Exercises 3

Jordan Despain and Jack Freeman

03/22/2023

What Causes What?

1

You can't just run the regression of *Crime* on *Police* because there are a ton of other things that affect both the crime rate and the number of cops in a city that need to be taken into consideration, like poverty rates or how populated an area is. The results from this regression would be biased. At best, you might be able to get some sort of understanding of the correlation between these two variables.

2

The researchers were able to isolate the effect of *Police* on *Crime* by taking advantage of the terrorism alert system. When the threat level is orange, police presence is increased due to the threat of terrorism. This situation is unrelated to amount of street crime in the area. So, they can look at the effect these extra police on orange alert days has on street crime. From Table 2, we see the increased presence did indeed have a negative effect on total daily crime. Crime was lower on high alert days and their finding is statistically significant at a 5% level.

3

The researchers had to control for Metro ridership because it could be that tourists and residents were just more afraid to go out on high alert days and there was fewer potential victims and less activity on the streets. Controlling for Metro ridership helped determine if this was in fact what was happening. They find there was no diminishing effect on activity level in the city on these high alert days. So, controlling for Metro ridership helps better capture the effect of police presence on crime levels, and also helps better their argument for causality.

4

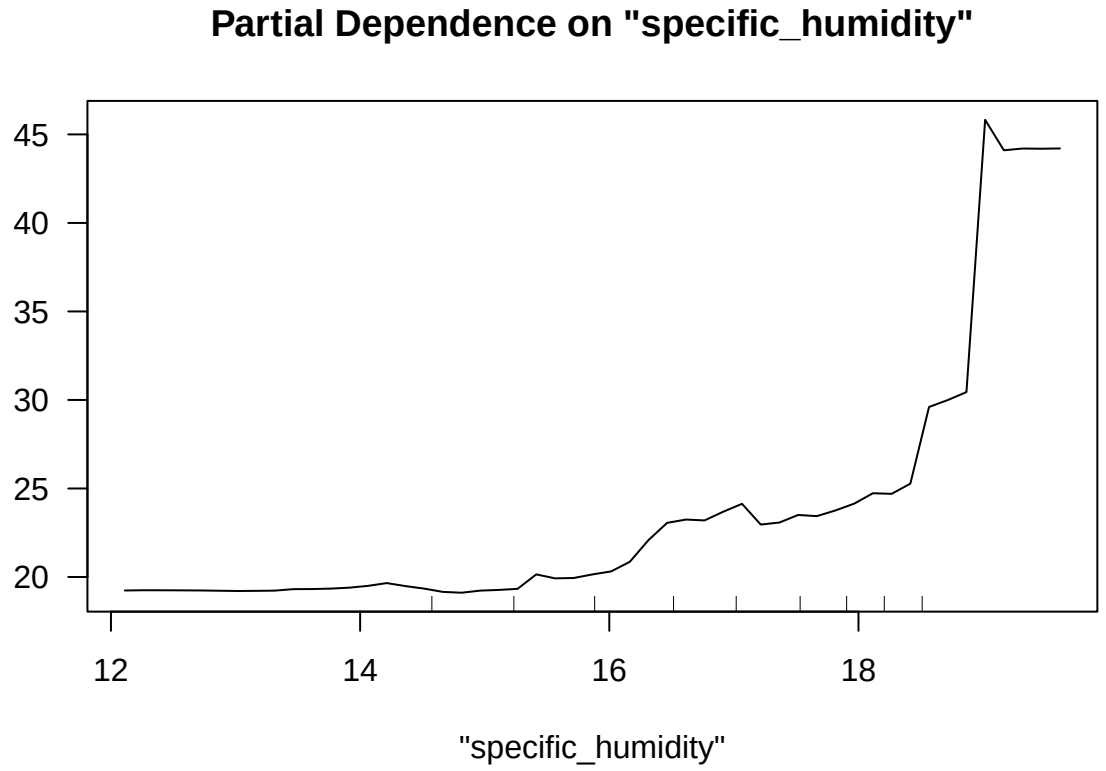
The results are estimating the effect of *Police* on *Crime* using high alert days and separating District 1 and the rest of the districts. District 1 is where the National Mall is located, which is in the same area as a lot of important buildings and monuments. So, on high alert days you would expect increased focus from police in this area as opposed to other districts. These results show there is a much larger statistically significant reduction effect on crime in District 1 compared to the rest of the districts. The effect on all the other districts is not statistically significant and cannot really tell us much about the effect in those areas. The researchers can use these results to provide more support to the argument that more police presence decreases crime levels.

