Problem Set 3

By Jae You (jhy479)

What causes what?

1. As the podcast mentioned, the issue is answering the question "What causes what?", and we wouldn't be able to know from the data whether "Crime" causes an increase in "Police" or whether an increase in "Police" causes a decrease in "Crime".
2. The researchers used a case where there were lots of Police in D.C. for reasons unrelated to crime in order to isolate correlation from causation. By law, Washington D.C. is required to have more police when there is a high terror alert, which is unrelated to street crimes. Thus, the researchers observed whether there was a decrease in crime on days of high alert, when there were more police patrolling D.C., and their result was that there is a causal relationship in D.C. between more police and less crime. In table 2, the coefficient is negative and statistically significant, meaning that crime decreased on days with High Alert.
3. The researchers considered a case where people wouldn't go out as often during days of High Alert, which would lead to less victims being potential targets of crime and thus a decrease in crime that was not related to the "Police" variable. So they had to control for Metro ridership levels to see if they were diminished on days with High Alert. The result was that there were the same number of tourists, so a decrease in crime rates would not be related with a drop in the level of people traveling outside during High Alert days.
4. The model now has interactions between high terror alert days and district 1 and between high alert and other districts. The result is that the effect of an increase in "Police" affecting "Crime" is more present in District 1 than in other districts, as the interaction between High Alert and District 1 is statistically significant at the 1% level, while the interaction between High Alert and other districts is not statistically significant even at the 5% level.

Tree modeling: dengue cases

I predicted the total_cases with the variables city+season+specific_humidity+tdtr_k+precipitation_amt.

I measured the performances based on the mean RMSE of each model, which was cross validated by getting the results for each model 10 times, which is the default.

Here is the RMSE for CART:

```
## [1] 30.01389
```

Here is the RMSE for Random Forests:

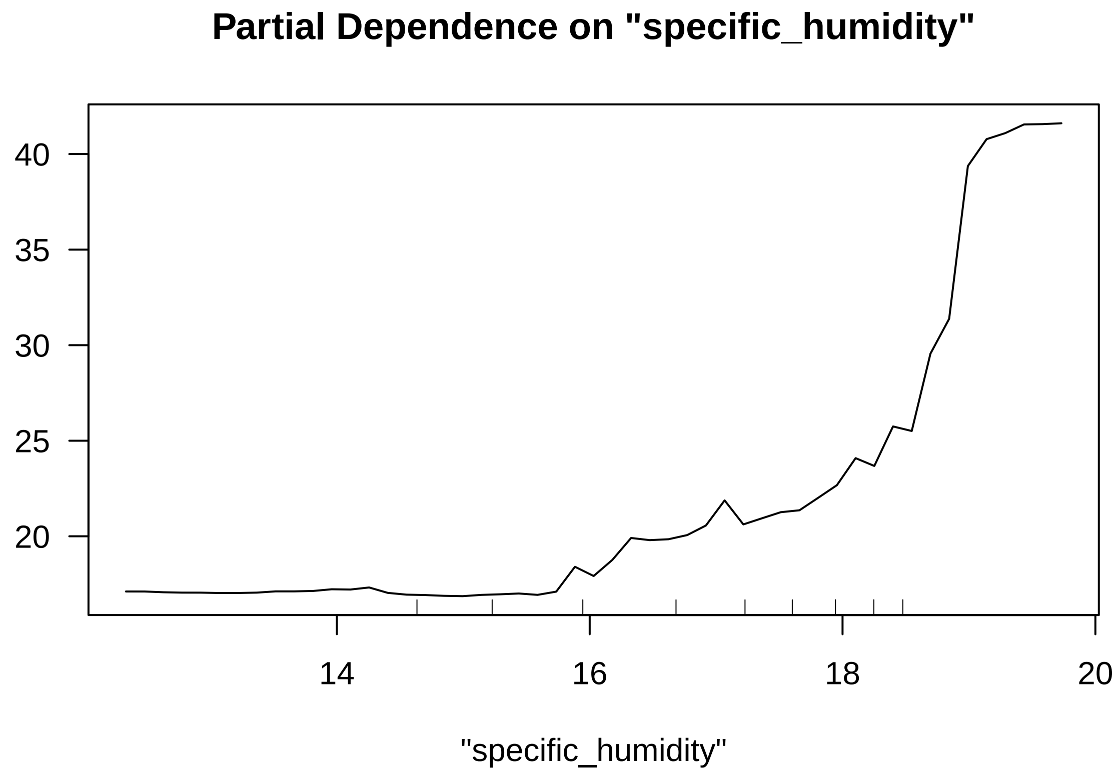
```
## [1] 30.40079
```

And here is the RMSE for Gradient-boosted trees:

