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.

There are many groups and organizations that might be interested in how much a house is worth. One example can be a company like Zillow that is a real estate marketplace. Another example would be a company located in the area that is paying workers extra to relocate to said area like Apple or Amazon. Finally, the local government in an area likely wants to have as much information about its city as possible, meaning they would be interested in whether housing prices in their domain would be high or low.

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 see C and D as unfair because they are systemic processes that directly affect owners and buyers of houses alike. For example, having cheap houses priced higher than they actually are would make it harder for normal citizens to live in houses they like. Similarly, having inexpensive properties priced lower than they actually are would make it so that normal citizens who invested in those properties lose large amounts of money. As for A and B, I don't really see them as unfair because pricing your house differently than you imagine can just be a mistake. It is very important however that houses are priced accurately and unbiasedly.

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 Cook County property tax system found out that as income level rose among residents in the county, their effective tax rate went down. For example, someone woh made 50,000 a year would effectively be paying more in taxes than someone who made 200,000. There were many reasons for this, but the main cause of this tax issue was that houses were not being priced correctly, leading to cheaper homes being priced higher than they should and more expensive homes being priced lower than they should. This means that people who lived in cheaper homes would proportionally pay more in taxes than people who lived in more expensive homes. The Chicago Tribune also found out that white homeowners had their houses priced cheaper than they should have been, adding to the issue presented earlier.

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?

In Cook County, a study done by the Chicago Tribune found that white homeowners had their houses priced lower than they should have been by around ~5%, while non-white homeowners had their houses priced higher than they should've been by as much as ~10%. This means that non-white property owners had to pay proportionally larger taxes than white property owners because of how property tax works in the county.

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?

```
In [22]: # 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
```

When we introduce an additional feature in our 2nd model, our θ_2 is likely compensating in our model's Log Sale Price. When we add Log Building Square Feet into our model, having a large Log Building Square Feet and a small amount of bedrooms likely makes the house more expensive than if the house has a large Log Building Square Feet and a large amount of bedrooms. This sentiment is reflected in our model. Essentially, because the Log Building Square Feet feature of our model is so important, it means that for any value of Log Building Square Feet that we have, when we add more bedrooms our house value will be less expensive than when there are fewer bedrooms. Therefore θ_1 is negative when we add Log Building Square Feet.

0.4 Question 7

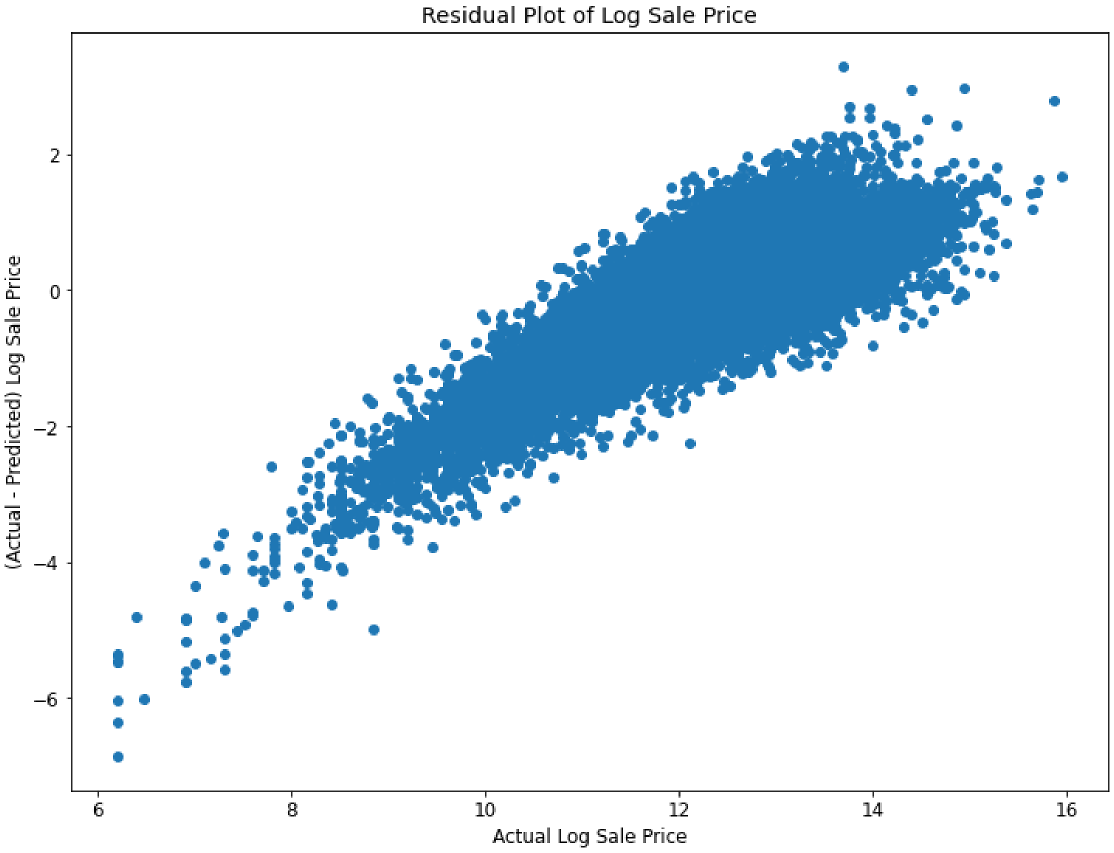
0.4.1 Question 7a

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 [23]: plt.scatter(y_test_m2, y_test_m2 - y_predicted_m2)
         plt.xlabel("Actual Log Sale Price")
         plt.ylabel("(Actual - Predicted) Log Sale Price")
         plt.title("Residual Plot of Log Sale Price")
```

```
Out[23]: Text(0.5, 1.0, 'Residual Plot of Log Sale Price')
```



0.5 Question 9

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?

The error in our model for the average homeowner is basically how different our model is going to be from the actual price of a home. If we had a perfect model, it would perfectly predict the prices of homes using the information that we give it, however since it is near impossible for a model like this to be perfect, root mean squared error represents on average how off our guesses are going to be from the actual price of a home.

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.

0.6 Question 10

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

I would define fairness in property assessments as having your home priced as acurately as possible, with the differences in predicted vs actual home assessments being only due to random chance rather than some sort of systemic bias. With taxes, I would define fairness as everyone paying their fair share that they owe to the government, with differences in predicted vs actual tax rates being due to random chance and not any bias in the system itself.

0.7 Question 11

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 feel that since the entire source code for the evaulation model is online and free for anyone to view, that this model should be considered "fair". However, even though everyone can see something doesn't mean that it is legible to some, if any few select people. In the documentation, this is clear as there are a lot of things which can be hard for people not interested in data science to understand. For example, trying to wrap your head around how this model was created is basically impossible for people that aren't formally trained in machine learning, or even adults with lots of technical knowledge about subjects like math or statistics. However, I believe that the documentation is still good for those who care, particularly the 2021 version of the model on GitLab.