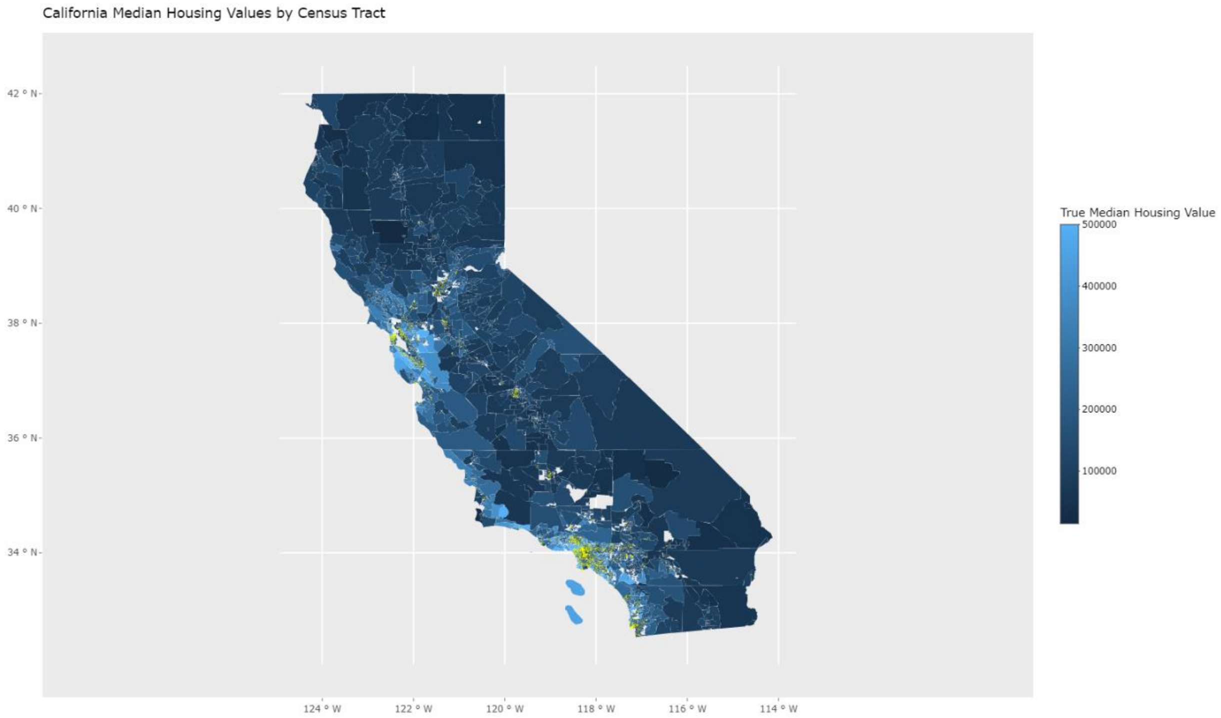
I was given data at the census tract level in California with metrics like latitude and longitude, median housing age, median income, and population, along with the true median housing value for the census tract. In order to predict median housing value given the data, I initialized a gradient boosting algorithm with a maximum number of 10,000 trees and an interaction depth of 6. I also performed two-fold cross validation on the model, which enabled me to find the optimal number of trees for my prediction based on out-of-sample performance. I found that the optimal model used 4,917 trees, with an out-of-sample RMSE of 46,856, which is a 23% percent average error when to the mean true median housing value of 206,868.

In order to map both my original data and the predictions, I matched the latitude/longitude pairs I was given with census-tract GEOIDs. In so doing, I noticed a problem with the original data where the coordinates were rounded to the hundredth decimal place. This was not much of a problem for the large census tracts, but many census tracts in California, especially in dense urban areas, are smaller in size than a hundredth of a coordinate. This lead to many rows having duplicated coordinates and being impossible to assign to a specific census tract, ultimately creating many holes in the graph especially in downtown Los Angeles and San Francisco. I fixed this problem by segregating out the non-specific rows and used a fuzzy joining algorithm to match them to census tracts within 0.5 miles. I then took the median of those matched rows to color the census tract on the map.

