0.1 Question 0: Human Context and Ethics

0.1.1 Question 0a

"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 to be high or low.**

the owner of the house, the potential buyer of house, a real estate agency/a real estate agent who is looking to put the house on the market

---

0.1.2 Question 0b

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.

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.

C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive properties.

This could lead to a situation where homeowners with inexpensive properties are unfairly taxed while homeowners with expensive properties are under-taxed, creating a regressive taxation system that disadvantages the less affluent by making them pay more tax than the amount they should.

0.1.3 Question 0d

What were the central problems with the earlier property tax system in Cook County as reported by the Chicago Tribune ? And what were the primary causes of these problems? (Note: in addition to reading the paragraph above you will need to watch the lecture to answer this question)

Cook County failed to value homes accurately for a long period of time. The result of this was a property tax system that harmed the poor and helped the rich. The Tribune's investigation found that assessments had been inaccurate for many years, with high-priced homes undervalued and low-priced homes overvalued, leading to a visible divide racially. The primary causes of these problems were outdated assessment methods, a lack of transparency and oversight, and political pressure to keep taxes low. Cook County relied on research conducted by the Illinois Department of Revenue. But that research comes out years after property taxes are calculated and doesn't provide granular details on individual neighborhoods.

0.1.4 Question 0e

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 the assessments were inaccurate and discriminatory. The investigation by the Chicago Tribune found that high-priced homes, which were typically owned by white residents, were undervalued, while low-priced homes, which were typically owned by non-white residents, were overvalued. This resulted in non-white property owners paying more in taxes than white people, even if they had similar or lower property values.

Moreover, the investigation found that non-white property owners were more likely to appeal their assessments than white property owners, but were also less likely to have their appeals granted.

0.2 Question 2a

Without running any calculation or code, complete the following statement by filling in the blank with one of the comparators below:

$$\geq$$
$$\leq$$
$$=$$

Suppose we quantify the loss on our linear models using MSE (Mean Squared Error). Consider the training loss of the 1st model and the training loss of the 2nd model. We are guaranteed that:

$$\text{Training Loss of the 1st Model} \geq \text{Training Loss of the 2nd Model}$$

This is because, the second model takes into account more features (Log Building Square Feet) than the first model and hence the second model will be a better fit to the data, and thereby will result in a lesser training loss.

0.3 Question 3b

You should oberseve that θ_1 change from positive to negative when we introduce an additional feature in our 2nd model. Provide a reasoning why this may occur. **Hint:** which feature is more useful is predicting `Log Sale Price`?

In the 2nd model, we added the Building Square Feet feature in addition to the Bedrooms feature that was included in the 1st model as well. The change in the sign of the coefficient on the Bedrooms feature, i.e. 0_1, from positive in the 1st model to negative in the 2nd model, means that the effect of the Bedrooms feature on the Log Sale Price was overestimated in the 1st model when we did not account for the effect of Building Square Feet feature.

In other words, the Building Square Feet feature was more useful in predicting Sale Price than the Bedrooms feature.

