0.1 Question 1: Human Context and Ethics

In this part of the project, we will explore the human context of our housing dataset. **You should watch Lecture 15 before attempting this part.**

0.1.1 Question 1a

"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.**

Home owners: They would be interested to see how their purchase of a house has increased or decreased as this can be used to directly affect their future resell price or their mortgage.

Investors: Those that invest into houses across the country to make money would be interested in seeing the growth of their investments over time.

Government: Whether this is state or federal governements, some form of government would be interested in seeing the price of houses in certain areas in order to get more information demographics

0.1.2 Question 1b

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. Would you consider some of these scenarios more (or less) fair than others? Why?

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.

In my opinion, they are all unfair. A is unfair for the seller, B is unfair for the buyer, C is unfair for poor buyers and rich sellers, and D is unfair for rich buyers and poor sellers.

I would consider some more unfair than others because generally if you are rich you are less affected by price changes than if you are poor, therefore those that affect the wealthier clients, who would not mind/care as much, would most likely be seen as more 'fair'.

0.1.3 Question 1d

What were the central problems with the earlier property tax system in Cook County as reported by the Chicago Tribune? 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)

Systematic racism. If they were white owned they were misvalued in order to assist their finances and what communities they lived in while black owned property where misvalued so that they had to pay higher levels of taxes and were forced into worse communities. They would overvalue low-end homes and overvalue high-end.

0.1.4 Question 1e

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

In addition to the regressive taxation on property value, there was also the graph shown in lecture that visualized the over/under valuation of homes depending on white ownership and redlining that occurred which led to discrimination in certain areas.

