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 13 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 who would like to sell their houses would be interested in seeing high housing prices, as it means they could sell their houses for a higher price and earn more. Federal Tax Authorities, like the IRS, could also prefer higher housing prices since they could benefit from high prices through capital gains taxes. People in the market to buy a home would want to see housing prices be lower so that they can spend less money on a purchase.

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 seems to be the most unfair, as overvaluing inexpensive properties and undervaluing expensive properties seems to be something like taking more from the poor and less from the rich. However, all options seem to be somehow equally unfair as it all leads to some group of people being negatively impacted.

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 13 to answer this question.

The problem was that the earlier property tax system carried a regressive tax system that disadvantaged working-class, non-white homeowners. The primary cause for such was that the residential assessments for the households lack of transparency and corruption.

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.

Cook County's property tax system historically placed a disproportionate burden on non-white property owners through biased initial assessments and an inequitable appeals process. The tax system caused the properties of the non-white owners to be overvalued through incorrect assessments that led to a disproportionate tax burden. Moreover, while there was a property tax appeals system, the process was found to be unfair for non-white homeowners.

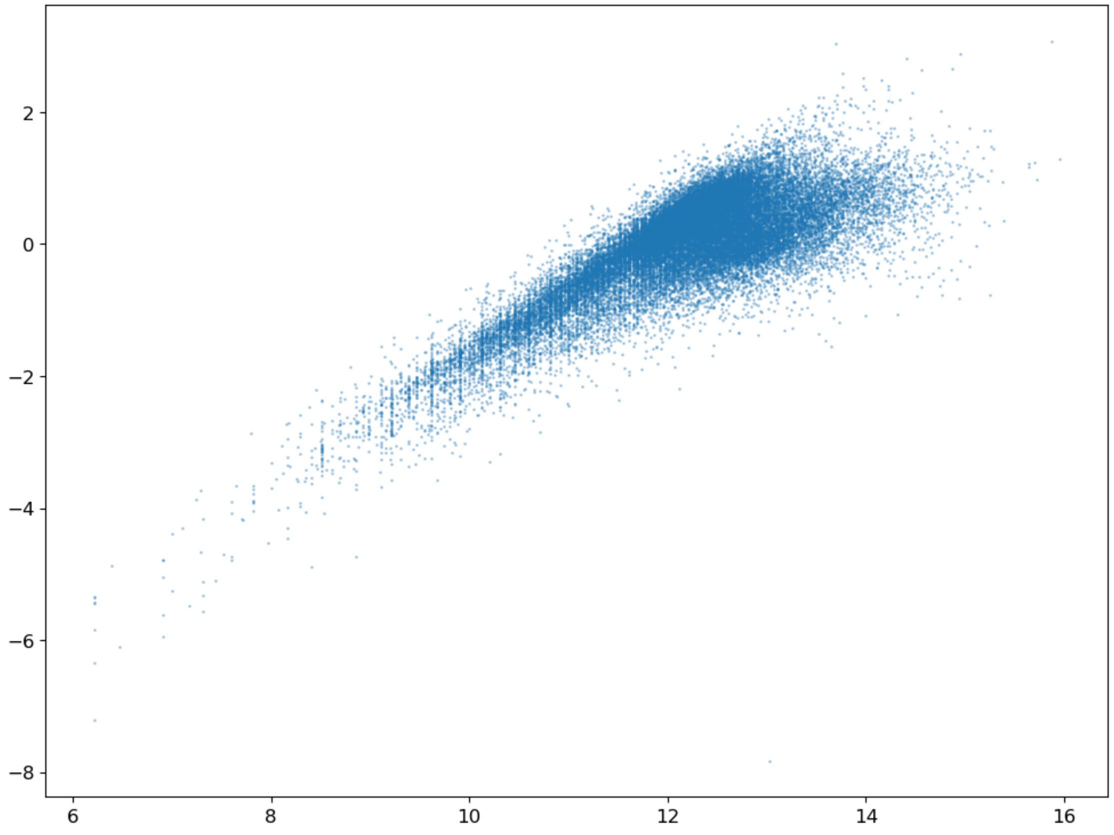
0.2 Question 4a

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 [24]: plt.scatter(x = Y_valid_m2 , y = Y_valid_m2 - Y_predicted_m2, alpha = 0.3, s = 1);
