Data_Mining_HW3_P1

Joey Herrera

3/27/2021

Data Mining Assignment 3

Question 1: What causes what?

1.1 Why can't I just get data from a few different cities and run the regression of "Crime" on "Police" to understand how more cops in the streets affect crime? ("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)

Answer: You can not just obtain data from a few different cities and run a regression of crime on police to understand how more cops in the streets affect crime. Various other factors might have a causal effect on the crime and police variables. These unobserved variables could cause the change in crime and police. Therefore, a regression of crime on police would be based on the correlation of the two variables. For instance, if only one city consistently has massive protests, there will be an increase in police, but there might be fewer citizens on the streets because they are scared of the mob. In this case, there would be a smaller number of victims on the street for criminals to target. As a result, the protest causes both an increase in the number of police and a decrease in the crime rate simultaneously. The other cities in the data most likely have other significant circumstances where the number of police in each city varies from day to day. Thus, comparing cities to one another without controlling for these unobserved factors will invalidate the final results.

1.2 How were the researchers from UPenn able to isolate this effect? Briefly describe their approach and discuss their result in the "Table 2", from the researchers' paper.

Answer: Researches from the University of Pennsylvania were able to isolate the effect of the number of police in a city on the crime rate in Washington D.C. by focusing on days where the fear of terrorism was significant. The researches focused on days labeled orange or above, which translates to days where a terrorism event is more likely to occur. As a result, more police are dispatched through the capital. Controlling for the level of terror on a given day allows the researchers to ensure that there would be an increased number of police in the city throughout that particular set of days. In Table 2, the first regression shows that the increase of police in the city on high alert days corresponds to approximately a 7.3% decrease in the crime rate with a 5% significance level. The second regression in Table 2 controls for midday ridership, which causes a less significant reduction in crime on high alert days. In the second regression, the high alert variable decreases the crime rate by approximately 6% at a 5% significance rate. Thus, midday ridership has a significant effect on the crime rate. Higher levels of midday ridership equate to an increased number of potential victims on the street.

1.3 Why did they have to control for Metro ridership? What was that trying to capture?

Answer: The researchers at UPenn controlled for Metro ridership to account for the possible lower number of people out in the city during high threat days. Controlling for Metro ridership captures the number of potential victims in the city during the day, which provides value because that number can be compared to Metro ridership on any other given day where there is no high alert. Days where Washington D.C. is not on high alert serve as a baseline for the regular number of police in the city and the typical amount of civilians riding the Metro midday.

1.4 Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the

table. Can you describe the model being estimated here? What is the conclusion?

Answer: The first regression in Table 4 is estimating the effect District 1 and Other Districts have on the crime rate while holding the high alert variable constant and controlling for midday Metro ridership compared to low alert days. The first column in Table 4 concludes that the increase in police caused by a high alert day in Washington D.C. in District 1 decreases the crime rate by approximately 2.6%. This result is significant at the 1% level. Thus, the increase of police on a high alert day in District 1 significantly decreases the crime rate.

Question 2: Predictive model building - green certification

Your goal is to build the best predictive model possible for revenue per square foot per calendar year, and to use this model to quantify the average change in rental income per square foot (whether in absolute or percentage terms) associated with green certification, holding other features of the building constant. Note that revenue per square foot per year is the product of two terms: rent and leasing_rate! This reflects the fact that, for example, high-rent buildings with low occupancy may not actually bring in as much revenue as lower-rent buildings with higher occupancy.