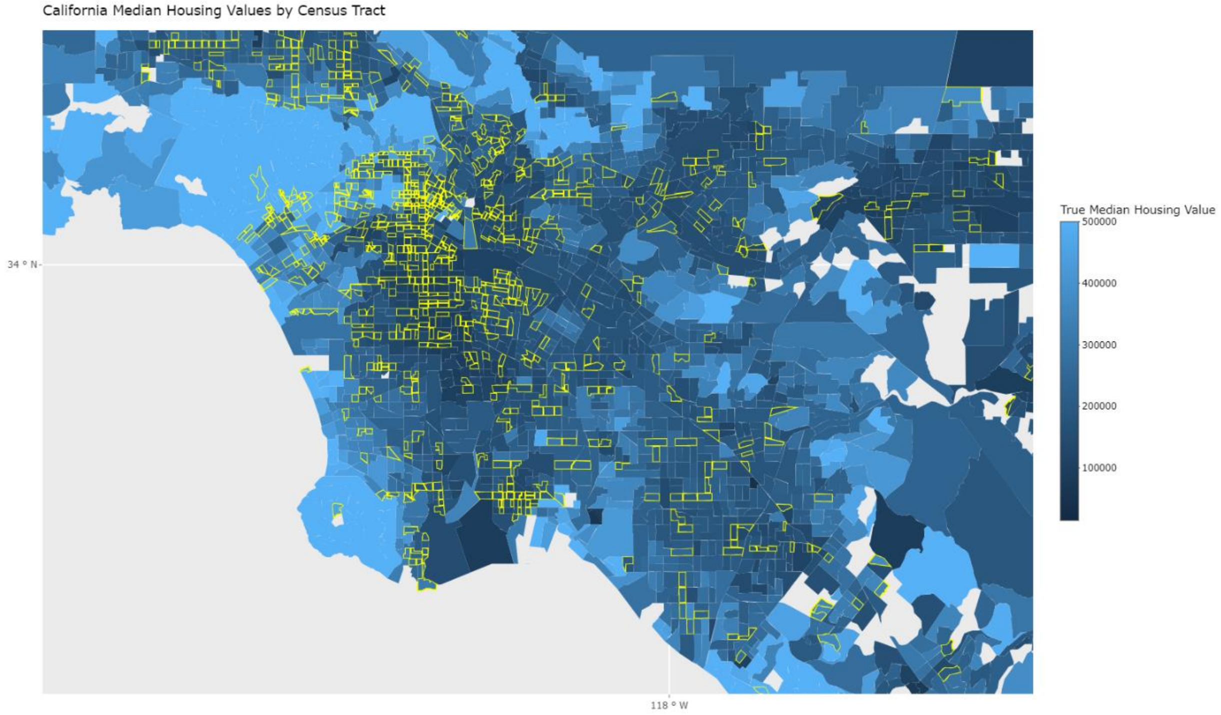
Below I have attached images from the various maps I created for this project, however, due to the scale of the data (many census tracts are very small and tightly packed, and therefore illegible when viewing the entirety of California) I believe that the best way to view this data is through an interactive map that can be zoomed and panned. The code for these maps is provided along with my other work, however creating them is quite resource intensive, so I have also provided HTML files which will allow you to load the already created maps in a web browser. I highly encourage opening those files and panning around the maps yourself rather than relying solely on the images below.

Original Data:

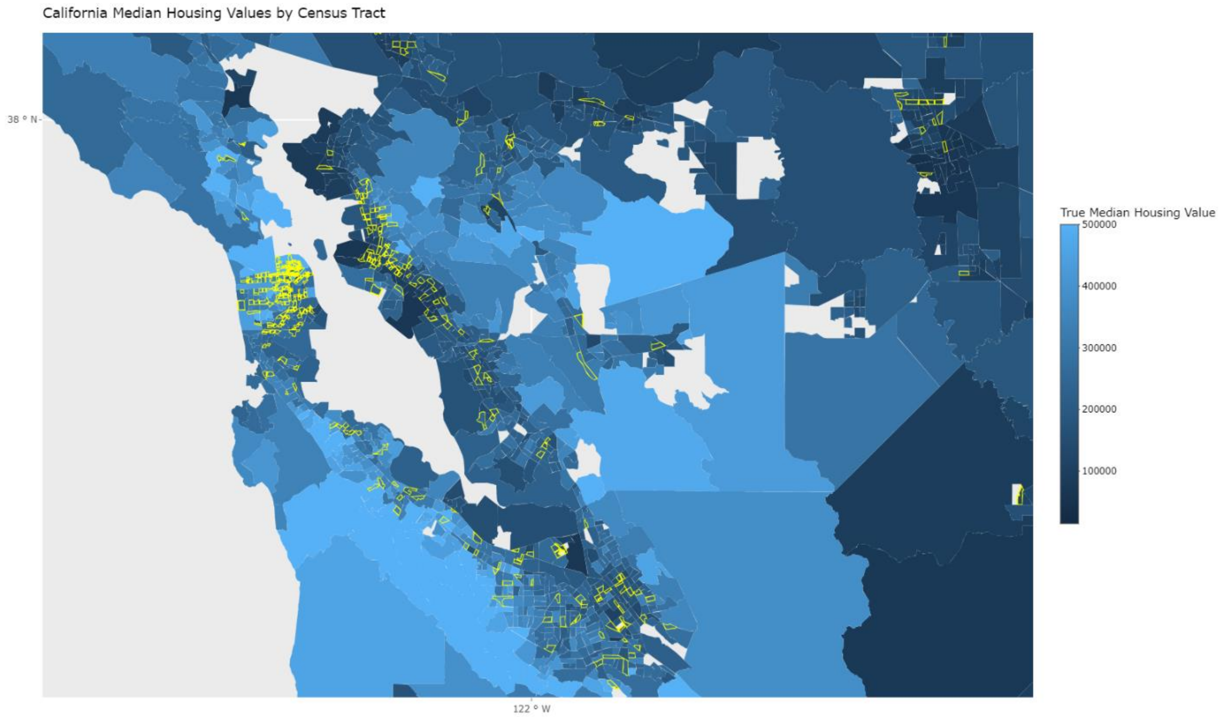
(Relevant file: “True California Median Housing Values by Census Tract.html”)