```



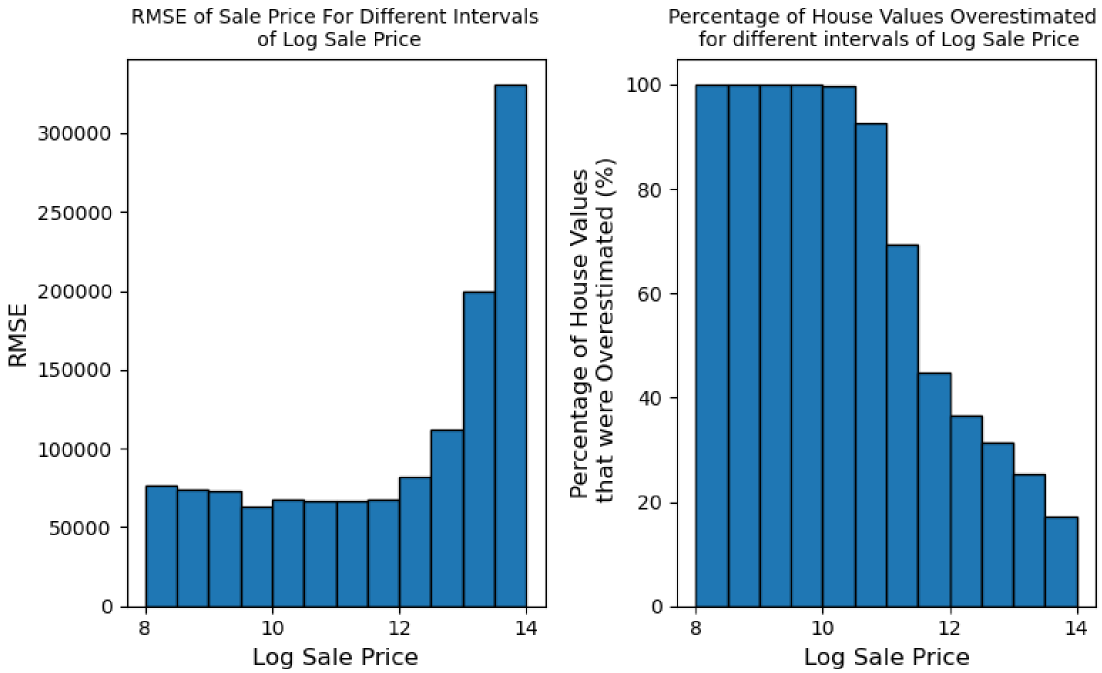
0.2.1 Question 6c

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

```
In [48]: # 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 of Sale Price 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 3 to 4 sentences.

The graph on the right would be more useful as we see that smaller log sale price gives higher percentage of house values that are overestimated. Therefore, as the log sale price increases, less houses' values are overestimated. This plot is closely aligned with what scenario 'C' is describing.

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 approximately 2 to 4 sentences.

For an individual homeowner, a positive residual means their home is being undervalued which means they would be paying less property tax than they should. A negative residual means their home is being overvalued and that the homeowner pays more property tax than they should.

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 would have a small RMSE by having the predicted property values similar to the true property value. However, it is also important to note that just having a small RMSE doesn't necessarily mean it is accurate, as having an even distribution over different intervals is also important. In a fair model, we would want to avoid having RMSE spikes on certain subgroups as the model's total RMSE might be “low,” but its accuracy for some datasets (like expensive homes) can be very low. Low RMSE doesn't mean high accuracy throughout the datasets, as RMSE measures average error, not distributed error. It is possible to have low average error but high distributed errors. This means that certain groups' inaccuracy might be ignored and cause unfairness when analyzing the data. Therefore, for a model's predictions of property values for tax assessment to be fair, we must have low RMSE with even distribution of high accuracy.