Tree Modeling: Dengue Cases

We decided to predict dengue cases instead of log dengue cases because we are aware of the size of the population of these two areas, and we thought the actual number of cases would give us a better idea of the severity of the disease in these areas. Basically, we just found it more interesting because it gave us a better idea of the entire situation. In our models we decided to include the variables, *precipitation_amt*, *avg_temp_k*, *max_air_temp_k*, *specific_humidity*, *tdtr_k*, *city*, *season*, *precip_amt_kg_per_m2*. We thought each of these were important based off the description given by the students who compiled it. We also played around with the different variables and found these gave us the best results. We created a loop to give us 10 sets of results for the CART, random forest, and boosted tree models. We included the train-test split in the loop to give us new sets each time. We then took the mean RMSE for each model to give us a better idea of which one actually performs better more consistently. Here are the results,

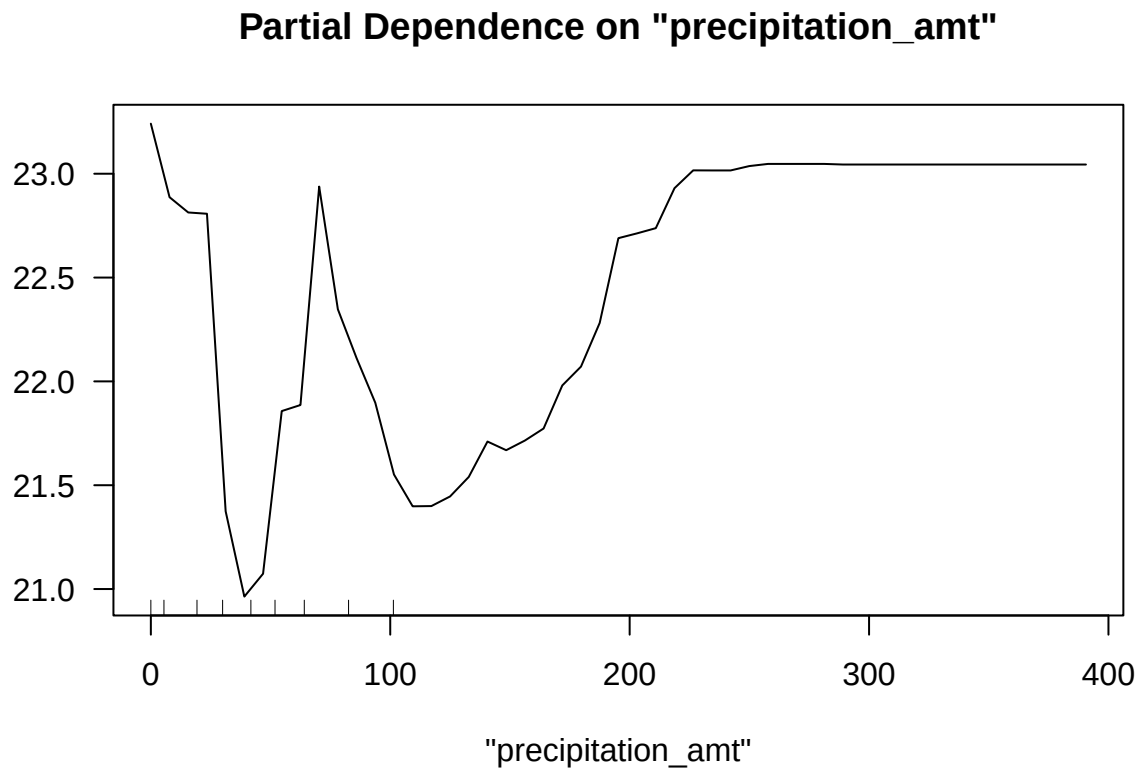
Model	***RMSE***
CART	27.51
Random Forest	26.33
Gradient-Boosted Tree	27.93