Los Angeles:


California Median Housing Values by Census Tract
34 ° N
118 ° W
True Median Housing Value
500000
400000
300000
200000
100000

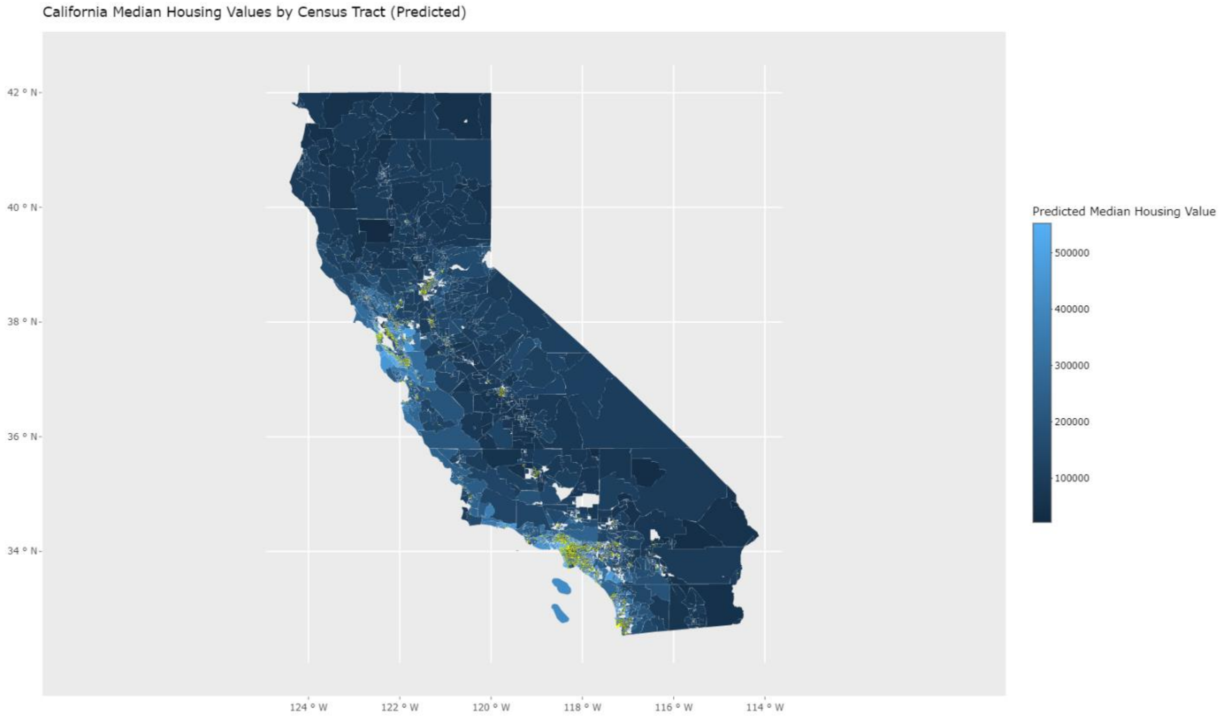
San Francisco:


California Median Housing Values by Census Tract
38 ° N
122 ° W
True Median Housing Value
500000
400000
300000
200000
100000

