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.**

Homeowners will be interested in knowing how much their house and other properties are worth, favoring if their own housing price is high due to home price appreciation. But this can also mean high property tax rates, in which lower housing prices would be ideally favored in a fair taxing system. Real estate agents would also be interested in high housing prices so that they can generate a profitable income on their properties when the sell/advertise the home. On the other hand, prospective house buyers will be interested in low prices for buying a place of residence. An HOA would also like to see low housing prices as its residents are made up of the resident themselves paying rent/mortgages.

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.

Scenario C seems unfair to me because is disparagingly disadvantages people of lower income and socioeconomic backgrounds. These individuals will not be able to acquire property at a fair price and will likely experience financial strain and homelessness while those who can afford “expensive properties” enjoy somewhat of a discount on acquiring propoerties. This scenario will only concentrate purchasing power amongst those from affluent backgrounds.

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)

The central problem was that Cook County's property tax system created an unequal financial burden on residents, giving large financial breaks to white homeowners who are well-off while discriminating and overcharging those in lower income brackets, particularly working class people living in minority communities. The problem stems from the fact that these rates became skewed in favor of wealthier residents and went unchecked for for fairness and accuracy. The method to determine tax rates used old, faulty models and computer programs to value residential properties in 2015.

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?

White areas of the city where racial gentrification had taken place saw values come in low, while homes just outside those neighborhoods populated by non-white property owners were more likely to be overvalued. The effective tax rate in Cook County differed widely by neighborhood between 2009 and 2015, even though the rate should have been roughly equal for everyone. In addition to this, residents of these areas disproportionately experienced redlining, a discriminatory practice that consisted of the systematic denial of housing services such as mortgages and insurance loans.

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:

Training Loss of the 1st Model_______Training Loss of the 2nd Model

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

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`?

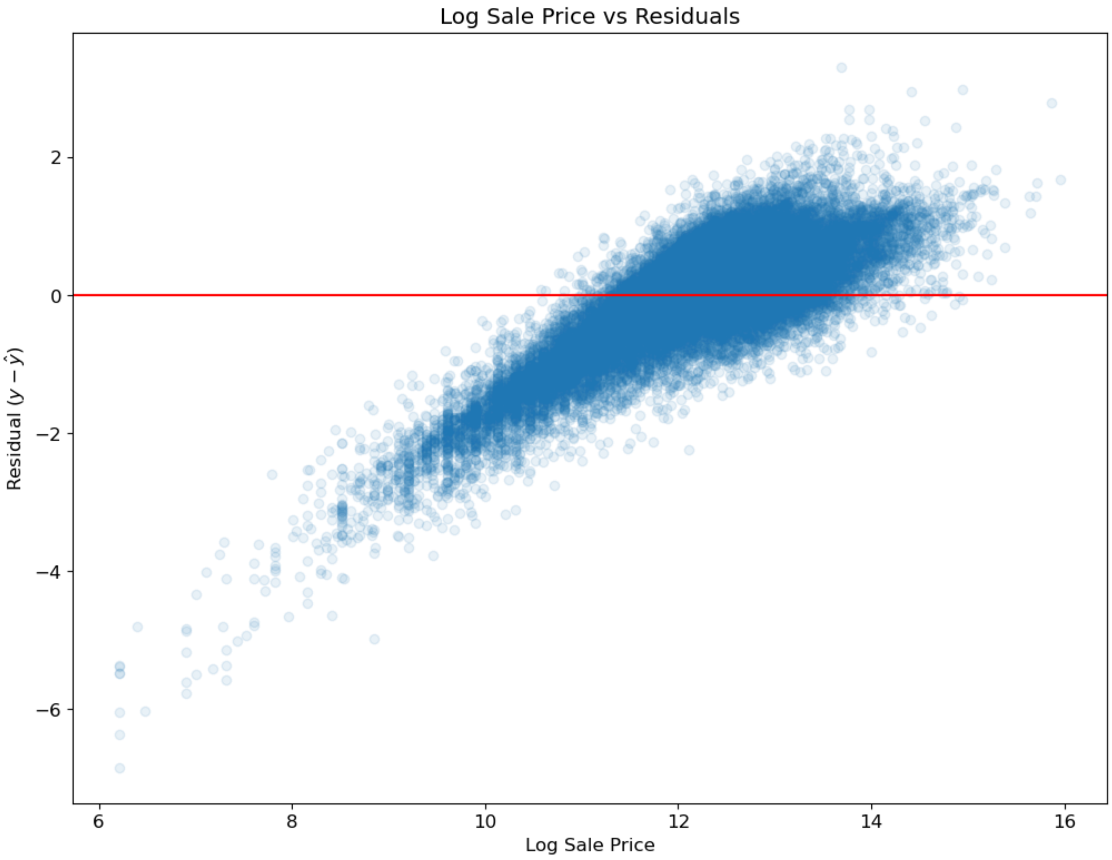
For our second model, we are using more features to predict Log Sale Price. By adding Log Building Square Feet, we recognize that it is a better predictor at predicting Log Sale Price compared to Bedrooms because it allows for more variability in our model than just Bedrooms and θ_0. To account for overpredicting, we can observe that θ_1 changes from positive to negative to minimize our RMSE.

