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.**

PARTY 1: Cook County. Property values are used to determine property tax liability in Cook County, so they might want housing prices to be *high* to increase revenue. **PARTY 2:** Homeowners who plan to keep their house for at least a year. Property values are used to determine property tax liability in Cook County, so they might want housing prices to be *low* to decreasae their tax laibility. **PARTY 3:** Homeowners who plan to sell thier house within the year. Property assesments provide a powerful anchor for real estate transations, those who plan to sell their home want their property value to be *high* so that they can ask for a higer price when selling their home.

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 don't think fairness is the right lens here. Fairness pertains to *equality*, meaning that all parties are treated equally by the system. In that case, all four scenarios are equally unfair, as they all systematically harm one group or the other. The thing we actually care about when we talk about Cook County is *equity*. We don't just want everyone to be treated equally, we want the burdens of our system to fall onto those who are most able to cope with them. Through the lens of *equity*, options A and C are worse because the effect of these errors could easily cause someone to lose their home, where B/D type errors, while still unfair, wont cause as much material harm to the groups affected by them.

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.

One problem was that the proccess to appeal am innacurate appraisal was only avaialble to those with the means to persue a costly and time-consuming beuracratic battle. In effect this meant that only high income individuals could petition to lower thier assesment. Another problem was that the model systematically overvalued less expensive 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?

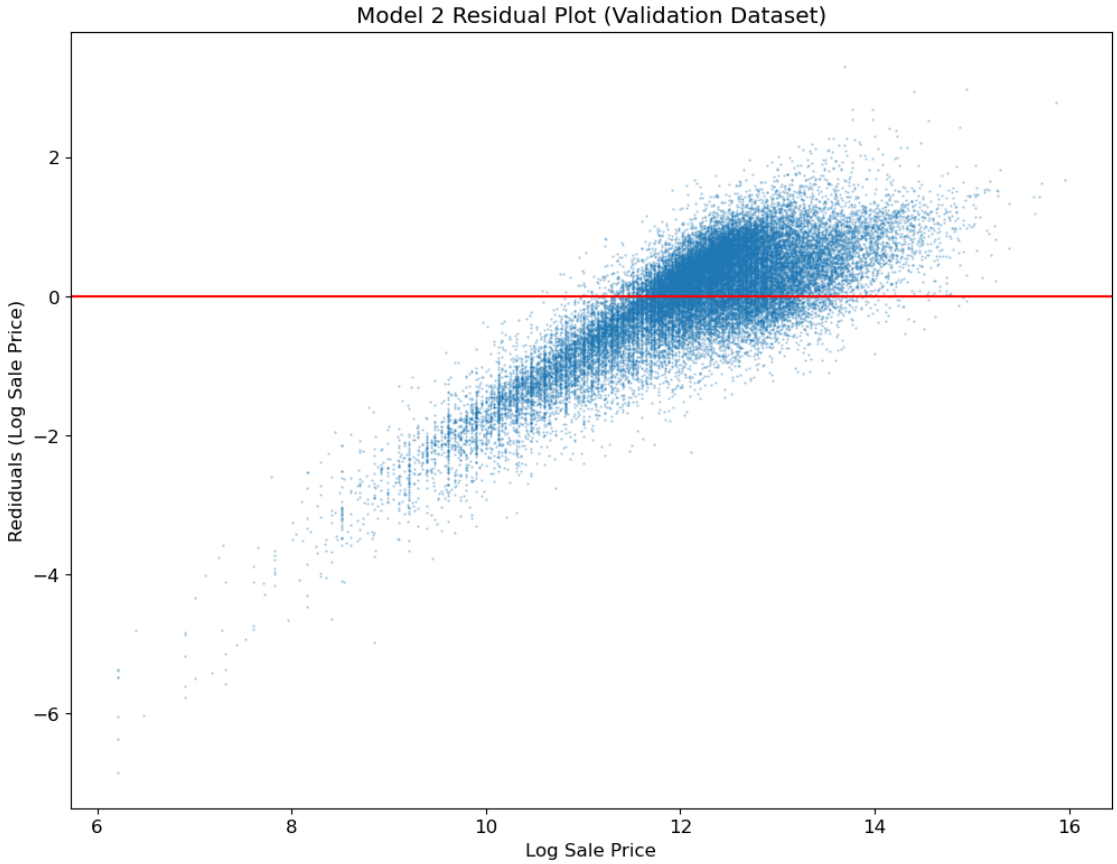
Due to Chicago's history of racially discriminatory housing practices (redlining, etc), an association was created between living areas with lower property values and being "non-white," making property value a loose proxy for race in Cook County. Thus the regressive system which placed a disproportionate burden on lower valued properties also dispropotionatley affected non-white property owners.