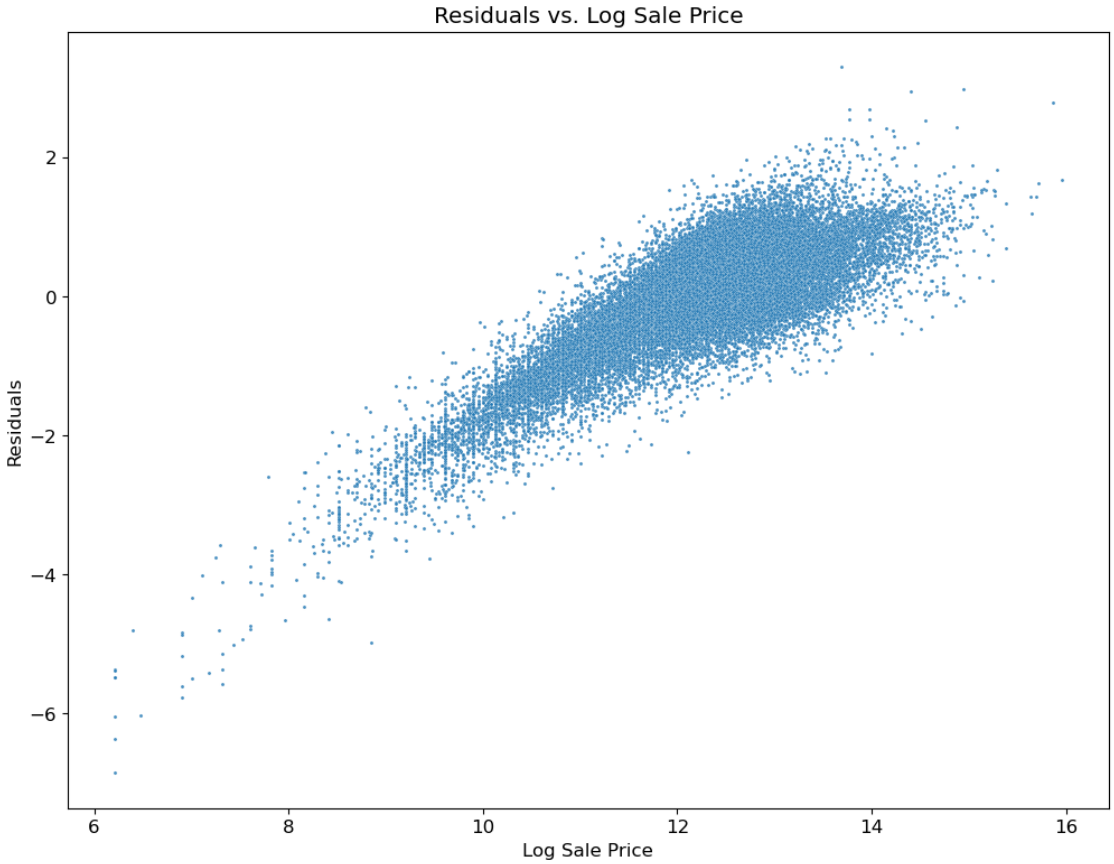
0.4 Question 3c

Another way of understanding the performance (and appropriateness) of a model is through a plot of the residuals versus the observations.

In the cell below, use `plt.scatter` to plot the residuals from predicting `Log Sale Price` using **only the 2nd model** against the original `Log Sale Price` for the **validation data**. With a data size this large, it is diffult 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 [25]: res= y_valid_m2-y_predicted_m2
         sns.scatterplot(x=y_valid_m2, y=res, alpha=0.8, s=5)
         plt.xlabel('Log Sale Price')
         plt.ylabel('Residuals')
         plt.title('Residuals vs. Log Sale Price')
```

```
Out[25]: Text(0.5, 1.0, 'Residuals vs. Log Sale Price')
```


Residuals vs. Log Sale Price
Residuals
2
0
−2
−4
−6
6
8
10
12
14
16
Log Sale Price

0.5 Question 5

In building your model in question 4, what different models have you tried? What worked and what did not? Brief discuss your modeling process.

Note: We are not looking for a single correct answer. Explain what you did in question 4 and you will get point.

First to select the best features, I started constructing plots for log sale price vs the features of interest, i.e., "Bedrooms", "Building Square Feet", "Estimate (Building)", and "Age". However, in the plots of "Building Square Feet" and "Estimate (Building)", there were bulges and so, I transformed those features to logs so that the relationship becomes linear. After selecting the features, I removed outliers from all of them, so that the model is more generalised and works well. Before that, an important step was to replace all the infinity and 0 values from the features with Nan values and then I replaced all those with the mdeian of that particular feature since log of these vales would result in Nan values, which does not work well with Linear Regression. The reason I chose median was to avoid the effect of outliers. After which, I created the training and testing sets respectively.

0.6 Question 6 Evaluating Model in Context

0.7 Question 6a

When evaluating your model, we used root mean squared error. In the context of estimating the value of houses, what does residual mean for an individual homeowner? How does it affect them in terms of property taxes? Discuss the cases where residual is positive and negative separately.

In the context of estimating the value of houses, a residual refers to the difference between the predicted sale price of a home and the actual sale price of the home. Specifically, the residual is the difference between the estimated sale price calculated by the model and the actual sale price of the home.

A positive residual means that the actual sale price of the home was higher than the estimated value predicted by the model. This means that the homeowner may have paid more in property taxes than they should have, based on the estimated sale price of their home.

On the other hand, a negative residual means that the actual sale price of the home was lower than the estimated value predicted by the model. This means that the homeowner may have paid less in property taxes than they should have, based on the predicted/estimated sale price of their home.

0.8 Question 6b

In your own words, describe how you would define fairness in property assessments and taxes.

According to me, fairness in property assessments and taxes should go beyond just technical measures of accuracy. It should take into account historical contexts and larger social and economic trends that have resulted in systemic inequities in property assessments. A fair system should aim to correct these inequities, particularly for marginalized communities such as Black and Hispanic neighborhoods that have consistently faced overvaluation resulting in higher property taxes.

Fairness should also consider the broader implications of property assessments on the community, particularly on issues of housing equity and social justice.

0.9 Question 6c

Take a look at the Residential Automated Valuation Model files under the Models subgroup in the CCAO's GitLab. Without directly looking at any code, do you feel that the documentation sufficiently explains how the residential valuation model works? Which part(s) of the documentation might be difficult for nontechnical audiences to understand?

The documentations consists of sufficient explanation for the Residential Automated Valuation Model which a detailed explanation of the features, the training set and the process. Along with that, it also includes reasons for model selection. However, I believe that it is written in a technical language that may be challenging for non-technical audiences to understand. The documentation uses several technical terms such as machine learning algorithms, feature engineering, and model validation that may be unfamiliar to non-technical audiences. In particular, the section with Hyperparameter selection would be tough for non-technical audiences to understand and comprehend.

Overall, while the documentation appears to provide a comprehensive explanation of the residential valuation model, it may require a certain level of technical knowledge to fully understand.