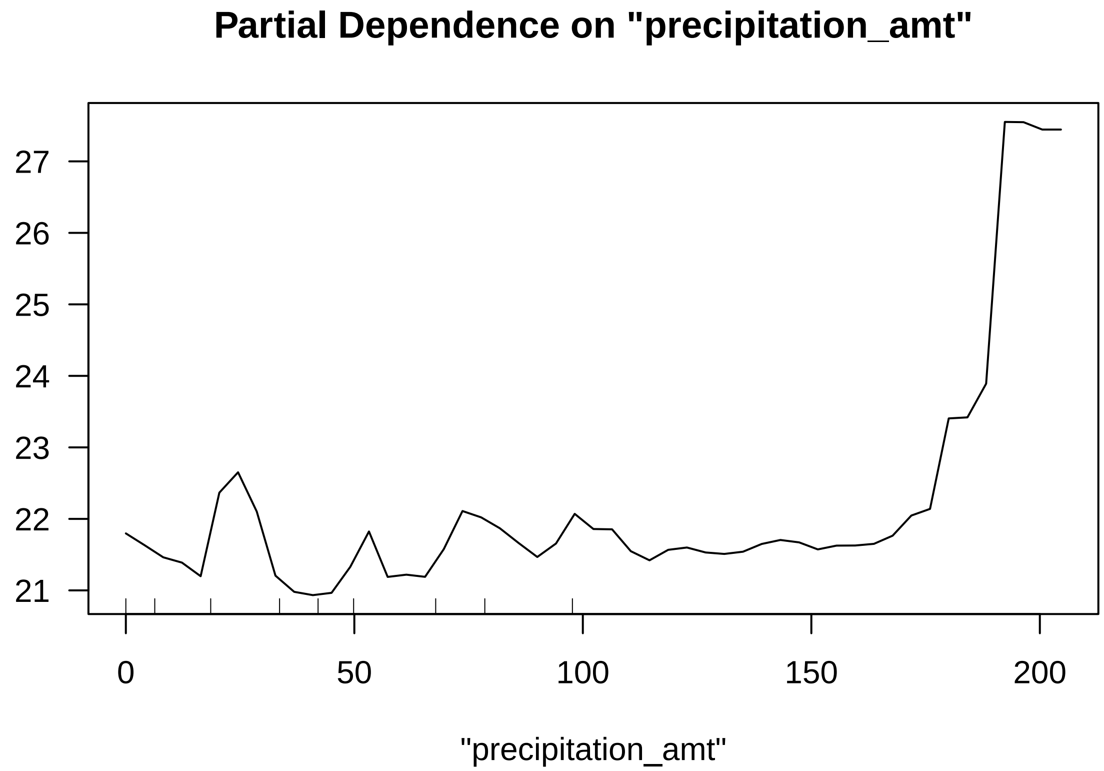
```
## [1] 25.11549
```

So I would conclude that the Random Forest model best fits the data out of the three models, as it has the lowest mean RMSE.

Next, I made three partial dependence plots for the Random Forest model.