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 [78]: validation_residuals = Y_valid_m2-Y_predicted_m2
         x_vals, y_vals = Y_valid_m2, validation_residuals
         plt.scatter(x_vals, y_vals, s=0.3, alpha=0.4);
         plt.title("Model 2 Residual Plot (Validation Dataset)")
         plt.ylabel("Rediduals (Log Sale Price)")
         plt.xlabel("Log Sale Price")
         plt.axhline(y = 0, color = 'r', linestyle = '-');
```


Model 2 Residual Plot (Validation Dataset)
Rediduals (Log Sale Price)
Log Sale Price
2
0
−2
−4
−6
6
8
10
12
14
16

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
```

My model creates a scenario like the one in **OPTION C PLOT 2** is more useful for detecting systematic overvaluing of inexpensive properties as it illustrates the distribution of "overvaluing" as it varies over home

price. It is clear from the graph that homes whith lower value are overvalued at a much higer rate than homes of higher value. This fact is less clear from the plot on the left as RMSE hides the direction of the error, and as such would be better suited to assesing the accuracy of predicions at each price band as opposed to a metric like “overvaluing” where the sign of each error is relevant.

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.

What does the residual mean for an individual homeowner? The residual represents the amount that the assesor's office estimate differed from the "true value" of their home (as determined by what it would/did sell for if sold within the year). When the **residual is positive**, it means that the assesor valued the house at a price which was lower than what the house was worth. If the **residual is negative** it means that the assesor valued the house at a price which was higher than it was actually worth. **How does the residual affect them in terms of property taxes?** The residual multiplied by the effective tax rate multiplied by -1 represents the increase or decrease in the property tax laibility of a given homeowner. For a person whose home was overvalued, their residual is negative, so they owe some amount more money to the government than they truly should given the property own. This might look like somone whose home is worth 120,000 being valued at 200,000 by the CCAO, leading to a residual of -80,000. In this case they owe (-1) * (-80,000) * (the tax rate) more in dollars. If the tax rate was 1%, they would owe 800 dollars more than they should at the end of the year. A non-insignificant burden for a lot of people. If the residual were much higher, the consequences could be great.

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?

What makes a model's predictions of property values for tax assessment purposes "fair"? In a perfect world, the model used by the assesors office would be acurate in terms of rmse AND have its errors equally distributed over the range of home prices. I think of our model as our measurment instrument, so I don't think that the modeling stage is the best place to implement equity. I'd hope that we can generate an instrument which is accurate and is free from systematic error over its range of inputs. Then once we have a reliable instrument at our disposal, we could implement a progressive tax system at the policy level by explicitly designing the property tax rate to increase with either home value, net household income, or some other metric of financial security. In the real world, however, where we might be stuck with a flat property tax rate, then the implemataion changes aimed at creating equity of can unfortunately only come from the design of our instrument. In that case, special attention needs to be paid to the accuracy of predictions for which errors will cause the greatest harm to those who they effect. In this case, that means minimizing the residuals–specifically the negative residuals–for homes whose value is low, as it is likely that the increase in property tax liability caused by these errors will place a greater burden on those homeowners than a similar error would to a homeowner whose property was worth more. This may come at the expense of increased error at the higer range of home prices. Within reason, this is acceptable.