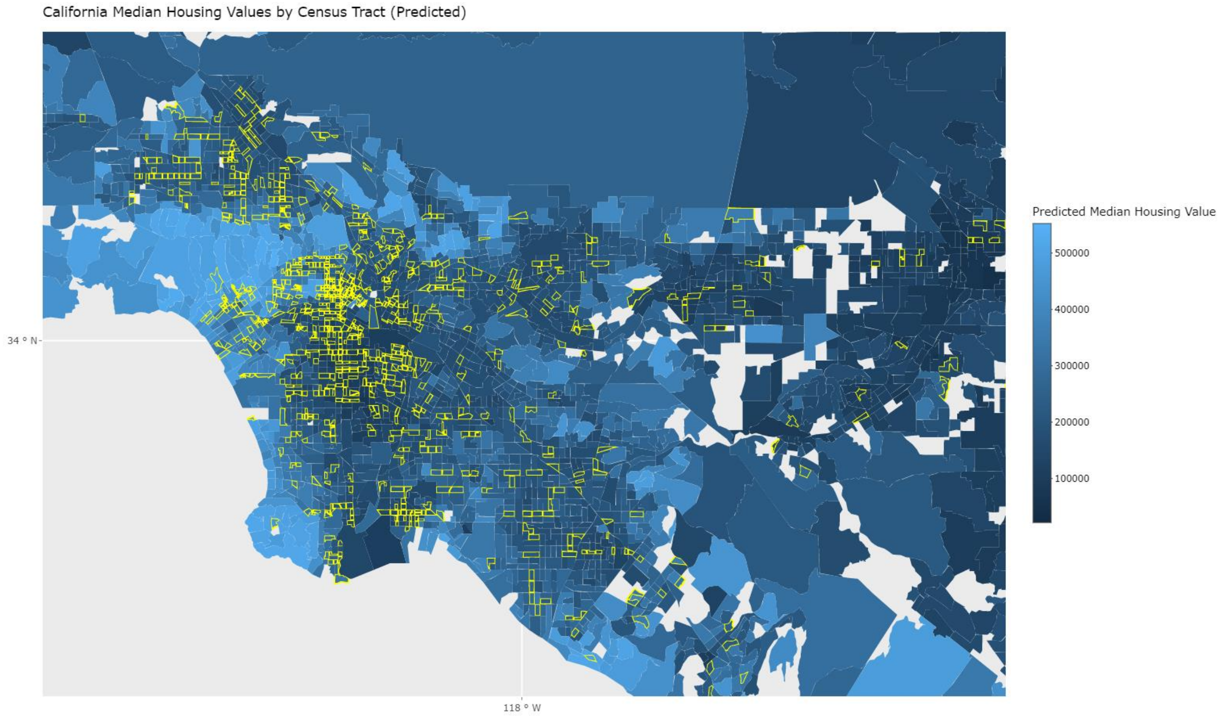
The yellow borders on these maps and also on the maps of my predictions below denote census tracts which were colored based on the fuzzy matching algorithm I described in the introduction.

Predicted Values:

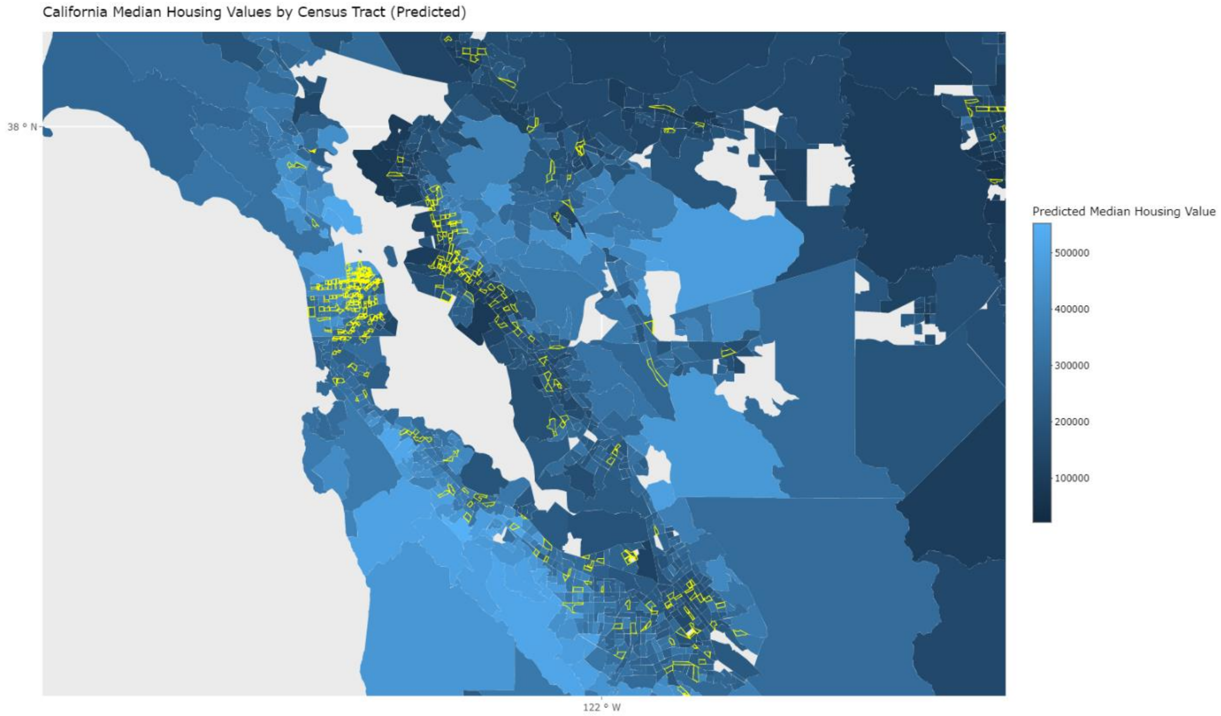
(Relevant file: “California Median Housing Values by Census Tract (Predicted).html)