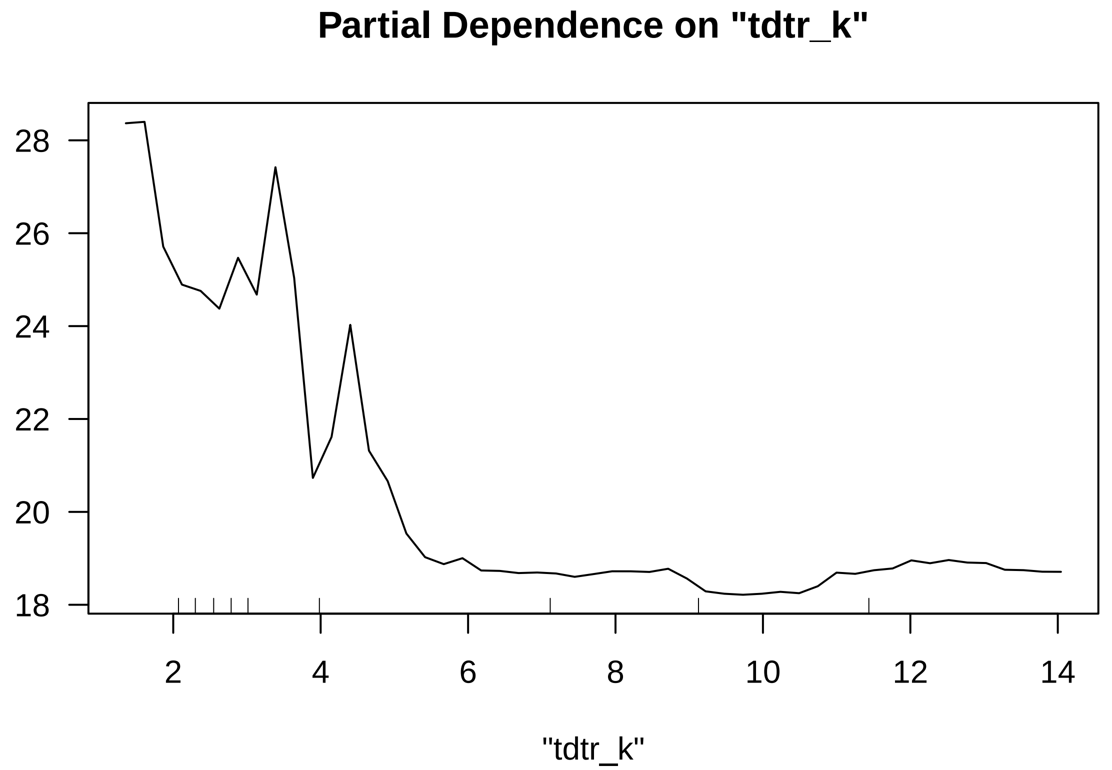


For the plot on specific humidity, there is a steady upwards trend until a spike at ~19g/kg of air, which may suggest that mosquitos are very prevalent at a high humidity.



For the plot on precipitation amount, it mirrors the previous plot of high humidity correlating with high cases, as the relative humidity will increase if there is a ton of precipitation.

There is a notable spike at very low amounts of precipitation, near 25mm and 75mm of rain. Although I am not sure what could cause these effects to be observed, I would assume that it has something to do with the phenomenon that humidity can be very high without raining if the water vapor does not cool.



For the third plot, I made a partial dependence plot on the "tdtr_k" variable, which is the Average Diurnal Temperature Range for the week. I found that the total number of dengue fever cases increased was higher when there is little difference between the maximum and minimum temperatures for the day and decrease as the difference between the max and the min grows wider.

This inverse relationship between average DTR and total cases was very interesting for me. Under the assumption that humidity and mosquitos have a strong, proven correlation, I would guess that days with huge temperature differences are also days that are not humid, while the days with fewer temperature fluctuations may be days with high humidity and thus would be prone to more mosquitos.

Predictive model building: green certification

In order to build the best predictive model for revenue per square foot per calendar year, I compared the performances of the linear regression, CART, and random forest models to see which model had the best fit for the data.

I ran each model 10 times and took the mean RMSE to measure the performance for each model. My dependent variable, "revenue_sqft", was Rent * leasing_rate, and I decided to remove several independent variables: rent and leasing_rate so that they did not "double down", cd_total_07 and hd_total07 because we have a variable "total_dd_07" that has the combined amount of both cooling and heating degree days, and the CS_PropertyID variable. I kept the indicators LEED and Energystar separate because the value of certain brands may have an impact on the revenue that the green building would earn, so these two indicators may bring about different revenues.

After back and forth testing, I decided to keep the variable "Precipitation" in the model despite it not really affecting the RMSE much, as the amount of rainfall in certain locations may impact the costs of constructing and maintaining a green building. Since the Precipitation variable did not affect the mean RMSE much, I suspect that the data was not from places with too much or too little rain, which does make sense because the United States is not know for it's extreme climates, unlike some places in Africa. However, I still decided to keep the variable in my model despite the data coming from properties across the United States, as the RMSE was indifferent between keeping and excluding the variable, and I preferred having a general model.