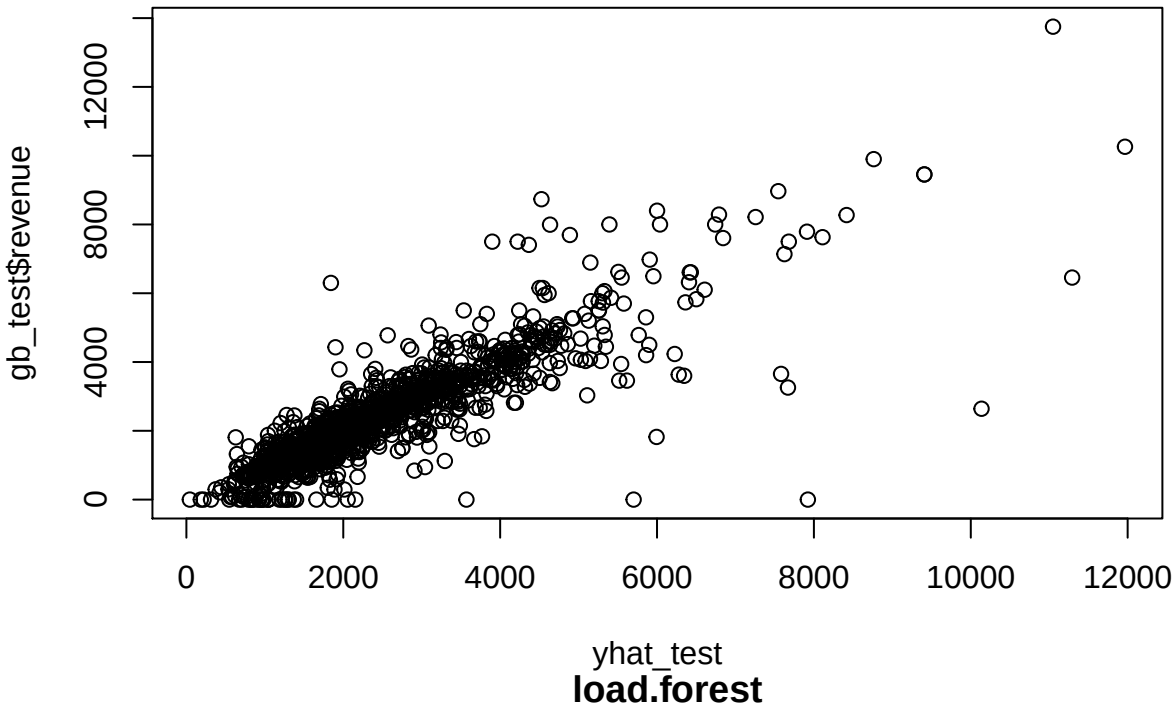
The purpose of this writeup is to identify the difference in revenue per square foot per year for buildings that have a green certification versus buildings that do not. Predicting the difference in revenue per square foot per year allows for the comparison of money made for buildings with relevant leasing levels depending on whether they have green certification.

