0.1 Question 0

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 value be high or low.

- home owners: seeing the house value high
- realestate agents: they might have an interest in seeing the value be high or low depending on their customers’ interests and what they are looking for
- local government: seeing the house value high for the purpose of property taxes

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.

- B: this is not an ideal situation for a homeowner as he/she loses money what they initially paid for
- C and D: this is not appropriate for an assessment process that either undervalues or overvalue properties whether they are expensive or not as it can lead to inflation and bad for economy and low income families.

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)

For those who own houses with lower property value (homes in working-class neighborhoods), when their homes were overvalued than the true market value, they are more likely to pay higher property tax rates than those with higher property value (wealthy homeowners). The primary causes of these problems are the inefficiency in these property tax assessment using the old model that would overvalue inexpensive properties.

0.1.4 Question 0e

In addition to being regressive, why did the property tax system in Cook County place a disproportionate tax burden on non-white property owners?

Although there is the appeals process for house valuation, most people whose were overvalued do not have resources to go through this appeal process. Therefore, they will still be paying high propety tax rates.

0.2 Question 2

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

$$\geq$$

0.3 Question 6

Let's compare the actual parameters (θ_0 and θ_1) from both of our models. As a quick reminder,

for the 1st model,

$$\text{Log Sale Price} = \theta_0 + \theta_1 \cdot (\text{Bedrooms})$$

for the 2nd model,

$$\text{Log Sale Price} = \theta_0 + \theta_1 \cdot (\text{Bedrooms}) + \theta_2 \cdot (\text{Log Building Square Feet})$$

Run the following cell and compare the values of θ_1 from both models. Why does θ_1 change from positive to negative when we introduce an additional feature in our 2nd model?

After adding Log Building Square feet in our 2nd model, Bedrooms is less useful than Log Building Square feet to predict sale price.

```
In [30]: # Parameters from 1st model
         theta0_m1 = linear_model_m1.intercept_
         theta1_m1 = linear_model_m1.coef_[0]

         # Parameters from 2nd model
         theta0_m2 = linear_model_m2.intercept_
         theta1_m2, theta2_m2 = linear_model_m2.coef_

         print("1st Model\n 0: {}\n 1: {}".format(theta0_m1, theta1_m1))
         print("2nd Model\n 0: {}\n 1: {}\n 2: {}".format(theta0_m2, theta1_m2, theta2_m2))
```

```
1st Model
 0: 10.571725401040084
 1: 0.4969197463141442
2nd Model
 0: 1.9339633173823696
 1: -0.030647249803554506
 2: 1.4170991378689644
```

