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.

People who are interested in answering the question are architects, landlords, and real estate. These parties of people are related to land and construction, which are highly related to the price of the house. In terms of low housing prices, architects do not care about the price, but landlords and real estate care about high price of housing.

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.

Scenario A is unfair because the value of a home is assessed on the location and the size of the land. If the price is higher than the market value, the homeowner may end up paying more rent fees than they should. This could disproportionately affect homeowners in declining markets or areas with inconsistent property valuations.

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 central problem with the earlier property tax system in Cook County is that standards of accuracy and fairness are not met and are often overvalued. In contrast, expensive houses are often undervalued, meaning that poor people are getting taxed more than they need to and rich people are getting taxed less. This is applied to low accuracy of the coefficient of dispersion and price-related difference.

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 real estate industry has objective rating systems for the value of the property which took racial value that non-white property owners due to systemic inequities in property assessments.

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).

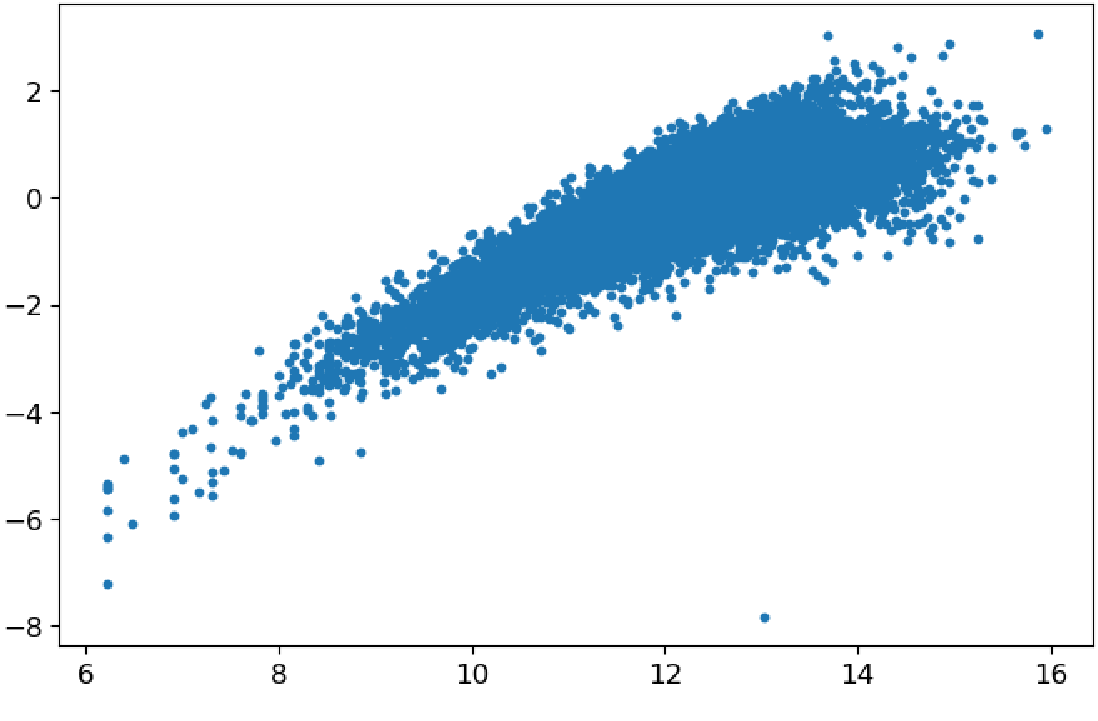
In the cell below, use `plt.scatter` (documentation) to plot the **model 2** residuals of `Log Sale Price` versus the original `Log Sale Price` values. For this part, you only need to plot the residuals and outcomes for the **validation data**.

- 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 [96]: residuals_m2 = Y_valid_m2 - Y_predicted_m2

         plt.figure(figsize=(8, 5))
         plt.scatter(Y_valid_m2, residuals_m2, s=10)
```

```
Out[96]: <matplotlib.collections.PathCollection at 0x7c48745a2d50>
```



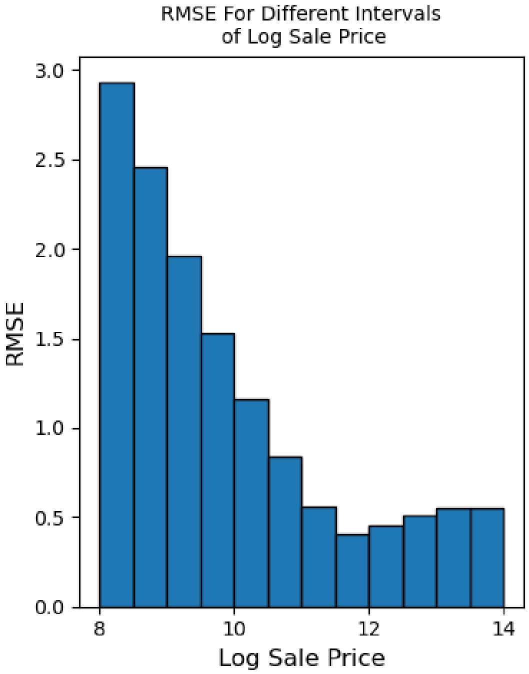
0.2.1 Question 6c

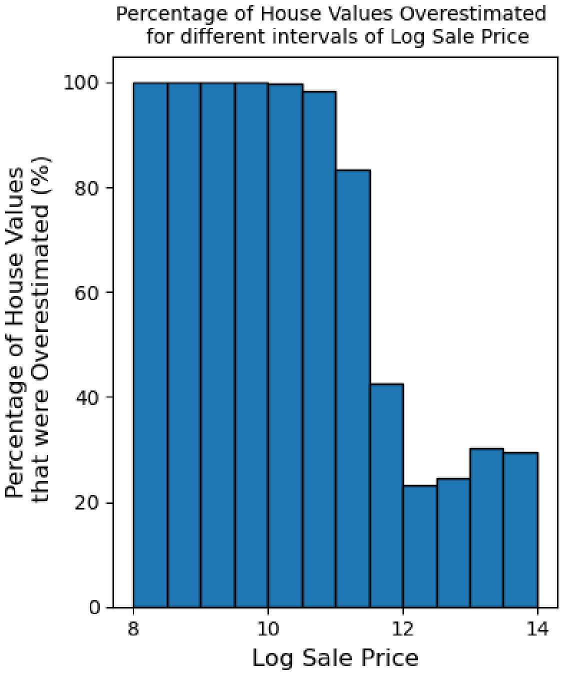
Using the functions above, we can generate visualizations of how the RMSE and proportion of overestimated houses vary for different intervals:

```
In [179]: # 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 above plot illustrates that the “RMSE for Different Intervals of Log Sales Price” tends to result in regressive taxation. This is evident as the graph shows a declining RMSE for higher sales prices of houses. From this, we can support Scenario C, as the RMSE is larger for houses with lower sales prices.

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.

To the individual homeowner, residual means show the difference between the official price and the predicted price, which is calculated by the prediction model and the actual value of the house. In terms of property taxes, a positive residual means that the predicted value is an underestimate, meaning that the homeowner pays less than they should have. A negative residual means the opposite, where the predicted value is an overestimate, meaning that the homeowner pays more than what they should have.

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.

To ensure that the model's property value predictions for tax assessment are fair, we must minimize the difference between the predicted price and the official value. A low RMSE indicates that the predictions are more accurate relative to the official values. As shown in the previous two questions, the gap between the predicted and official values tends to be larger (i.e., higher RMSE) for lower-priced properties. Therefore, minimizing RMSE is crucial for improving fairness in the model.