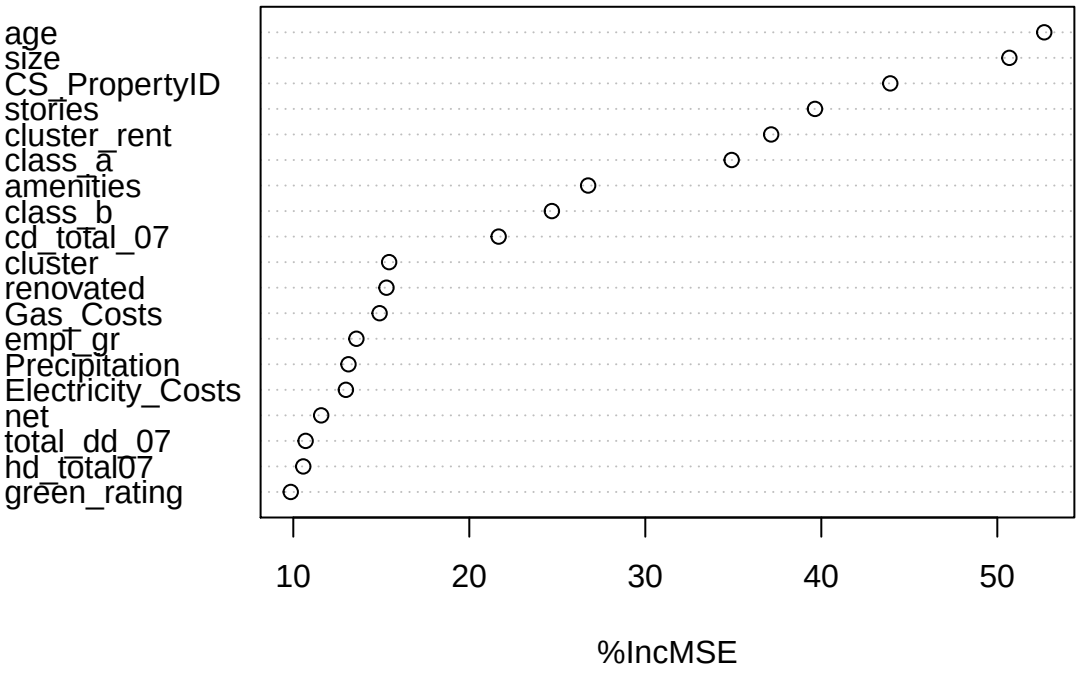
The data used to evaluate this problem is a set of 7,820 buildings split into clusters with at least one green-certified building and one non-certified building. In this dataset, other relevant indicators of revenue include the number of stories, age of the building, and if it has been recently renovated. Before creating a predictive model for revenue per square foot per year, I had to create the variable (revenue) using the product of the leasing rate and rent price for each building. Next, I decided to fit a single tree model for revenue with all of the indicators included as a base model. Creating this tree model gives me a baseline to compare the accuracy of future models with. Next, I fit a random forest model to the training data, consisting of 80% of the total observations. Finally, I predict the revenue per square foot per year for the remaining 20% of the data and compare the random forest model's out-of-sample accuracy to the single tree model via comparing RMSEs.

```
## [1] 885.9207
```

