on non-white property owners because of the fact that most working-class, non-white homeowners owned the lower value properties, which is systematically unfair because in proportion to their income, lower income homeowners had to pay more property taxes than their wealthy counterparts.

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 [ ]:
```
residuals_m2 = Y_valid_m2 - Y_predicted_m2

        plt.figure(figsize=(8, 6))
        plt.scatter(Y_valid_m2, residuals_m2, alpha=0.5, s=10)  # Adjust alpha and size to reduce overp
        plt.axhline(0, color='red', linestyle='--')
        plt.xlabel("Observed Log Sale Price")
        plt.ylabel("Residuals")
        plt.title("Residual Plot for Second Model")
        plt.show()
```

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 [ ]: # 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 Price
        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
```

Our model aligns more closely with Scenario C because it suggests that the model tends to overvalue inexpensive properties and undervalue expensive ones. The Overestimation Percentage Plot would be more useful

in determining whether the assessments tend towards progressive or regressive taxationbecause it depicts the direction of the bias across different property price ranges, helping identify if the model systematically favors lower- or higher-valued 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 estimating the value of houses, the residual represents the difference between the actual value (true market value) of a house and the value predicted by the model. For an individual homeowner, the residual indicates whether the model has overestimated or underestimated the value of their property. This residual can directly impact their property tax burden if property taxes are based on the estimated value.

When the residual is positive: The true market value of a house is higher than the predicted value. The homeowner might end up paying lower property taxes than they would if the estimate were accurate, as their property appears to be worth less than it actually is. Generally beneficial for homeowners paying property taxes, as they would pay less, but bad if they were trying to sell their property because an underestimated valuation might negatively impact the perceived worth of their home.

When the residual is negative: The predicted value is higher than the true market value. The homeowner might pay more in property taxes than they would if the estimate were accurate, as their property appears to be worth more than it actually is. This is financially disadvantageous for the homeowner, as they incur higher tax expenses.

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?

1. Accuracy: The model's estimates should be as close as possible to the actual market values
2. Homogenous Across Different Properties: Fairness also means that properties of similar value are assessed similarly
3. Lack of Systematic Bias: The model should not consistently overestimate or underestimate properties in specific price ranges or neighborhoods.
4. Transparency: Homeowners should be able to understand how their property's value was determined.