We found that the random tree model was the most accurate on the testing data. So, we used this model to make our three partial dependence plots. Here is the first for *specific_humidity*,



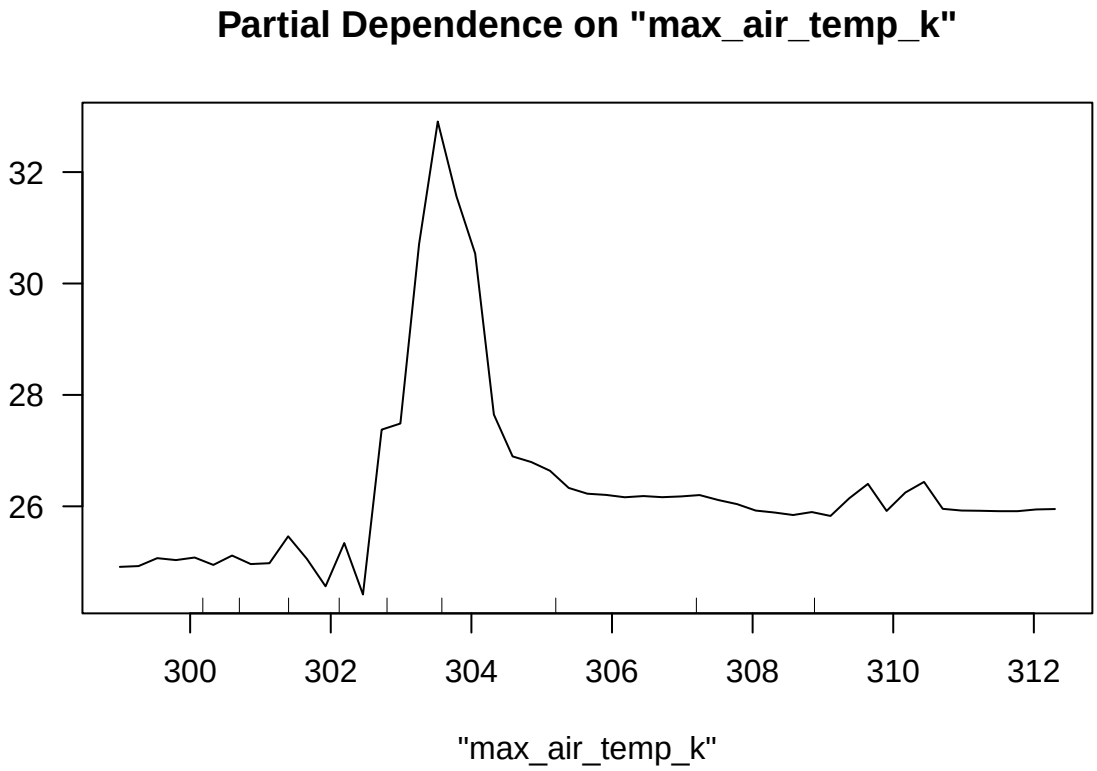
It is interesting over 18 you see a huge spike. It is as if this is the magic number for mosquitoes. Anything over 18 grams of water per kilogram of air is a good sign to really make sure your skin is covered and protected.

Here is the second for *precipitation_amt*,



We find the large dropoff in effect at low precipitation amounts really interesting. The rise at higher precipitation amounts is easier to understand, but it is very interesting to think about why we see a larger effect at the lower amounts as well. We know, in hot Texas days at least, when it just sprinkles and the sun comes out immediately after, you really feel the nasty humidity. So, maybe when it only rains a small amount, it is not enough to decrease the water vapor in the air. So, you still have high levels of humidity. These are just our thoughts however because we are definitely not meteorologists!

For the third, we decided to use *max_air_temp_k* because we found the huge spike from around 302-305 very interesting. Here is this plot,



We wonder if this is an ideal temperature for humidity levels or if this temperature is preferred by mosquitoes. It definitely had us very curious, so that is why we decided to included this variable as our wild card.