Here is the mean RMSE of the linear model:

```
## [1] 1051.798
```

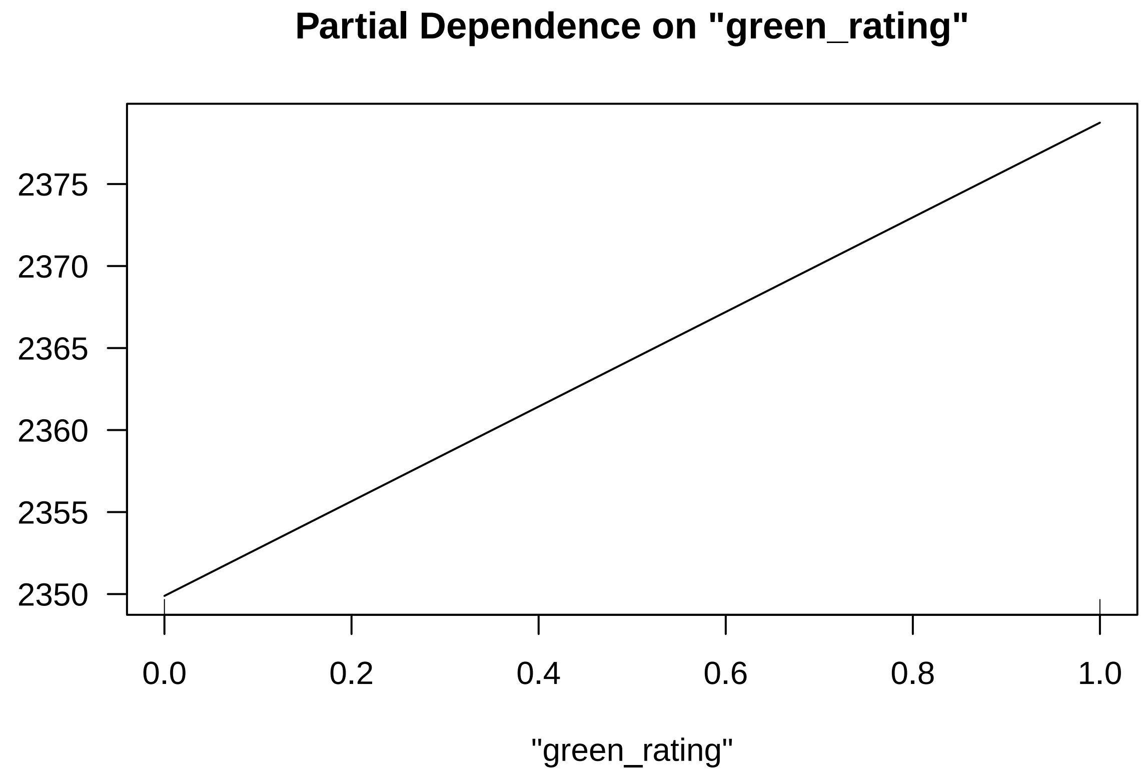
Here is the mean RMSE of the CART model:

```
## [1] 1240.507
```

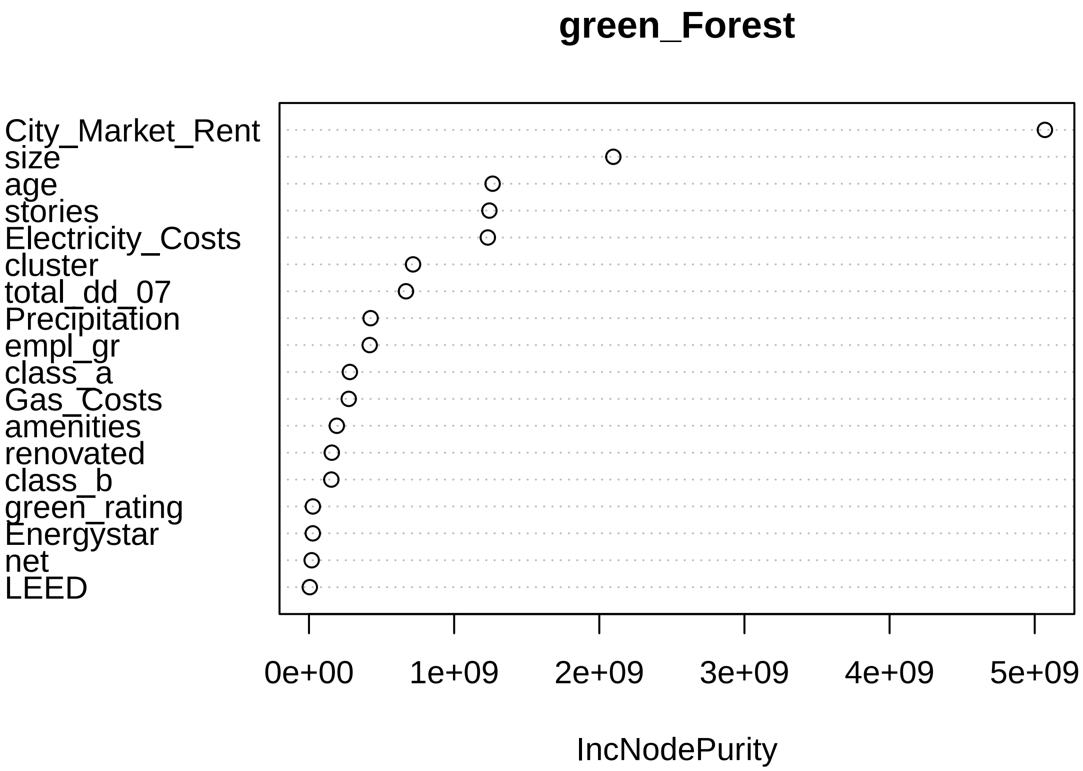
Here is the mean RMSE of the random forest model:

```
## [1] 715.1446
```