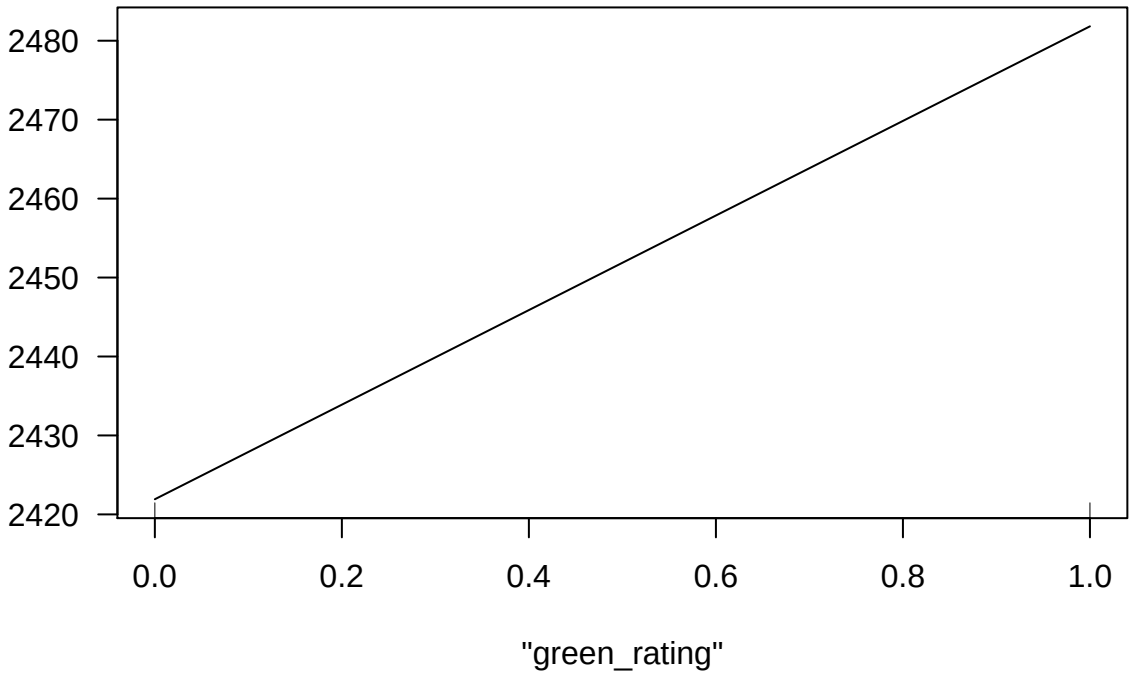
```
## [1] 702.4912
```

The root mean squared error from the random forest model is consistently approximately 200 dollars of revenue per square foot per year more accurate than the single tree model.

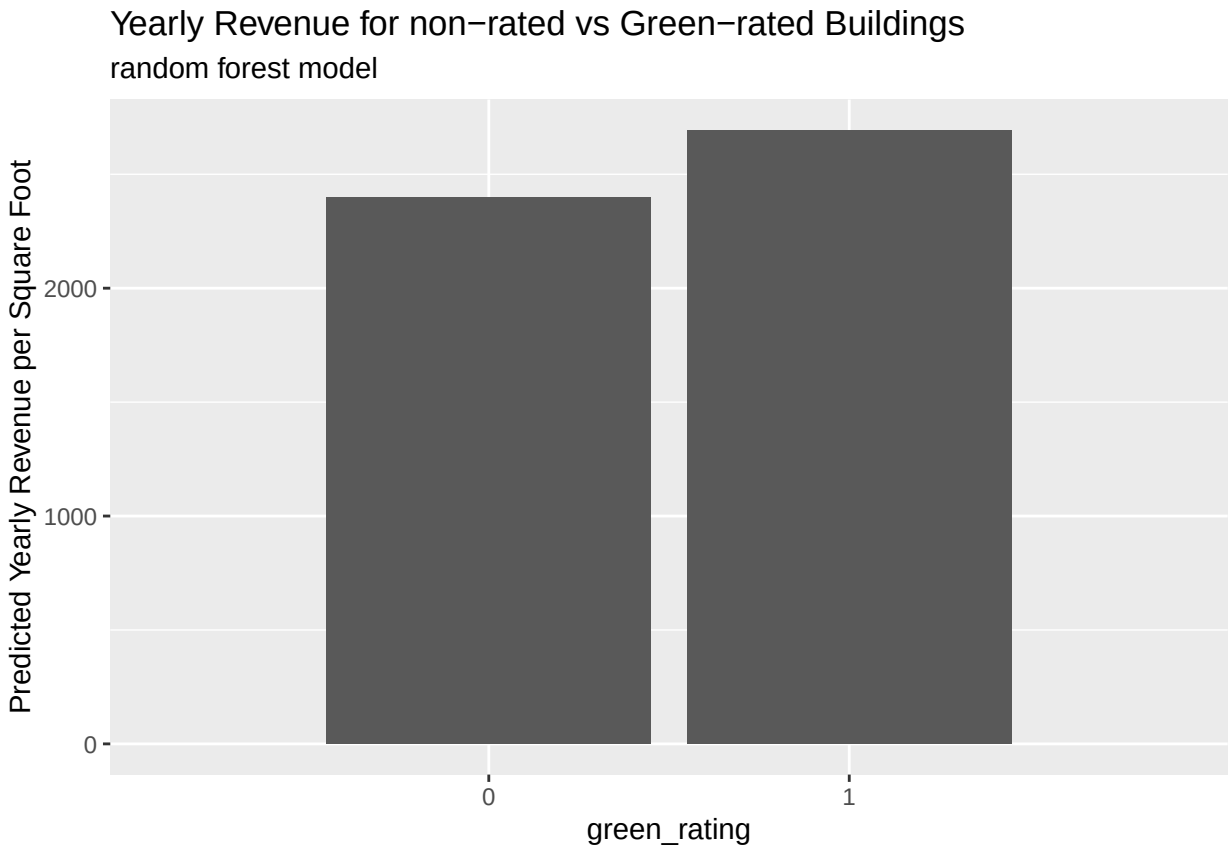




The variable importance plot above illustrates that a building's green rating does not have as significant an impact on the RMSE as most other variables in the random forest model. As a result, green rating is not a good predictor of revenue per square foot per year.