Predictive Model Building: Green Certification

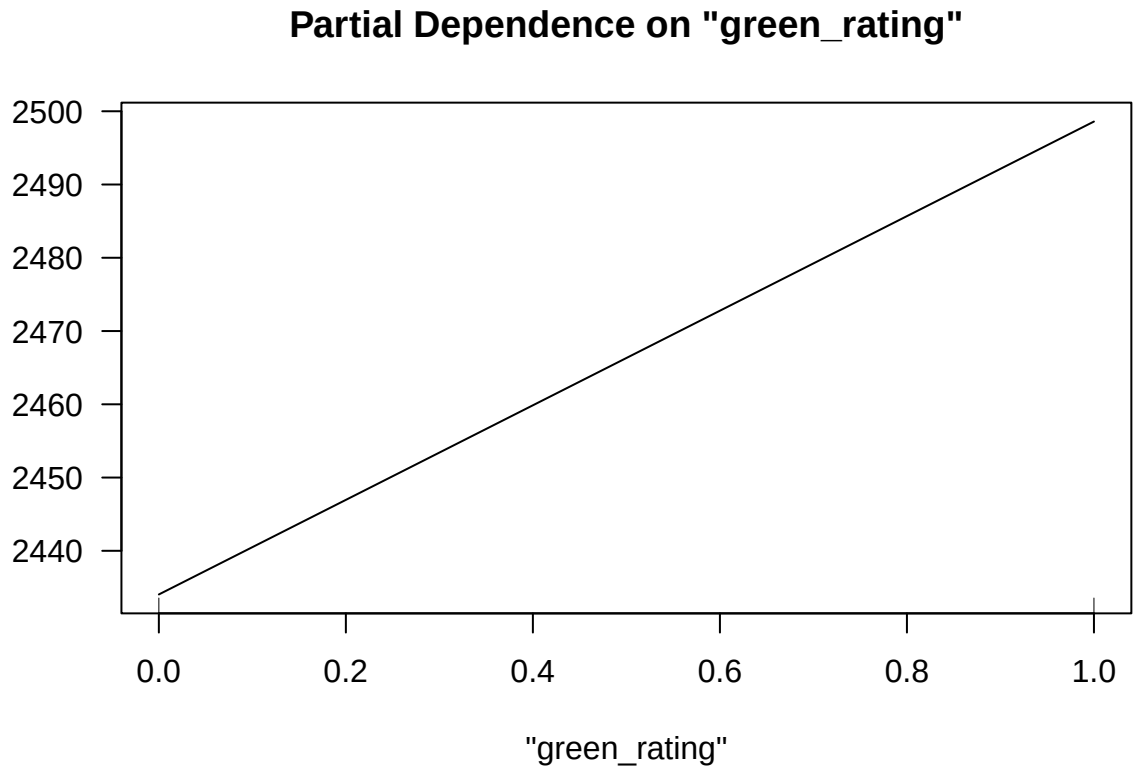
We want to build a model to try best predict revenue per square foot per calendar year, and then use this to try and get an idea of the effect a green certification has on rental income per square foot. To get started, we first prepare the data by creating the variable *revenue*, which is the product of *Rent* and *leasing_rate*. This variable represents the revenue per square foot per calendar year, so this is what we are trying to predict. We want to try multiple models to compare the results of each, so we choose a simple linear regression model, a random forest model, and a gradient-boosted tree model. Now, we start looking at the variables to try and decide what to include. We first remove *Rent* and *leasing_rate* because of their correlation with the outcome variable *revenue*. We also removed *LEED* and *Energystar* because we decided we were more interested in using the collapsed *green_rating* category variable. We also chose to remove *Electricity_Costs* and *Gas_Costs* as well, but then added them both in an interaction variable with *net* because we thought it would make the effect more interesting. Lastly, we removed the unique identifier variable *CS_Property_ID*.

Now that we have the variable we want to include in our models, we need to run the models and test their accuracy. We run each type of model 10 times in a loop, then take the average RMSE to try and get a better idea of their actual performance. Here are the results,

Model	***RMSE***
Linear Regression	1030.71
Random Forest	788.52
Gradient-Boosted Tree	956.23

