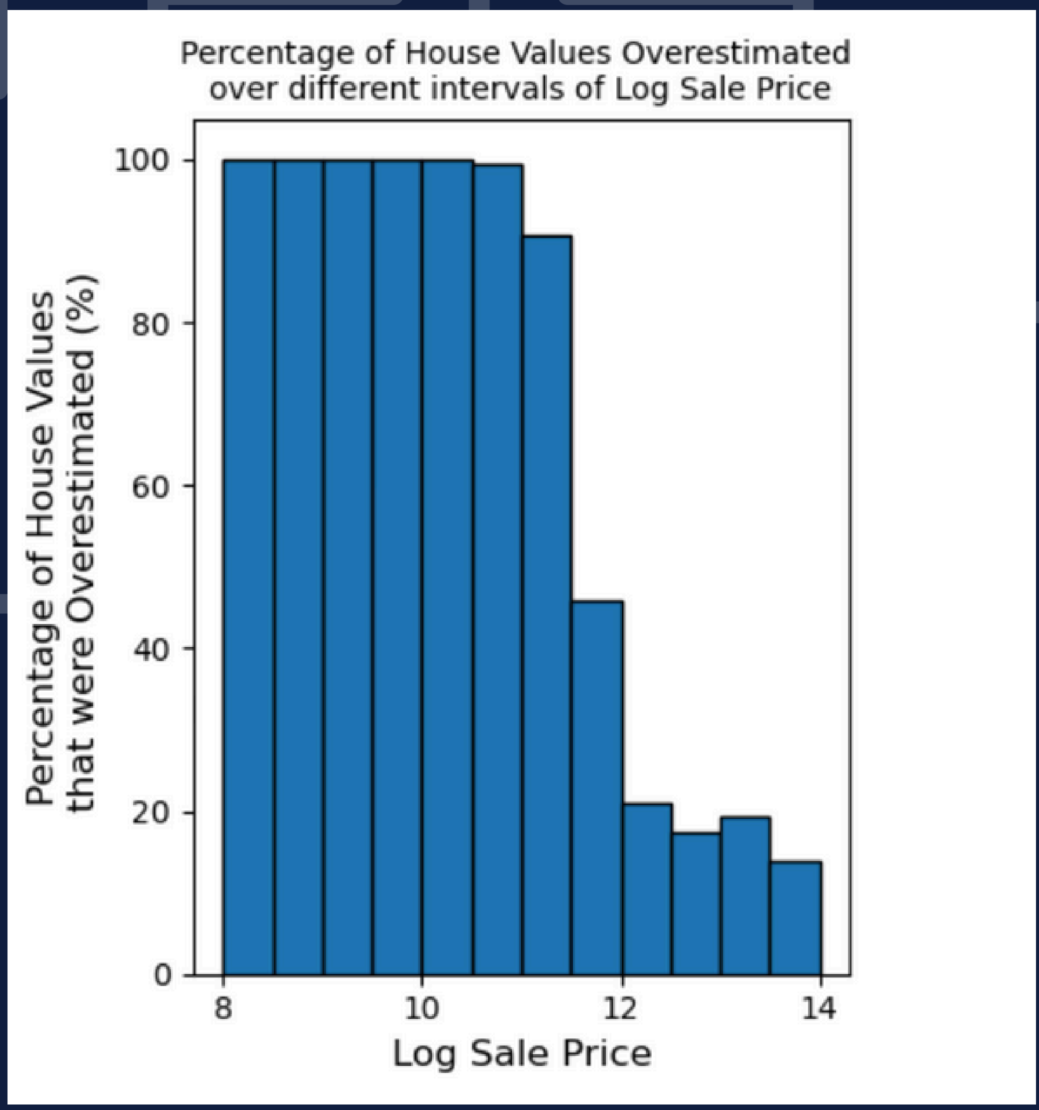

Predicting Housing Prices

Python

pandas, numpy, scikit-learn

Steps and Challenges:

The goal of this project was to predict housing prices in a region with known issues of redlining and overpricing in certain neighborhoods. Achieving this required a comprehensive exploratory data analysis (EDA) process, which involved filtering, cleaning, and separating useful data from less relevant information. This groundwork was essential to prepare the data for a linear model that could predict housing prices based on various factors.

To build this model, I used Python libraries like Pandas and Scikit-learn, which allowed me to fit the model to the county's official dataset and explore the relationships between different variables. One of the most challenging aspects of the project was identifying which variables had the most significant impact on housing prices. This required a deep dive into analytics, using techniques such as heatmaps and scatterplots to visualize correlations and trends. I also experimented with the training loss to fine-tune the model and determine which variables to include for the most accurate predictions.

The complexity of the data, combined with the need to account for historical biases in housing practices, made this project particularly demanding. It was a delicate balance between cleaning the data enough to make it usable and retaining enough information to capture the nuances of the market. In the end, this project provided valuable insights into the factors that influence housing prices and highlighted the importance of thorough data analysis in making accurate predictions.

Takeaways:

- <u>EDA</u>'s importance.
  - Working with so much information can make way for irrelevant, duplicate, or blank details to come in the way. From this project I really understood the potential consequences that come from using python packages without EDA, as the models may focus too much or too little on certain things that then decrease the efficacy of the predictor.
- <u>Multiple</u> visualizations are useful.
  - These datasets came with varying information, which required an intensive understanding of each variable and how it may impact my predictor's outcome. I initially limited the number of method's in which I was to find patterns in the data which ultimately led to a weak predictor. After taking my time to understand when and why to use visualization, my accuracy increased my magnitudes.
- <u>Testing techniques</u> matter.
  - When tuning my linear model, I recognized the importance of choosing the right testing methods. In this particular case with the quantitative values, I found use in and trusted the results after calculating the RMSE. Being able to identify when to use different techniques is an incredibily important takeaway from this project that I've implemented in all coding practices.

Given this is a class project, I cannot put my entire code online. However, I would be more than ready to discuss the code by request.