The partial dependence plot above shows the return for revenue per square foot per year for any green rating, holding all other variables constant. The plot depicts that green rating only accounts for about 60 dollars of revenue per square foot per year.




```
## # A tibble: 2 x 2
##   green_rating yhat_mean
##          <dbl>     <dbl>
## 1            0     2399.
## 2            1     2693.
```

The bar chart above displays a 200 dollar per square foot per year difference between buildings with a green rating and non-rated buildings. This difference is caused by the green rating variable and a combination of other factors correlated with the green rating variable.

Revenue per square foot per year for buildings with a green certification consistently is higher than buildings without a green certification. In conclusion, I recommend that the building owner weigh the cost of renovating to obtain a green certification versus the predicted yearly revenue per square foot and make their decision accordingly.

Question 3: Predictive Model Building - California Housing

Your task is to build the best predictive model you can for medianHouseValue, using the other available features. Write a short report detailing your methods. Make sure your report includes an estimate for the overall out-of-sample accuracy of your proposed model.

The goal of this writeup is to accuractely predict the median housing value for homes in California by census tract. Accurately predicting median housing values allows for predicting future housing values.

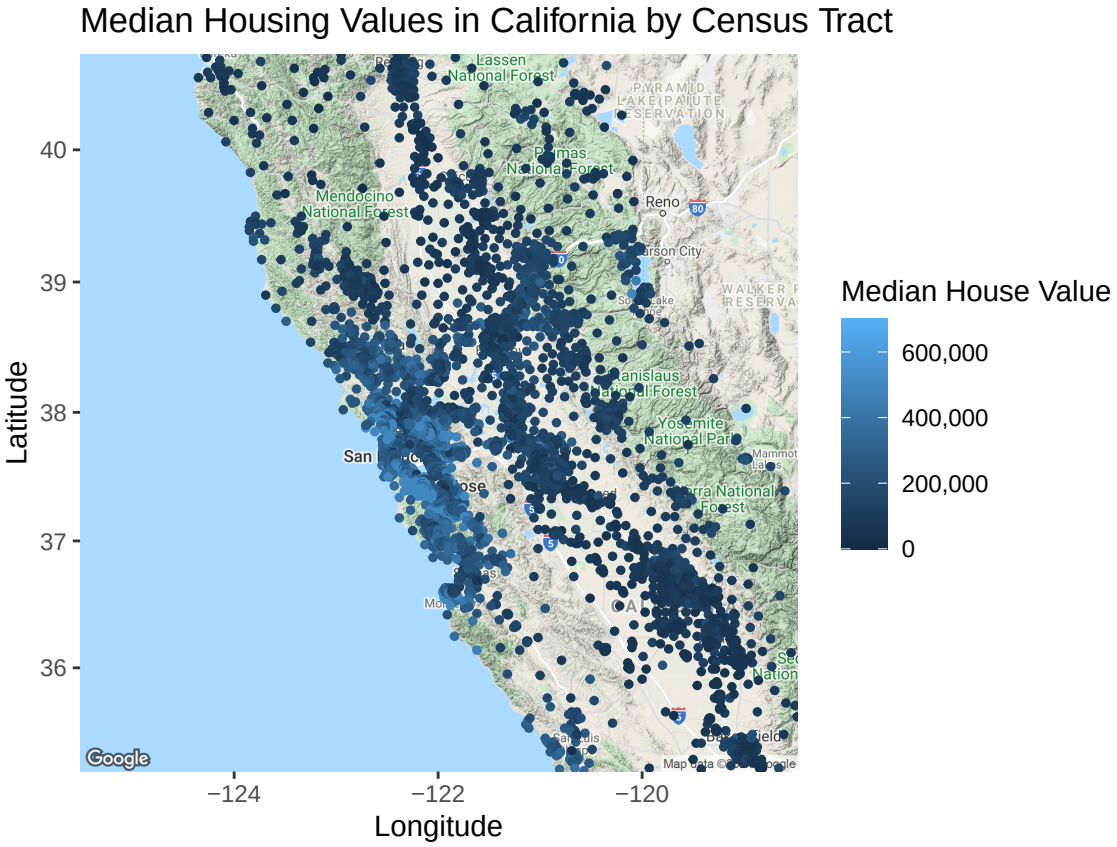
The available dataset includes observations of California's median housing values based on different census tracts, which are determined through each tract's unique latitude and longitude coordinates. This dataset includes other relevant housing value indicators such as the total number of bedrooms and population in each census tract. Before creating predictive models, I standardized the total number of rooms by dividing by the number of households in each census tract. The new units for the totalRooms variable are rooms