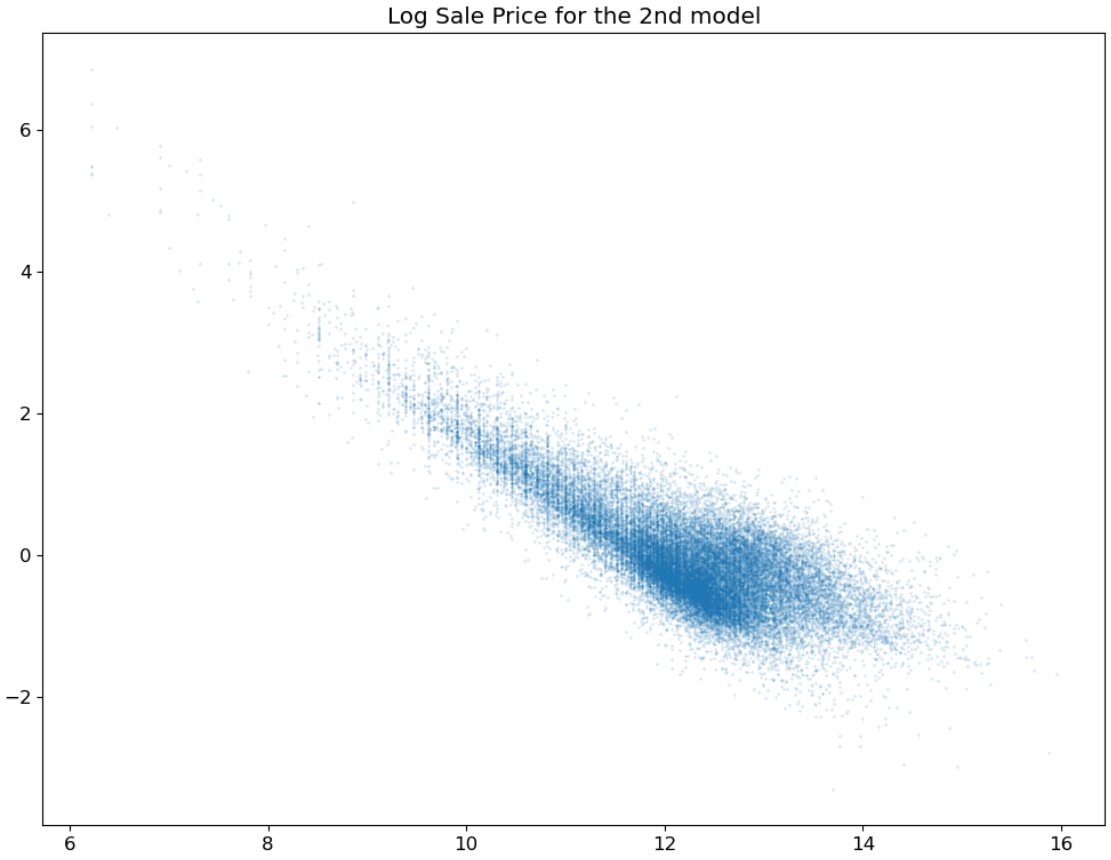
0.2 Question 4a

One way of understanding a model's performance (and appropriateness) 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 such a large dataset, it is difficult 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 [24]: plt.scatter(Y_valid_m2, (Y_predicted_m2-Y_valid_m2), s=0.05, alpha=0.6)
         plt.title('Log Sale Price for the 2nd model')
```

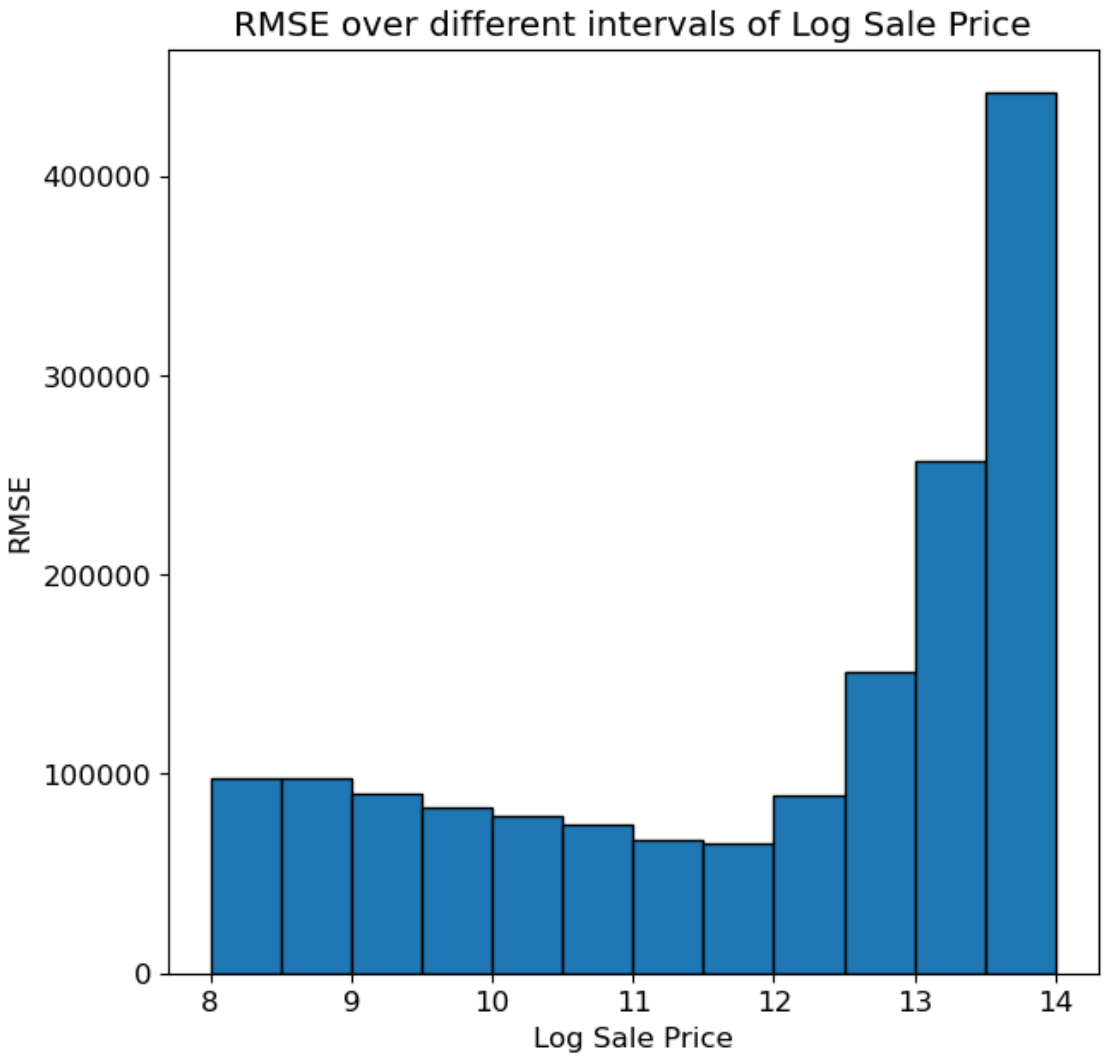
```
Out[24]: Text(0.5, 1.0, 'Log Sale Price for the 2nd model')
```



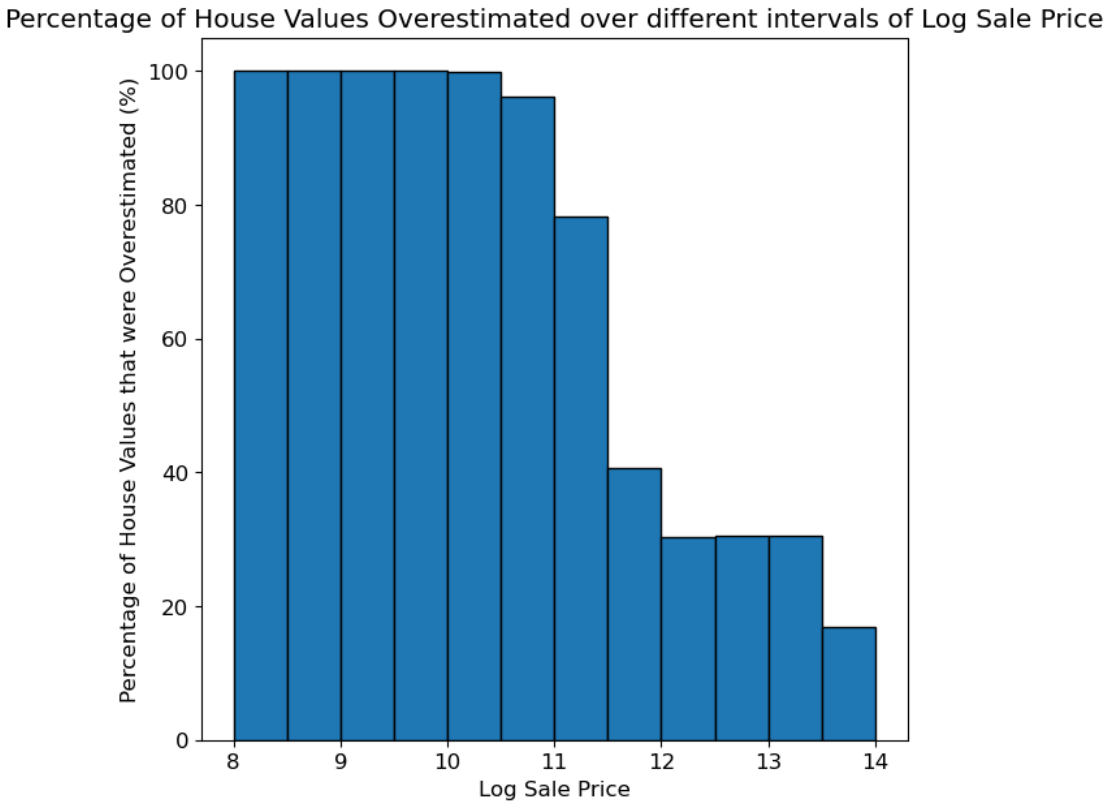
0.2.1 Question 6c

Now that you've defined these functions, let's put them to use and generate some interesting visualizations of how the RMSE and proportion of overestimated houses vary for different intervals.

In [52]:
```
# Run the cell below to generate the plot; no further action is needed
rmses = []
for i in np.arange(8, 14, 0.5):
    rmses.append(rmse_interval(X, Y, i, i + 0.5))
plt.figure(figsize = (7, 7))
plt.bar(x = np.arange(8.25, 14.25, 0.5), height = rmses, edgecolor = 'black', width = 0.5)
plt.title('RMSE over different intervals of Log Sale Price')
plt.xlabel('Log Sale Price')
plt.ylabel('RMSE');
```



In [53]:
```
# Run the cell below to generate the plot; no further action is needed
props = []
for i in np.arange(8, 14, 0.5):
    props.append(prop_overest_interval(X, Y, i, i + 0.5) * 100)
plt.figure(figsize = (7, 7))
plt.bar(x = np.arange(8.25, 14.25, 0.5), height = props, edgecolor = 'black', width = 0.5)
plt.title('Percentage of House Values Overestimated over different intervals of Log Sale Price
plt.xlabel('Log Sale Price')
plt.ylabel('Percentage of House Values that were Overestimated (%)');
```



Explicitly referencing **any ONE** of the plots above (using `props` and `rmses`), explain whether the assessments your model predicts more closely align with scenario C or scenario D that we discussed back in `1b`. Which of the two plots would be more useful in ascertaining whether the assessments tended to result in progressive or regressive taxation? Provide a brief explanation to support your choice of plot. For your reference, the scenarios are also shown below:

```
C. An assessment process that systematically overvalues inexpensive properties and undervalues expensive
D. An assessment process that systematically undervalues inexpensive properties and overvalues expensive
```

C. because if you look at hte percetnage of house values that were overestimated (%) you can see that the lower end of log sale price is at 100% or close to 100%, whereas the more expensive houses are only overvalued at 30%, no where near the cheaper houses.

0.3 Question 7: Evaluating the Model in Context

0.4 Question 7a

When evaluating your model, we used RMSE. In the context of estimating the value of houses, what does the 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.

A positive residual indicates to an individual home owner on average we will overevaluate the estimate of their house and cause higher property taxes. A negative residual would mean the inverse of this in which we would underevaluate their house and cause lower property taxes.

0.5 Question 7b

Reflecting back on your exploration in Questions 6 and 7a, in your own words, what makes a model's predictions of property values for tax assessment purposes "fair"?

This question is open-ended and part of your answer may depend upon your specific model; we are looking for thoughtfulness and engagement with the material, not correctness.

Hint: Some guiding questions to reflect on as you answer the question above: What is the relationship between RMSE, accuracy, and fairness as you have defined it? Is a model with a low RMSE necessarily accurate? Is a model with a low RMSE necessarily "fair"? Is there any difference between your answers to the previous two questions? And if so, why?

'Fair' is never an easy word to define. But in my opinion, if people have to be disrupted, then those that are negatively affected relatively less (in this case if wealthy people have to pay slightly more property taxes), then I would consider that more 'fair'. In the end, I want as few people disrupted as possible to the most minimal extent.