Los Angeles:

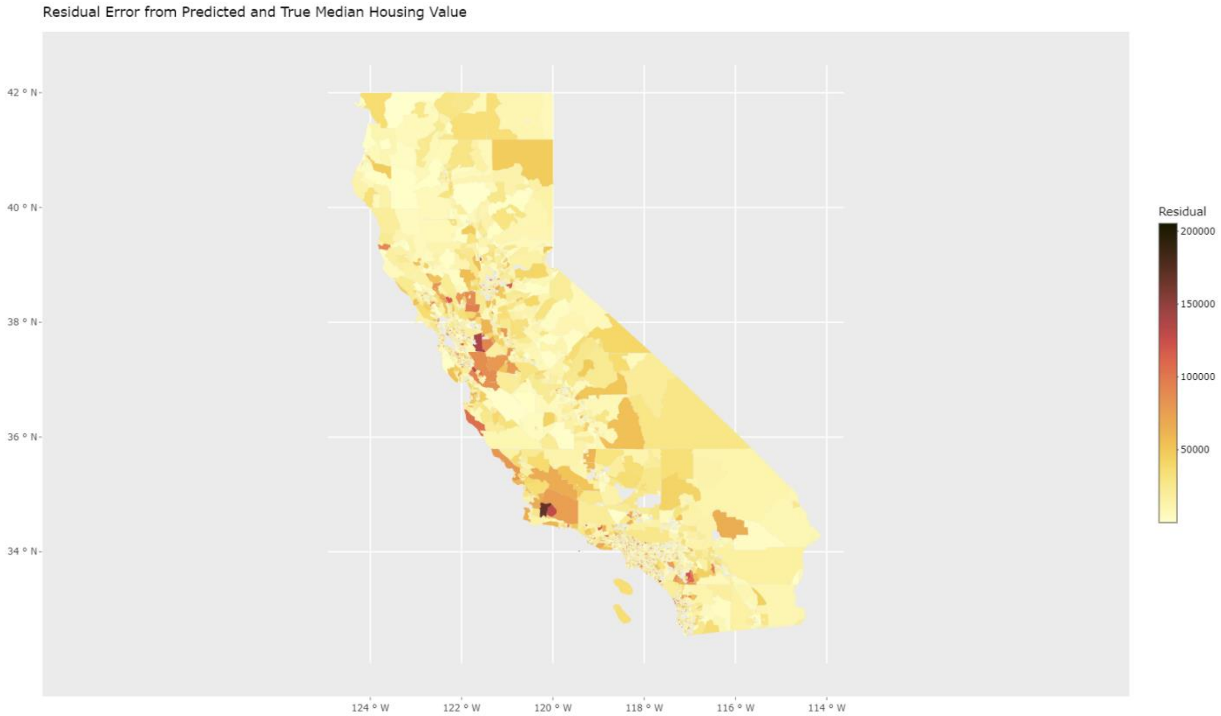

California Median Housing Values by Census Tract (Predicted)
34 ° N
118 ° W
Predicted Median Housing Value
500000
400000
300000
200000
100000

San Francisco:

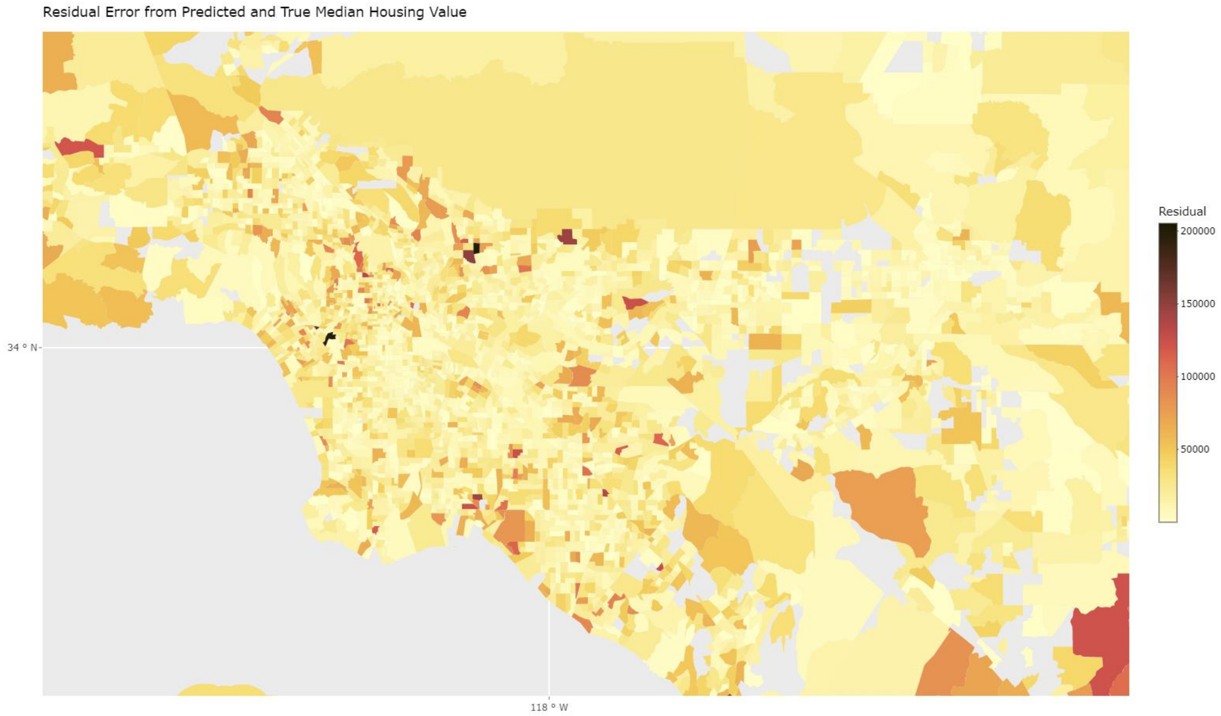

California Median Housing Values by Census Tract (Predicted)
38 ° N
122 ° W
Predicted Median Housing Value
500000
400000
300000
200000
100000

Residual Errors:

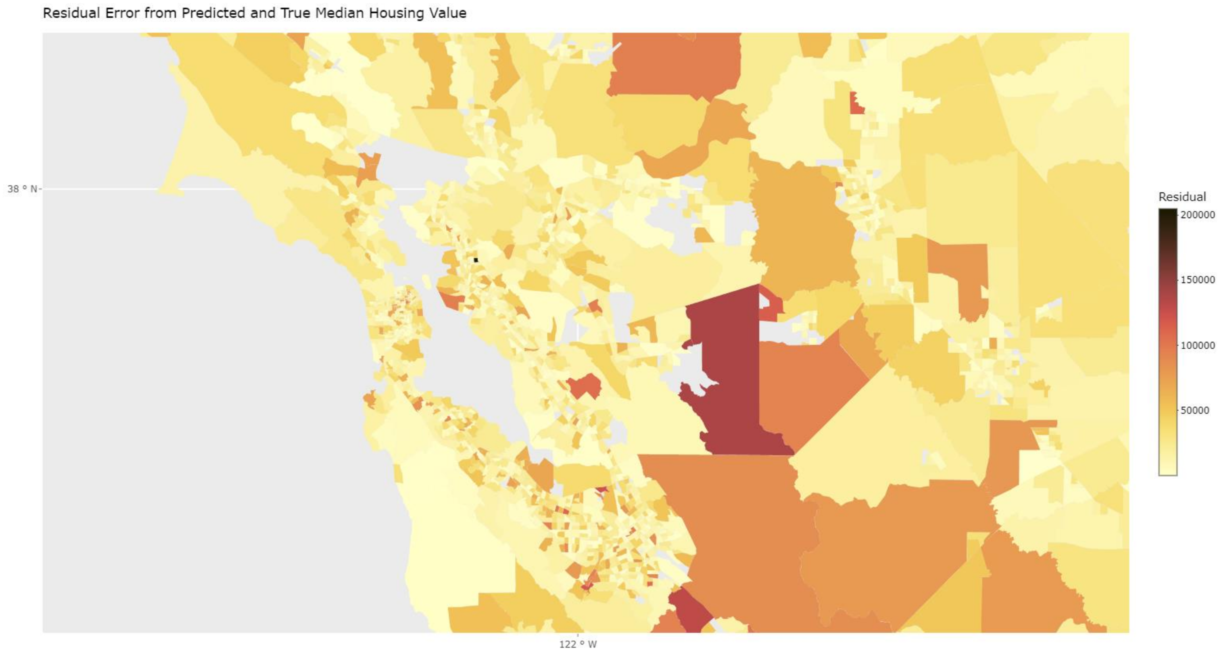
(Relevant file: Residual Error.html)