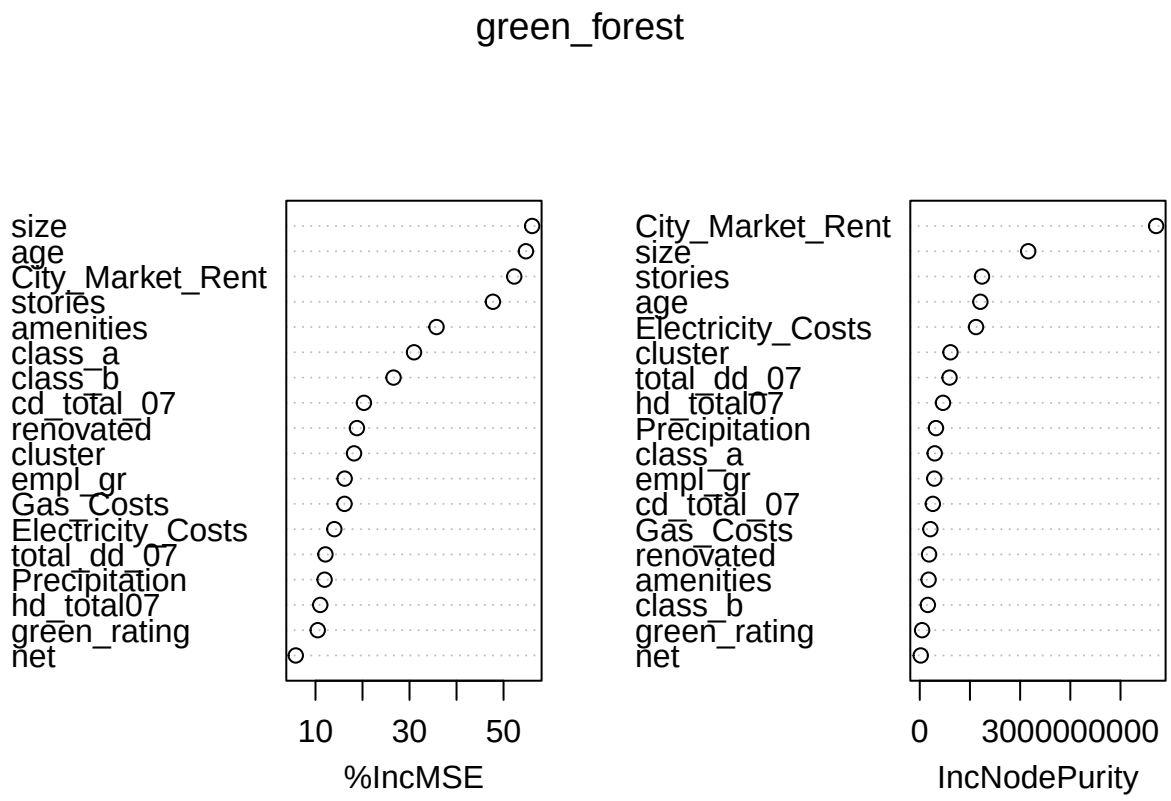
We see that our random forest model seems to give us more accurate predictions. So, we will use our random forest model to now try and see what the effect of *green_rating* is.

We can get a rough idea of the average change in rental income per square foot associated with green certification by looking at a partial dependence plot of *green_rating*. Here is what this looks like,



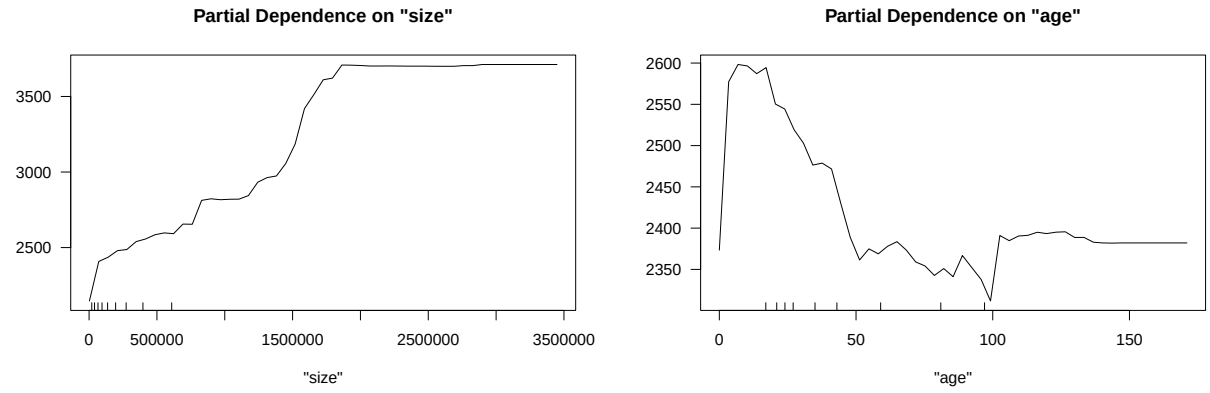
The variable *green_rating* is a binary variable, so we are interested in the difference in the values at 0 and 1. We see there is only about a difference of 60. This isn't much of an effect at all. It seems as those having a green certification may not be very significant on the amount of revenue received. We want to get a better idea of this variables importance by looking at the variable importance plot from our random forest model on the next page.