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 [31]: plt.scatter(y_valid_m2, y_valid_m2 - y_predicted_m2, alpha=0.1)
         plt.ylabel("Residual $(y - \hat{y})$")
         plt.xlabel("Log Sale Price")
         plt.title("Log Sale Price vs Residuals")
         plt.axhline(y = 0, color='r');
```



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.

I started by feature selection by calculating the correlation between a selected features and Log Sale Price to see if there existed somewhat of a linear relationship between the two variables. I would then transform the selected feature by squaring, logging, or square rooting it. This was effective in that I was able to more accurately fix any bulges in the data and weigh appropriately. Visualizing the relationship between the variables proved to be less effective as such a large dataset resulted in overplotting despite opacity adjustments and transformations. I then built my feature engineering pipeline by removing extreme outliers of Sale Price to better generalize my linear regression model. In addition to the features I selected initially, I also created variables using One Hot Encoding for roof material, garage, sale month, and fireplaces . From this, I observed that OHE worked better price related qualitative ordinal variables as there is somewhat of a rank that can be fit to scale the variables. I also made use of the Description column and extracted data involving the property rooms, bedrooms, bathrooms, and stories using RegEx.

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.

The residual is a measure of the actual price of the house - the predicted price of the house, or the difference in valuation for their home. For an individual homeowner, a positive residual means that the actual price is more than the predicted price of the house. As a result, the homeowner is subjected to pay a lower property tax than what they should actually pay for since there property is valued for less. On the other hand, a negative residual indiciates that the actual price is less than the predicted price of the house. As a result of this, the homeowner is subjected to pay a higher property tax than what they should actually pay for since their property is overvalued.

0.8 Question 6b

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

In the case of property assessments and taxes, I think fairness takes on a both a computated and ethical stance. I believe the overall goal is to minimize our metric of loss in order to best produce accurate predictions for housing prices based on relative properties with similar characteristics. But we also have to recognize that there exists systematic biases in place that undermines marginalized communities due to historical housing discrimination, gentrification, and other factors. In other words, while the main goal is to promote an equal assessment of statistical modeling, social equity, which can take the form of reparations and subsidies, must be taken into consideration since not everyone is impacted to the same degree within systematic institutions.

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?

I think that the documentation does a good job explaining how the residential valuation model works because it outlines how the model estimates the sale price in the "How it Works" section. It also makes transparent what features are used to evaluate the model in the "Features Used" which allows people to understand from what data is used to determine a property's assessed value. I also appreciate how it outlines ongoing issues surrounding data quality and lack of property characteristics. For non-technical audiences, I feel like information regarding the methodology of cross-validation and hyperparameter selection may be difficult to interpret.