per household in a given tract. To begin creating the best predictive model, I fit a baseline linear model for median housing value using only the main factors. Next, I used a step function to determine a more optimal combination of main factors, pairwise interactions, and polynomials. Finally, I compared the baseline model's RMSE to the stepwise function's RMSE.

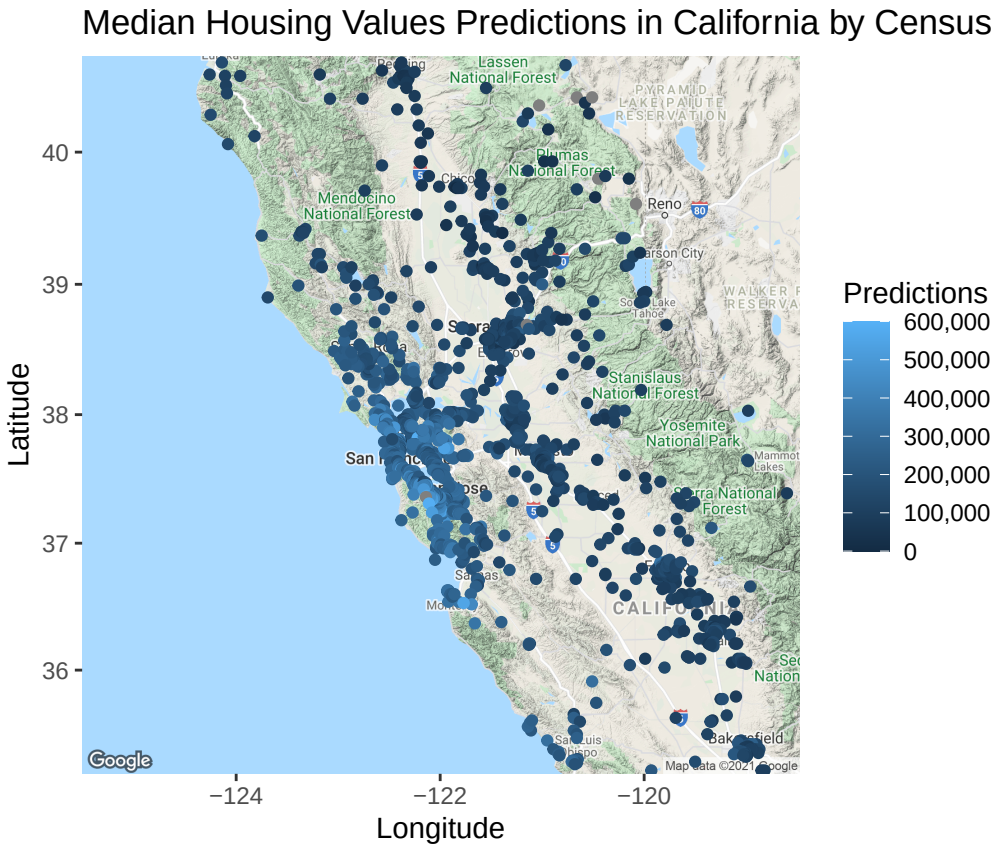
```
## [1] 65919.23
```

```
## [1] 69330.93
```