Here is the variable importance plot,



We see *green_rating* is near the bottom which means it's not very important and doesn't have much of an effect on our model's accuracy whether it is included or not. We notice some of the more important variables like *size* and *age* and we are curious what the effects of these variables look like. So, we want to check out the partial dependence plots for these and see if there is any interesting effects. We look at the plots on the next page.

Here are the partial dependence plots for *size* and *age*,



The effects are more along the lines of what you would expect to see. As the buildings increase in size, you would expect more rental income. Also, you would expect higher rental income from newer buildings and this decrease as the buildings get older. This is what we see in these plots. However, it is interesting in the plot for *age* how after 100 years you see a jump in revenue. We wonder why this is and if it has to do with some sort of historical effect or maybe the buildings have been renovated. We liked seeing this effect because it made us very curious about the cause of it.

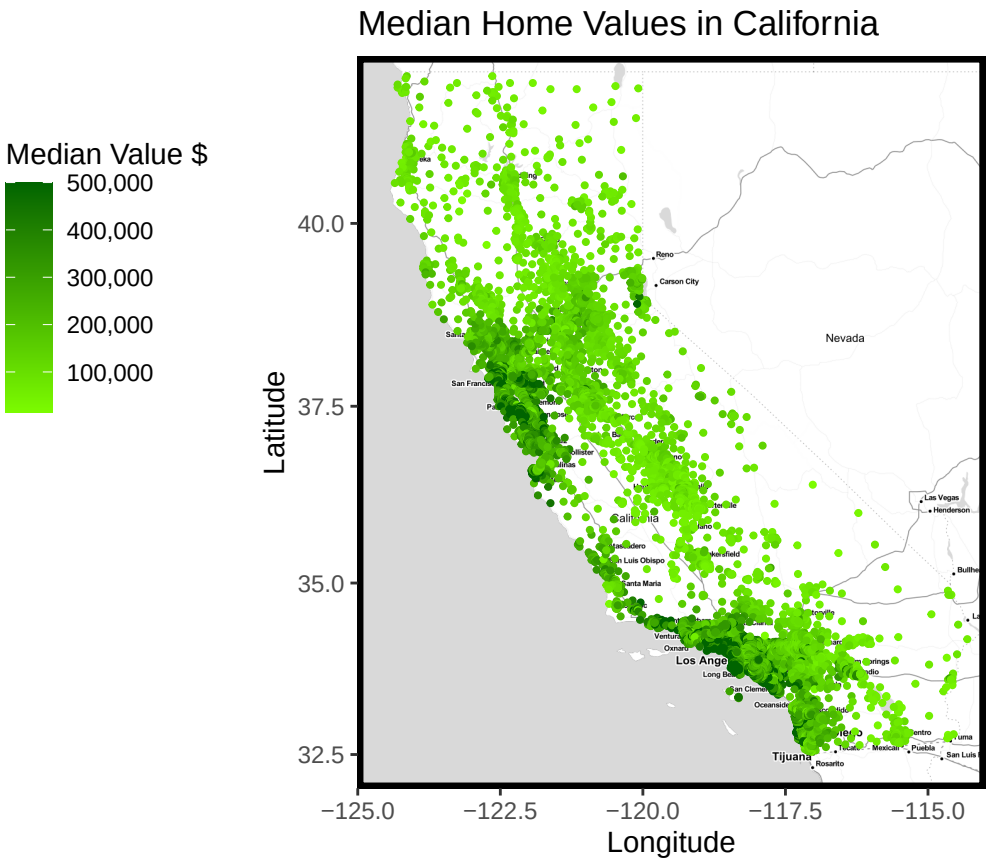
So, from our results we found a building having a green certification doesn't really do much to increase rental income per square foot. We definitely thought it would have more of an effect, so we found our results very interesting. There are much more important variables like *size* and *age* that seem to make a significant difference of *revenue*.