0.4 Question 7

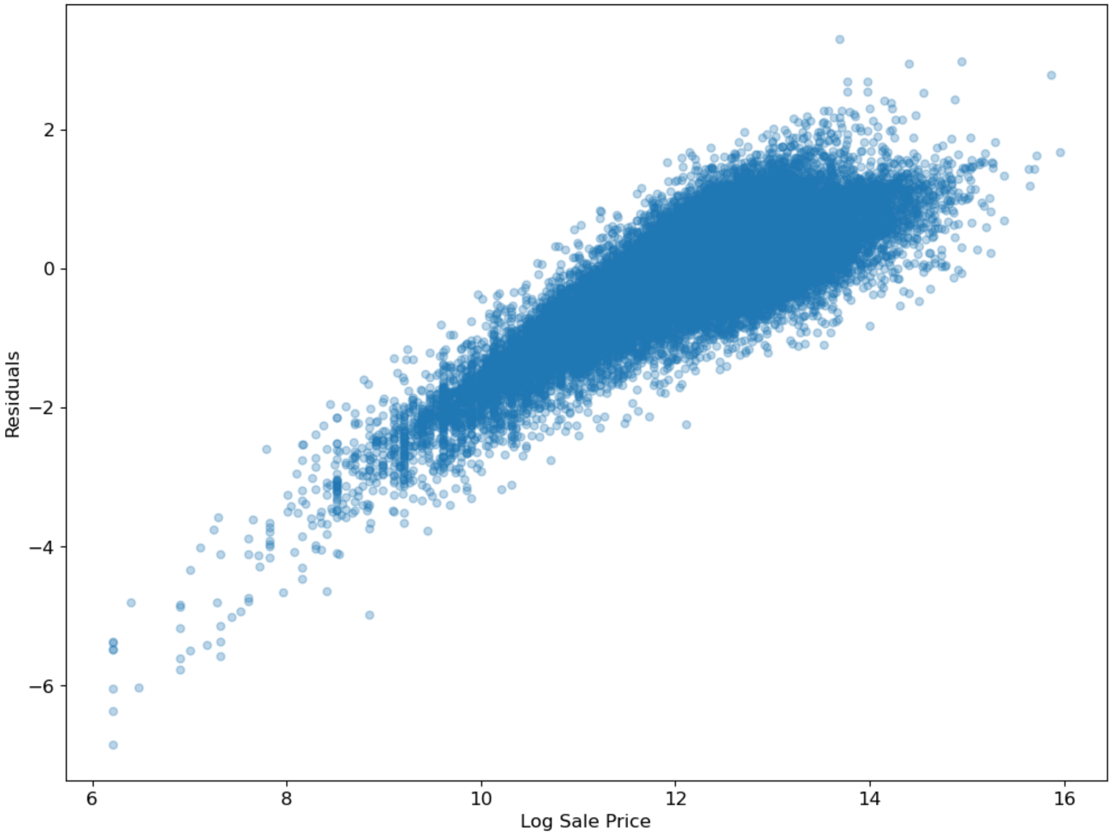
0.4.1 Question 7a

Another way of understanding the performance (and appropriateness) of a model is through a plot of the model 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 **test data**. You should also ensure that the dot size and opacity in the scatter plot are set appropriately to reduce the impact of overplotting.

```
In [31]: plt.scatter(y_test_m2, y_test_m2-y_predicted_m2,alpha=0.3, s=25)
         plt.xlabel("Log Sale Price")
         plt.ylabel("Residuals")
```

```
Out[31]: Text(0, 0.5, 'Residuals')
```



0.5 Question 9

In building your model in question 8, what different models have you tried? What worked and what did not? Brief discuss your modeling process.

Note: We are looking for a single correct answer. Explain what you did in question 8 and you will get point.

The following are not changed in the model selection: - Predictor Variables: Bedrooms, Building Square Feet - Remove outliers in sale price

3rd model: - Additional Predictor Variables: - Roof Materials - One-hot encode roof materials - RMSE: 312835

4th model: - Additional Predictor Variables: - Age - Plot a scatter plot and find negative association between Age and Log Sale Price - Plot box plot and distribution plot (using plot_categorical method) between Age and Log Sale Price to look at counts and outliers - Drop property records with age over 100 years old - Log the age - I tried one-hot encoding on age and got a better result than log age, but was afriad of overfitting the model, so kept the log age. - RMSE: 268682

I tried adding more features (wall materials, basements, porch, garage indicator, sale month of year) but results were not favorable. (The decrease in RMSE in very low due to the lack of variability in these features.)

5th model: - Additional Predictor Variables: - Fireplaces, Garage 1 Size - Plot a box and distribution plot between fireplaces and log sale price to look at the variability and their distribution - Again, not much decrease in training error - RMSE: 264754

6th model: - Additional Predictor Variables: - Property class, neighborhood code, town code - Plot a box and distribution plot - compare the two code features - Although neighborhood code seemed to be a good choice to predict the sale price, the town code gives a decent decrease in training error than I expected. - One hot encode the two categorical variables
- RMSE: 261164 (with neighborhood code) - RMSE: 242776 (with town code; much less training error)

7th model: - Additional Predictor Variables: - pure market filter (deterministic factor) - This removes a lot of outliers in sale price - RMSE: 123487

0.6 Question 10

When evaluating your model, we used root mean squared error. In the context of estimating the value of houses, what does error mean for an individual homeowner? How does it affect them in terms of property taxes?

If the error is high, we are either over-estimating or under-estimating the value of houses, and that has effects on low-income families. These houseowners have to pay high property taxes rates if their houses are overvalued than they would be sold for.

In the case of the Cook County Assessor's Office, Chief Data Officer Rob Ross states that fair property tax rates are contingent on whether property values are assessed accurately - that they're valued at what they're worth, relative to properties with similar characteristics. This implies that having a more accurate model results in fairer assessments. The goal of the property assessment process for the CCAO, then, is to be as accurate as possible.

When the use of algorithms and statistical modeling has real-world consequences, we often refer to the idea of fairness as a measurement of how socially responsible our work is. But fairness is incredibly multifaceted: Is a fair model one that minimizes loss - one that generates accurate results? Is it one that utilizes "unbiased" data? Or is fairness a broader goal that takes historical contexts into account?

These approaches to fairness are not mutually exclusive. If we look beyond error functions and technical measures of accuracy, we'd not only consider *individual* cases of fairness, but also what fairness - and justice - means to marginalized communities on a broader scale. We'd ask: What does it mean when homes in predominantly Black and Hispanic communities in Cook County are consistently overvalued, resulting in proportionally higher property taxes? When the white neighborhoods in Cook County are consistently undervalued, resulting in proportionally lower property taxes?

Having "accurate" predictions doesn't necessarily address larger historical trends and inequities, and fairness in property assessments in taxes works beyond the CCAO's valuation model. Disassociating accurate predictions from a fair system is vital to approaching justice at multiple levels. Take Evanston, IL - a suburb in Cook County - as an example of housing equity beyond just improving a property valuation model: Their City Council members recently approved reparations for African American residents.

0.7 Question 11

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

I think that the property assessments and taxes will evaluate the value of a house without discriminations or bias and eliminate regressivity while taking account of the socioeconomic status of homeowners.

0.8 Question 12

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 documentation has done its job well in providing the process of the residential valuation model, but model selection, hyparameter selection, and framework selection might be difficult for nontechnical audiences to understand as they contain high-level knowledge of developing machine learning model.