Los Angeles:



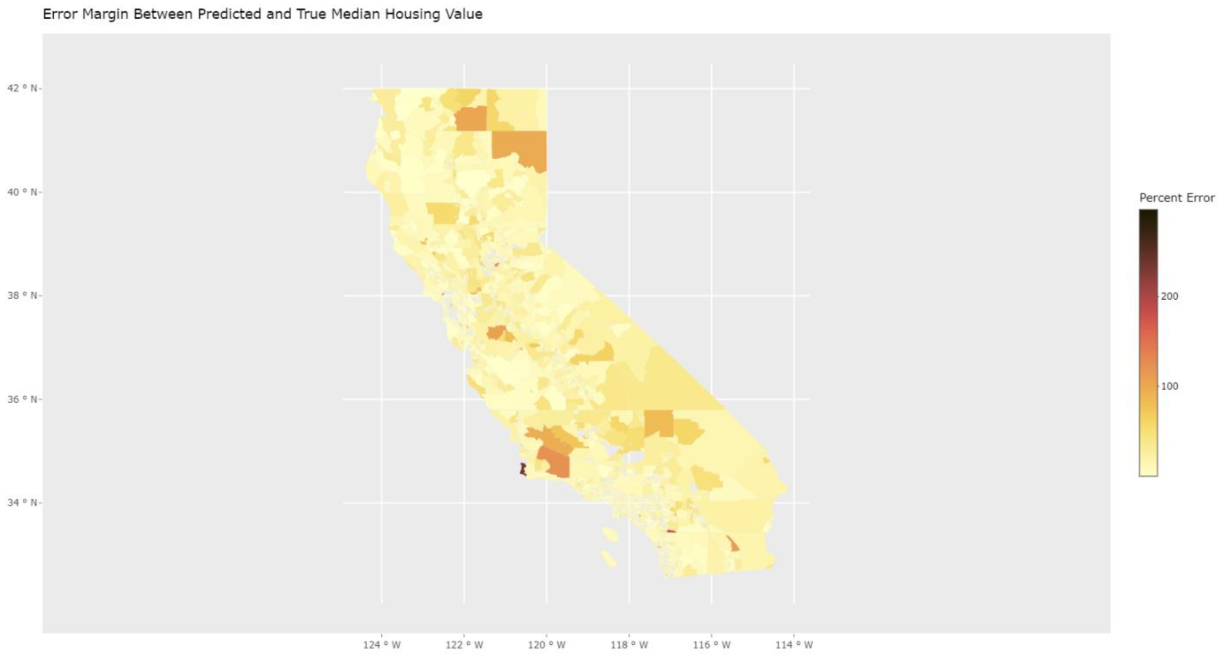
San Francisco:



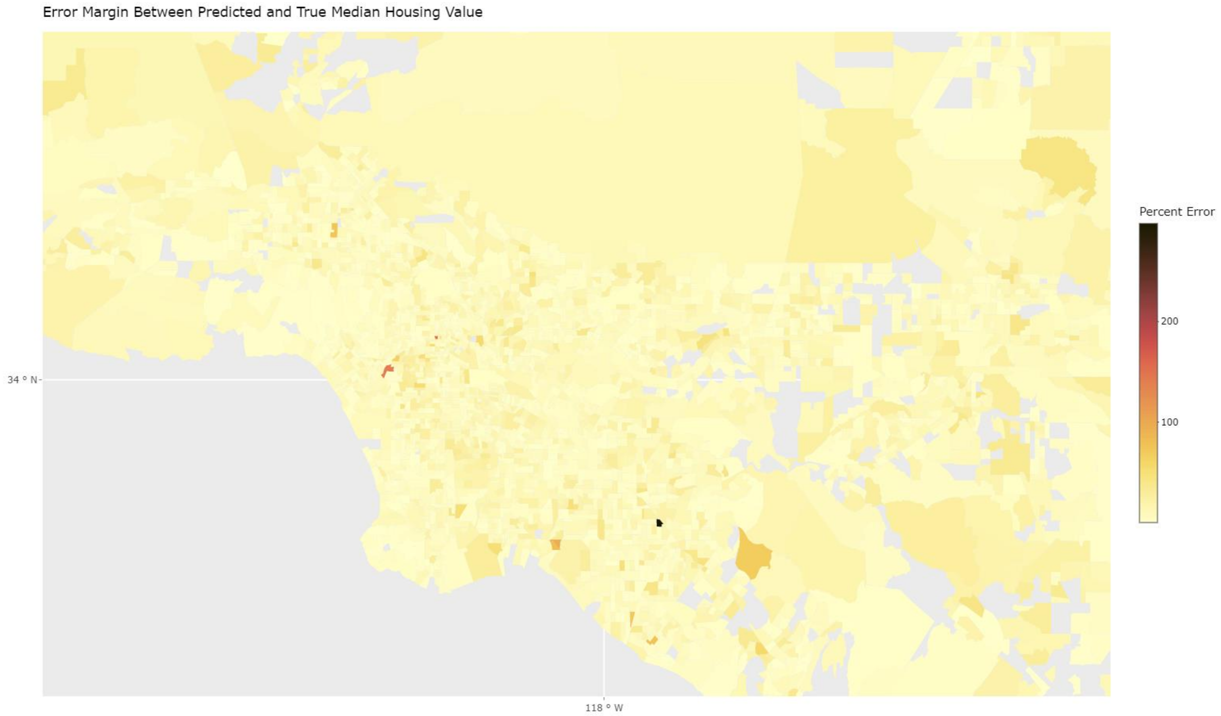
I have also provided maps of the percentage error (residual divided by true median housing value), which I believe is a more useful assessment of accuracy—overvaluing a home by $20,000 in a market where the average home sells for $2,000,000 is much less severe than overvaluing by the same amount a home in a market where the average home value is $200,000, but the above graphs would color both districts the same.

Percent Error:

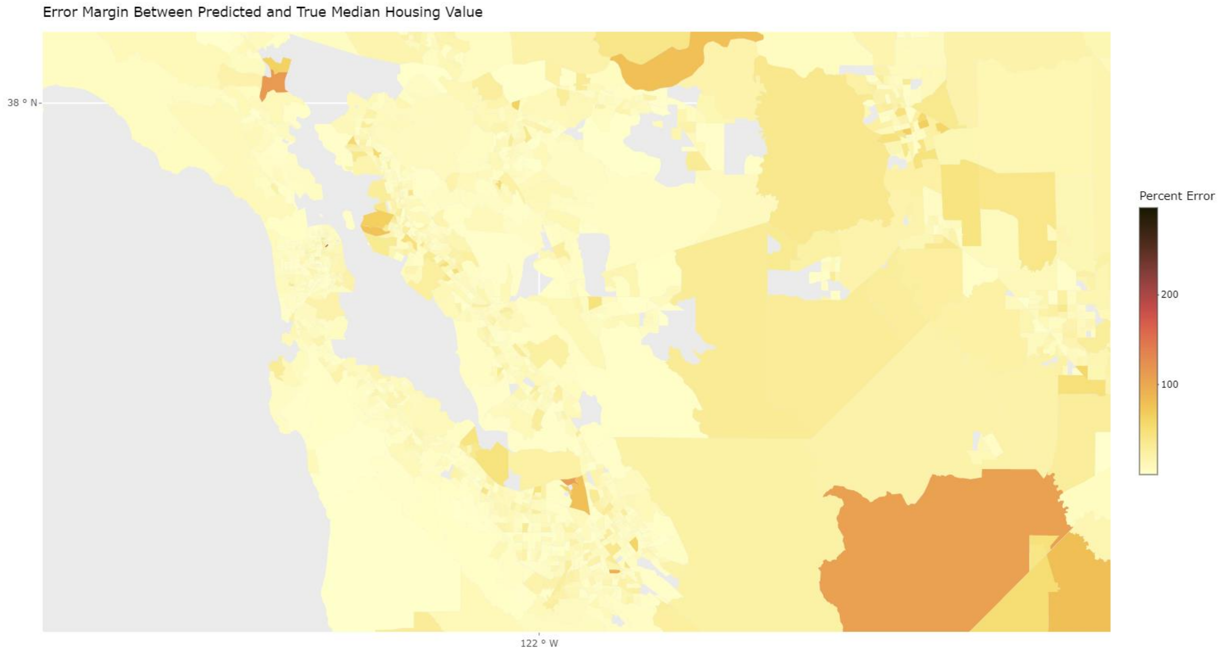
(Relevant file: “Percent Error.html”)



Los Angeles:


Error Margin Between Predicted and True Median Housing Value
Percent Error
200
100
34 ° N
118 ° W

San Francisco:


Error Margin Between Predicted and True Median Housing Value
Percent Error
200
100
38 ° N
122 ° W