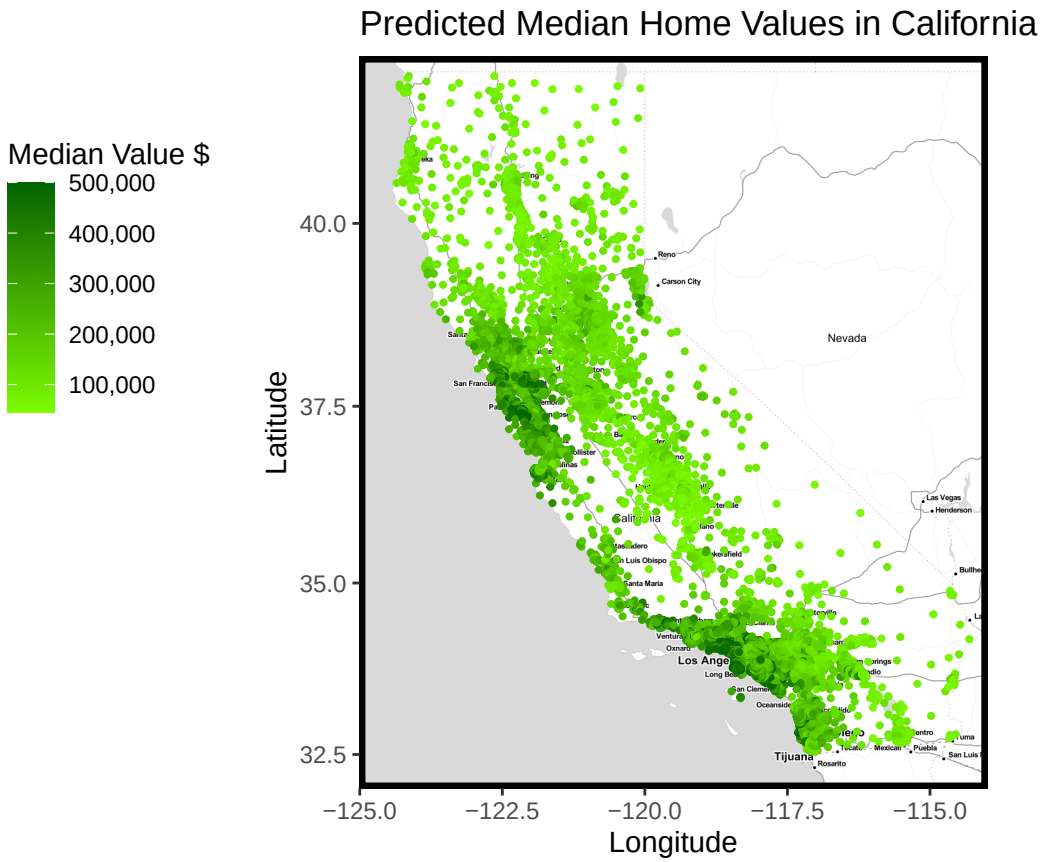
Predictive Model Building: California Housing