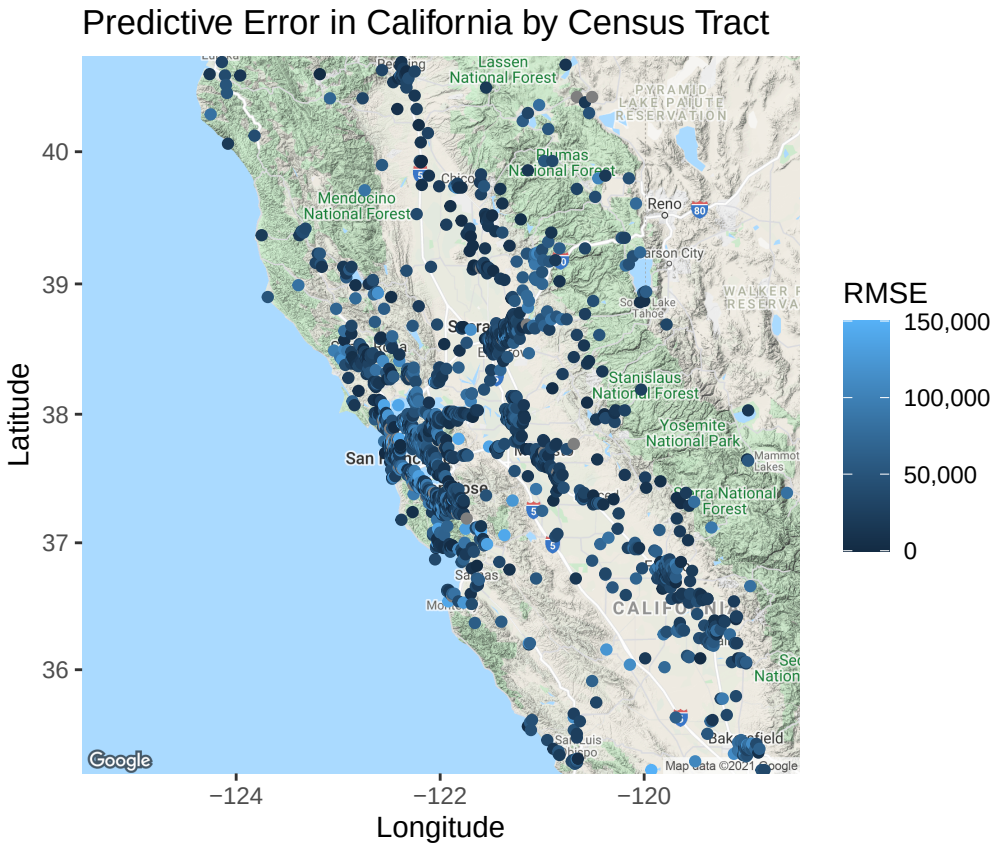
The RMSE for the stepwise function consistently outperformed the baseline model by approximately 6000 dollars closer to the actual housing price. To further visualize the map below's data is by plotting the median house value for each census tract in California. The map below illustrates the rise in median house value the closer the house is to the ocean.



The next plot shows the stepwise function's predictions on the 20% of the test set data. The plot below depicts the same trend as the plot above. The median housing value rises the closer a home is to the ocean.



The third and final plot depicts the RMSE for each observation in the test set. The RMSE varies regardless of location and includes grey points, which have an RMSE so large it is an outlier. The error seems to be randomly distributed among census tracts. Thus, the error does not impact the trend that median house values in California rise the closer the census tract is to the ocean.



In conclusion, in California, the closer the census tract is to the ocean, the higher the median home value. For potential home owners looking to buy a house in California, I recommend, taking distance from the ocean in to consideration when thinking about housing prices.