The random forest model is the best model with an RMSE of 858.55, beating the linear and CART models which are both around 1000. Thus, I used the random forest model for my partial dependence plot on green_rating.



Since the "green_rating" variable is binary, the difference in the y-axis values is 30, which is not that large on the amount of revenue per square foot that green certified buildings receive compared to non-certified green buildings.



The variable importance plot shows green_rating near the bottom, which means that the variable "green_rating" is not that impactful on the model's accuracy.

My conclusion is that companies may want to focus on other meaningful variables such as rent and size instead of using resources trying to become a green-certified building, as it seems that having the certification may not bring in as much revenue as you have invested, let alone profits.

Predictive model building: California housing

In order to build the best predictive model for the median market value of all the households in California, I compared the performances of the linear regression, CART, and random forest models to see which model had the best fit for the data.

I ran each model 10 times and took the mean RMSE to measure the performance for each model. I standardized the "totalRooms" and "totalBedrooms" variables by households and decided to remove "totalRooms" and "totalBedrooms" from the model so that they did not "double down." I kept the rest of the variables because I felt that they were important for predicting the median house value, as age and income are likely factors that influence what type of house one would get.

Here is the mean RMSE of the linear model:

```
## [1] 69131.45
```

Here is the mean RMSE of the CART model:

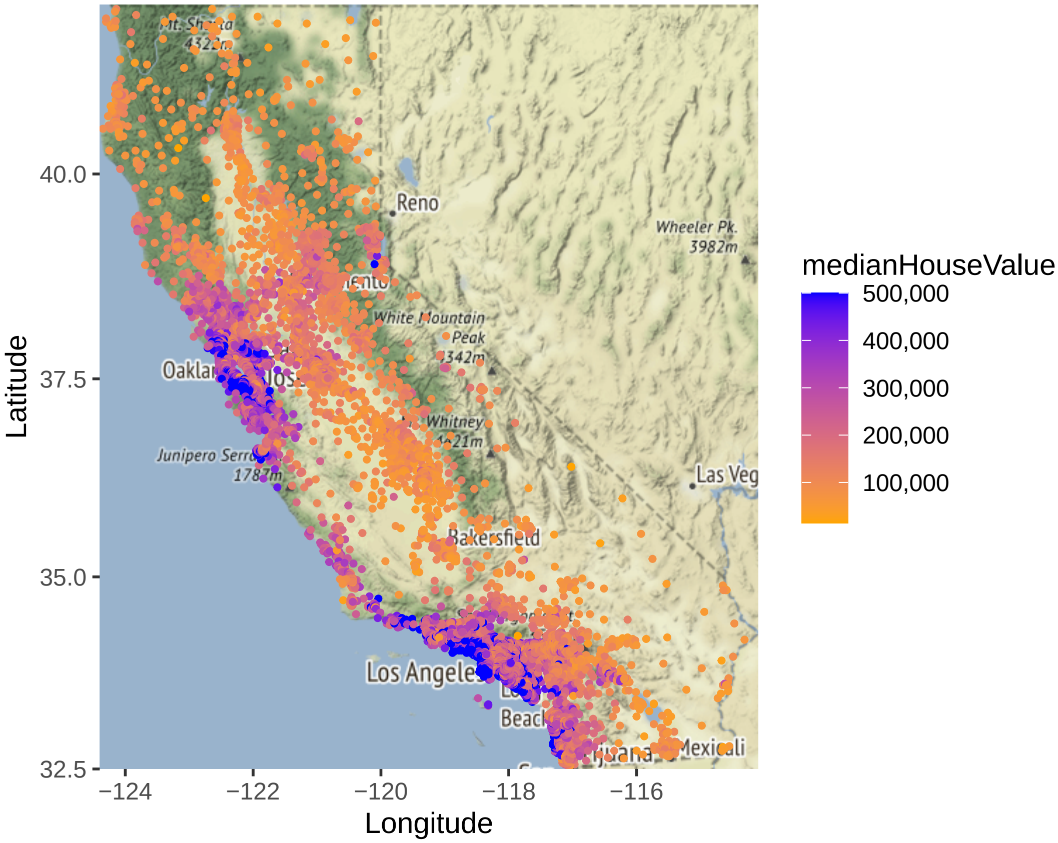
```
## [1] 64314.46
```

Here is the mean RMSE of the random forest model:

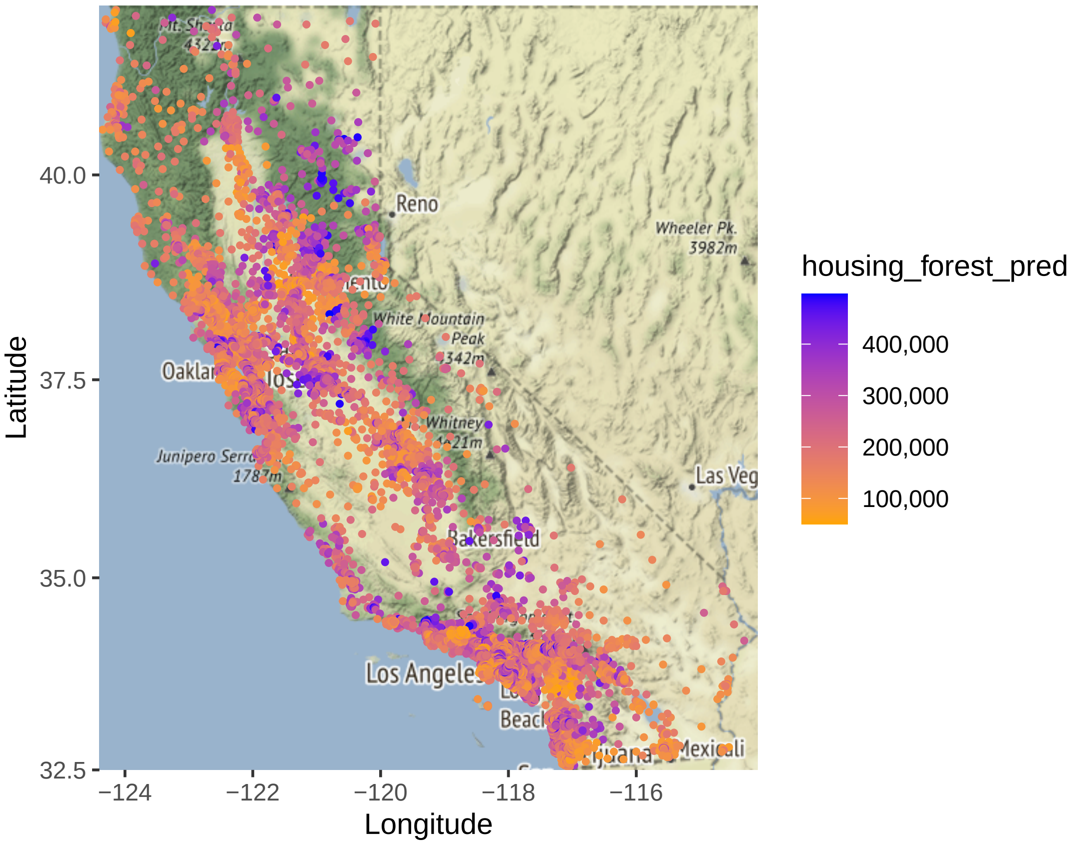
```
## [1] 51359.43
```

The random forest model is the best model with an RMSE of 51300, beating the linear and CART models which are around 69000 and 63000 respectively. Thus, I used the random forest model for my plots of the predictions of "medianHouseValue" as well as the errors versus longitude and latitude.