0.6 Question 8: Finding Better Metrics For Our Model

0.7 Question 8a

As discussed in Project A1, RMSE—while a widely used and powerful error metric introduced in earlier lectures—is not always the most appropriate choice when evaluating the *fairness* of a property appraisal system. In Question 7, you already encountered some of RMSE's limitations, particularly its tendency to disproportionately emphasize errors in high-priced properties due to the squaring of residuals.

In this question, rather than relying on RMSE, we will train and evaluate our model using another custom fairness metric. Specifically, we will examine the **Mean Absolute Percentage Error (MAPE)**, which measures the average error as a percentage of the true value. This allows us to better assess whether the model is making relatively fair predictions across different segments of the housing market.

In the code cells that follow, we'll explore how MAPE varies across price ranges—comparing the model's relative performance on inexpensive versus expensive housing. This helps identify whether the model systematically favors or disadvantages certain groups of properties based on their price.

Note: You'll notice that we're no longer using `lm.LinearRegression()`, but instead have started using `minimize`. This is because `scikit-learn` does not allow customization of the loss function in their standard linear regression model. The approach shown below provides an equivalent way to train a linear regression model using a custom error metric.

Warning: These cells take quite a long time to run. Please be patient, wait, and avoid restarting the kernel or rerunning these cells more than necessary.

```
In [49]: data = pd.read_csv("cook_county_train.csv", index_col='Unnamed: 0')
         trainX, trainY = feature_engine_final(data)
```

```
In [50]: # X is the design matrix (including bias column), y is the vector of true outputs, theta is th
         def mape(theta, X, y):
             y_pred = X @ theta  # compute predicted values using linear combination of features
             percentage_error = np.abs((y - y_pred) / y)  # calculate element-wise percentage errors
             return np.mean(percentage_error)  # return the MAPE
```

```
In [ ]: from scipy.optimize import minimize

        # Add bias (intercept) column to the design matrix
        trainX_with_bias = np.column_stack([np.ones(trainX.shape[0]), trainX])

        # Initialize parameter vector with zeros
```

```
theta_0 = np.zeros(trainX_with_bias.shape[1])

# Use scipy's minimize to find weights that minimize MAPE
res = minimize(mape, theta_0, args=(trainX_with_bias, trainY), method='BFGS')

# Optimal weights after training
theta_opt = res.x

theta_opt
```

```
In [ ]: new_preds_df = pd.DataFrame({
            'True Log Sale Price'     : trainY,
            'Predicted Log Sale Price': trainX_with_bias @ theta_opt,
            'True Sale Price'         : np.e ** trainY,
            'Predicted Sale Price'    : np.e ** (trainX_with_bias @ theta_opt)
        })

        plt.figure(figsize=(8, 5))
        plt.subplot(1, 2, 1)

        mape_values = []
        for i in np.arange(8, 14, 0.5):
            mape_values.append(mape_interval(new_preds_df, i, i + 0.5))

        plt.bar(x=np.arange(8.25, 14.25, 0.5), height=mape_values, edgecolor='black', width=0.5)
        plt.title('MAPE of Sale Price Across\n Log Sale Price Intervals', fontsize=10)
        plt.xlabel('Log Sale Price')
        plt.ylabel('Mean Absolute Percentage Error (MAPE)')
        plt.xticks(fontsize=10)
        plt.yticks(fontsize=10);
```

What can you infer from this graph using our new custom error metric? Write your findings in the space below in three to five sentences. Consider addressing the following points:

- How does this metric differ from RMSE?
- What is the purpose of this custom error function?
- How does it relate to the idea of underpriced expensive housing mentioned in Question 6a?
- Why would this potentially be better than the RMSE in terms of the CCAO dataset?

The metric differs in a way that the metric is a relative percentage error rather than an absolute dollar error. This custom error function shows us the proportion of inaccuracy. It shows the issue of underpriced expensive housing by showing that while their dollar error was high for the RMSE plot, their percentage error is tiny. This metric seems to be better for the CCAO dataset because it reveals the proportional fairness issue of different properties' sale prices.