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.**

1. Potential home buyers: As a potential homeowner, the person is interested in buying a house with cheaper price so naturally the person would like to see house price going low. But the person is likely also interested in a safer neighborhood with good community and good public schools that will drive the house price higher, so it’s a comprehensive decision.
2. The internal revenue services (IRS) or assessor’s office: The government is interested in collecting tax to keep the government operation going. They would like to understand how much tax they can collect from the house sale and the property taxes.
3. The mortgage providers would like to understand the local markets and evaluate the investment. The house prices will affect their rates and mortgage numbers, etc. They need to make decisions with all information, with housing price being a very important factor.

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.

I'm mostly unhappy with: C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive properties.

I think rich people should naturally pay more taxes to support the society. If the assessment systematically overvalues inexpensive properties, it's exploiting poor people which might lead to homelessness.

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)

Real estate has been a key motor of racial inequality in modern US History. Racial segregation has been adopted by federal policy and developed by real estate professionals, hence leading to appraisal of property values which encode race as a factor of valuation. As a result, neighborhoods with more whites are constantly undervalued in Cook County and neighborhoods with less whites are overvalued.

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?

As a result of the racial inequality and unfair appraisal, neighborhoods with more whites are constantly undervalued in Cook County and neighborhoods with less whites are overvalued. In addition, neighborhoods with more whites tend to hire professionals to appeal whereas the less white neighborhoods don't have access to services like that. It's part of a deeper, institutional pattern, potential corruption.

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

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`?

This is likely because the inclusion of the log of the building square feet as a feature has changed the relationship between the number of bedrooms and the sale price.

In the first model, which only includes the number of bedrooms as a feature, the parameter _1 represents the change in the log sale price for a one-unit increase in the number of bedrooms, holding all other variables constant. A positive value of _1 means that as the number of bedrooms increases, the log sale price also increases.

In the second model, which includes both the number of bedrooms and the log of the building square feet as features, the parameter _1 represents the change in the log sale price for a one-unit increase in the number of bedrooms, while holding the log of the building square feet constant. The negative value of _1 in this model means that as the number of bedrooms increases, the log sale price decreases, holding the log of the building square feet constant.

The actual values of _1 and _2 provide an indication of the strength of the association between a predictor variable and the outcome variable. However, the magnitude of a coefficient alone is not sufficient to determine which predictor variable is more useful in predicting the outcome variable.

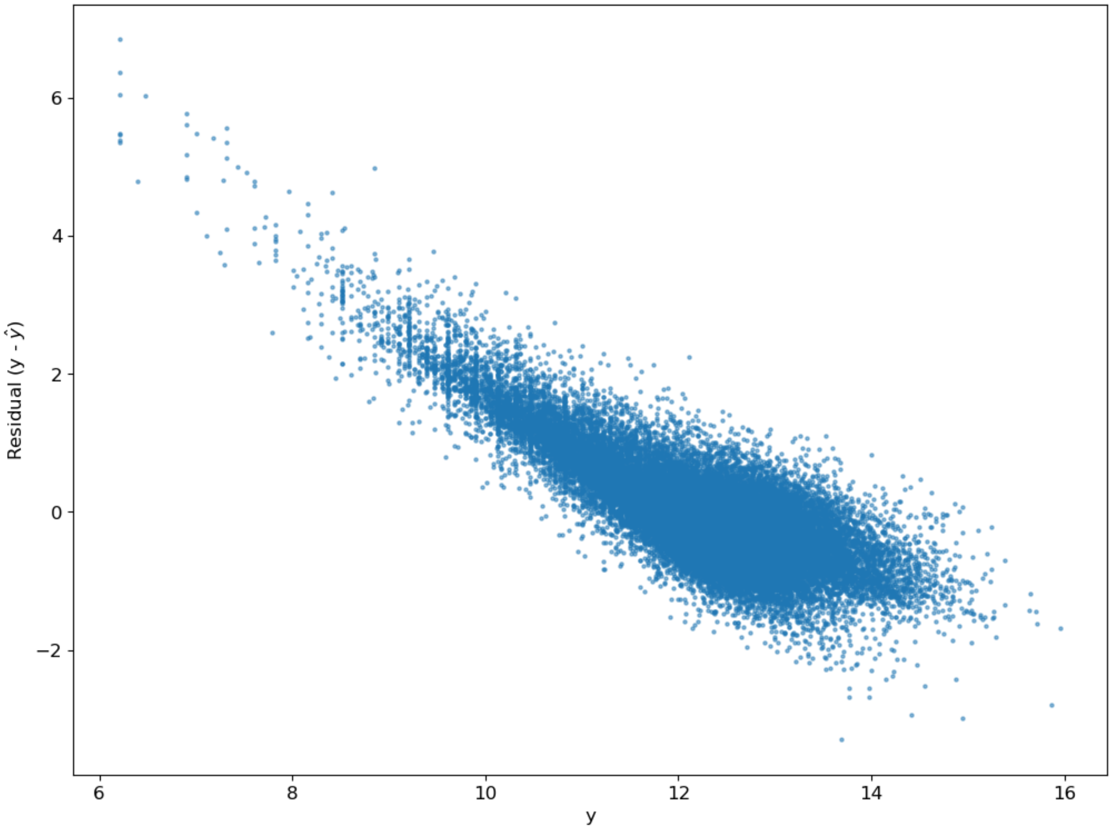
0.4 Question 3c

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 [36]: plt.scatter(y_valid_m2, y_predicted_m2 - y_valid_m2, s = 5, alpha = 0.5)
         plt.xlabel('y')
         plt.ylabel('Residual (y - $\hat{y}$)')
```

```
Out[36]: Text(0, 0.5, 'Residual (y - $\\hat{y}$)')
```



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.

In question 4, I first picked out 27 features intuitively from the codebook.txt descriptions. Then I plotted them out against Log Sale Price to look for any trend. I picked several numerical features and several categorical features. For the numerical features I basically did log transformation to all of them. For categorical features I did OneHotEncoder. The I combined the features into a new dataframe as feature engineered datasets for linear regression models. The end result was quite disappointing as I managed to reduce RMSE from 380000 to 290000 but could never push it down any further. I also couldn't debug the Sale Price column missing issue.

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.

Residual means how much is the property tax overvalued or undervalued. A positive residual, meaning actual price - predicted price > 0, hence the house is undervalued and the homwowner should pay more tax. A negative residual means the homeowner should pay less tax.

0.8 Question 6b

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

I think fairness means after taking all factors into account, the predicted house price reflects the house's market price and no one is overpaying or underpaying the property tax.

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 it listed all necessary calculations as how CCAO reached their new model. I think the model itself that includes all the codes should be the most difficult for nontechnical audiences.