0.1 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

$\geq$

As the number of features increases, meaning our model becomes more complex, our training error goes down.

0.2 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 [19]: # 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: 1.9339633173823714
 1: -0.030647249803554506
 2: 1.4170991378689641
```

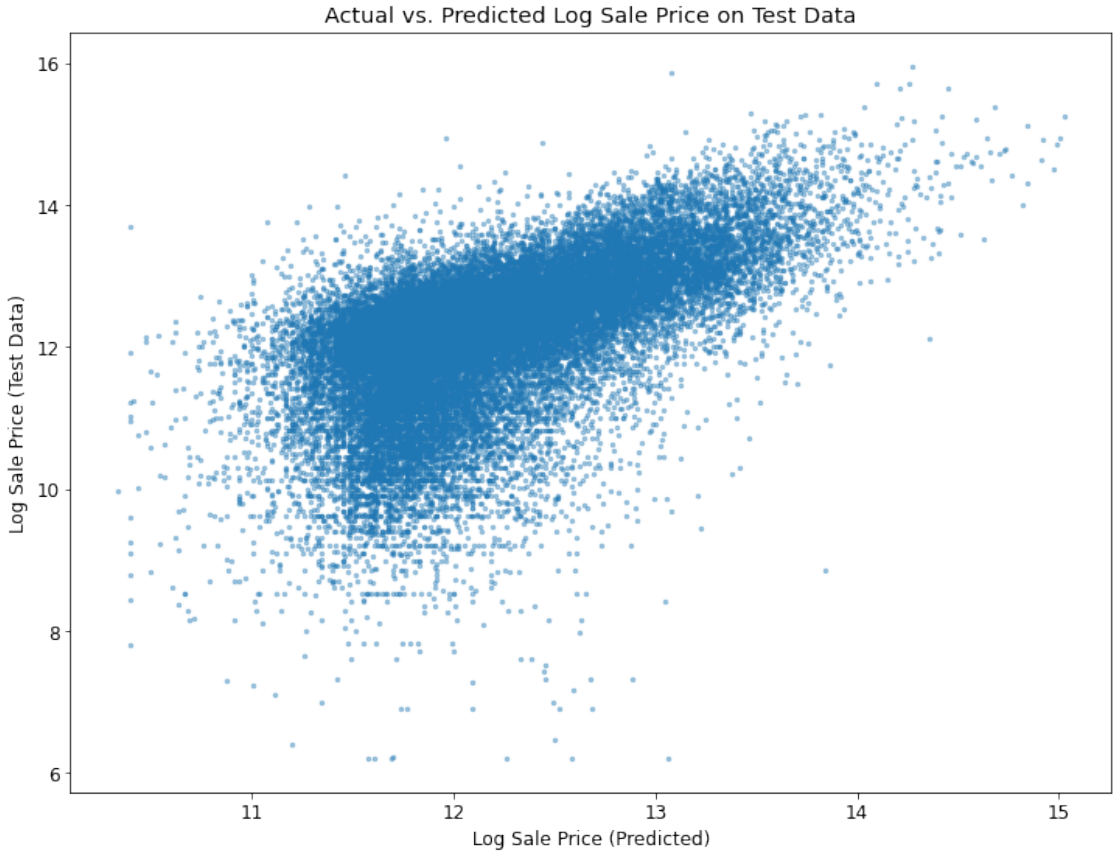
As we see before with our first model we only had 1 feature added, indicating that the relationship between bedrooms and log sales price is positively correlated. As we add an extra feature, it now considers the correlation or relationship between bedrooms in a house and square feet. When we fit our new model labeled m2, we have a problem now because there is some relationship between our two variables in which these two variables are highly correlated. This may lead to a sign change in our coefficients.

0.3 Question 7

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

In the cell below, use `plt.scatter` to plot the predicted `Log Sale Price` from **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 [20]: plt.scatter(y_predicted_m2,y_test_m2,s=6,alpha=0.4);
         plt.xlabel("Log Sale Price (Predicted)");
         plt.ylabel("Log Sale Price (Test Data)");
         plt.title("Actual vs. Predicted Log Sale Price on Test Data");
```



0.4 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?

Our root mean squared error means how far we were in estimating the true value of a house vs our prediction. For an individual homeowner, it means how the house was evaluated vs what the system for Cook County deems how properties are valued and how it fits into the actual value of one's house. It might affect the amount of property tax one receives, as the system depicted to evaluate one's house is not very accurate. We might have a property who's assessed value is lowered as well because of this or rated higher.

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.

0.5 Question 10

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

Fairness in property assessments and taxes are assessed by the value of one's property, contingent on on ones facots such as size,structure,age and location. Tax rates are leveraged based on the assessed value of your property and based on these characteristics alone.

0.6 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?

No, for an average documentation for the average reader it uses hyperparameters selection consisting of a tree structure. This data structure LightGBM, as said in the hyperparameter section is said to be useful, however it is difficult to understand what each parameters purpose is for the model. For example, we aren't able to know why min_data_in_leaf works, only that we know it prevents overfitting in our data. LightBGM has the potential to be accurate and useful, but the model and iteration is difficult to understand.