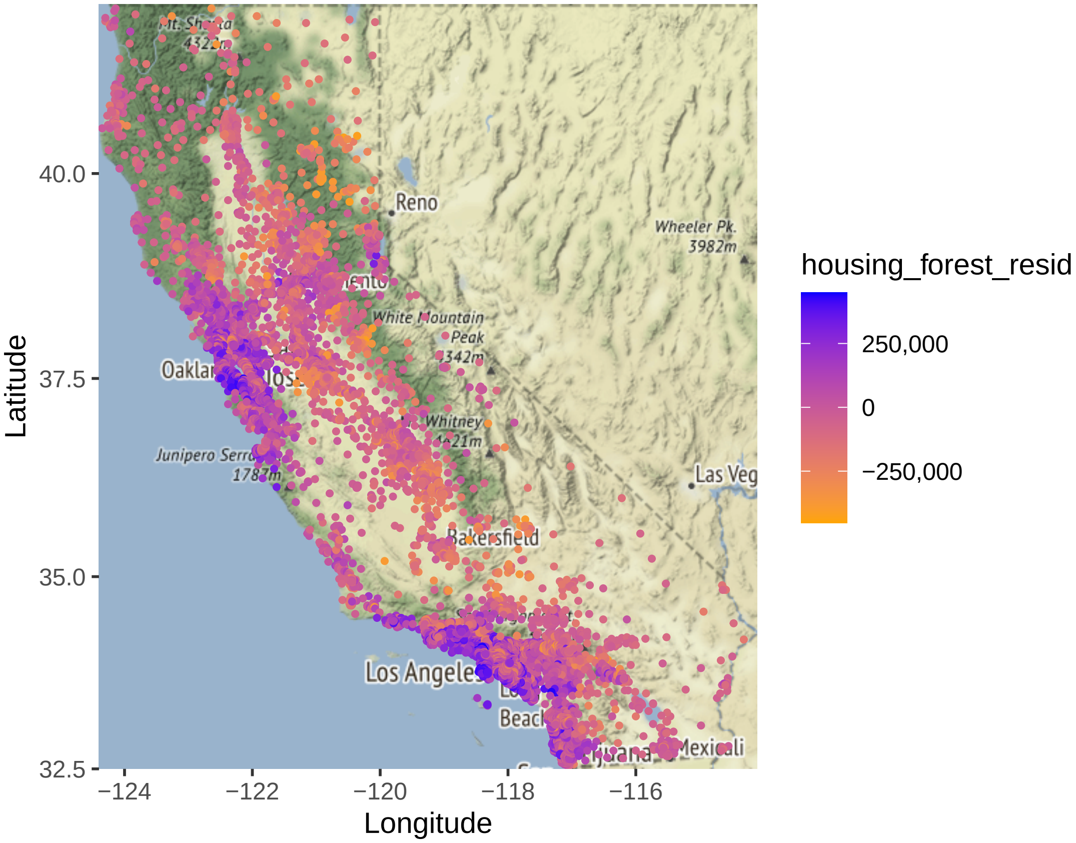
For the plots, I tried to use colorblind-friendly colors that also showed the contrast between low and high housing values.



On the plot of the original data, I noticed that the homes with the highest market value are on the coast, which makes sense because people may wish to have a nice view of the ocean, or tourists may come to California for their beaches. Additionally, the homes near the cities are relatively more expensive than those in rural or mountainous areas.



The second plot is of the random forest model's predictions, which has far less homes dotted than that of the original data's plot. This prediction plot also seems to suggest that the southern coastal area near Long Beach has higher housing valuations than the north-west coast. As there is both Los Angeles and Long Beach, this does make sense from a tourist perspective: more tourists in the area would lead to a rise in housing prices.



I plotted the random forest model's residuals, which was from medianHouseValue - predictions, versus longitude and latitude. Here blue represents a high residual, so the actual value is higher than the predicted value, while orange would mean that the predicted value was higher than the actual value. I noticed that, although the model did get a decent amount "correct" with the actual values from all the purple on the plot, there was still much to be desired. Specifically, the areas with large population such as the coastal regions and the Northern cities had many misses with the predictions. Los Angeles, in particular, had a lot of over-prediction on the values of homes.