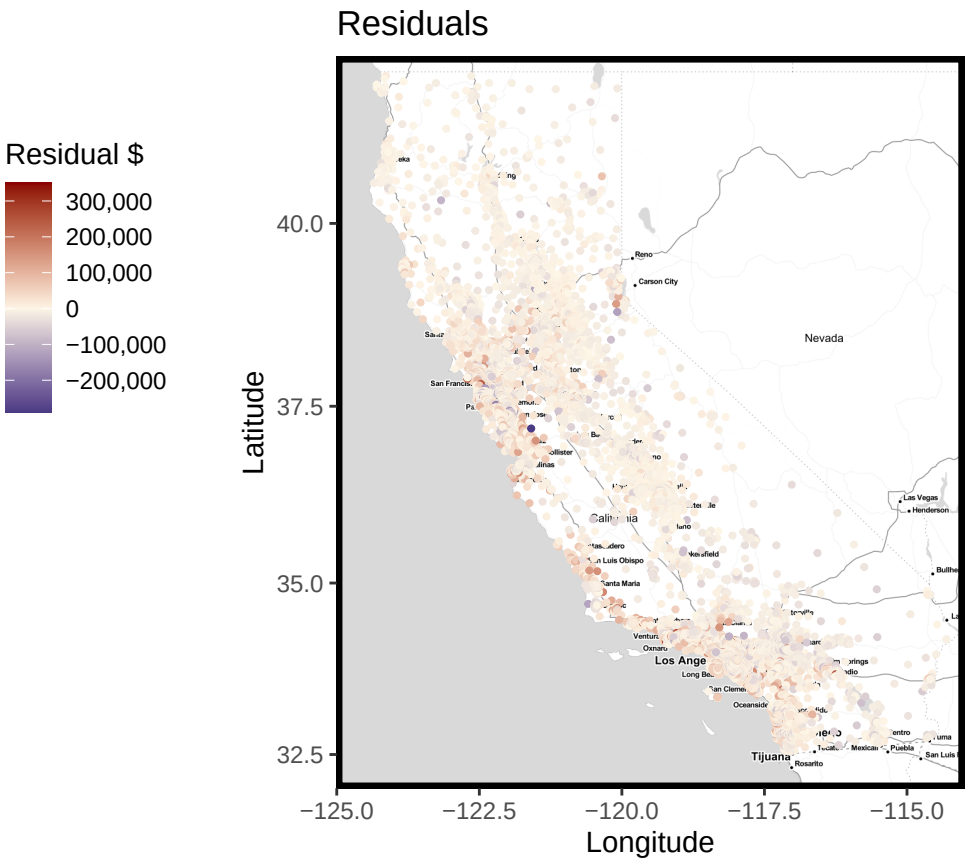
We now want to build a model to predict median home values in the state of California. We first take the variables *totalRooms* and *totalBedrooms* and divide these both by *households* to create two new variables *avg_rooms_per* and *avg_bed_per*. These variables give us the average number of rooms and bedrooms per household in each tract. We also create the variable *avg_house_size*, which is the average household size in each tract. We now want to create a linear model and a random forest model and compare the accuracy of each. We use all the variables include except *totalRooms* and *totalBedrooms* and include the variables we created. We run each model in a loop 10 times and take the average RMSE from both to help get a better idea of the real accuracy of the model. Here are the results,

Model	***RMSE***
Linear Regression	70046
Random Forest	48784.45

We see the random forest model does a much better job, so we will use this to create our predicted values to map using the package "ggmap". So, we run our model then predict median home values. We also want to create a map with residuals, so we create a residual variable by subtracting the predicted values from the real values. We display each of our maps over the next three pages. The first map is a map of the real data. As you would expect, the higher home values are mostly near the coast and near the large cities. Then second map is our predicted median values. We see it is pretty similar, so it seems like the model has does a decent job just by comparing the maps. Our last map is the residuals. We use two dark colors for the extremes and a light color for close to zero to try and better identify where the higher error is coming from and if these are more over-predictions or under-predictions. A positive value in this case would mean the actual value is higher than the predicted value, so red represents under-predictions. Just from looking at the map, it seems there are more under-predictions than over-predictions. Also, most of the inaccuracy seems to fall near the more populated areas.


Median Home Values in California
Median Value $
500,000
400,000
300,000
200,000
100,000
Latitude
40.0
37.5
35.0
32.5
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Nevada
Reno
Carson City
Las Vegas
Henderson
Los Ange
Tijuana


Predicted Median Home Values in California
Median Value $
500,000
400,000
300,000
200,000
100,000
Latitude
40.0
37.5
35.0
32.5
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Nevada
Reno
Carson City
Las Vegas
Henderson
Los Ange
Tijuana


Residuals
Residual $
300,000
200,000
100,000
0
−100,000
−200,000
40.0
37.5
35.0
32.5
Latitude
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Nevada
Reno
Carson City
Las Vegas
Henderson
Los Ange
Tijuana