Problem 1: What Causes What?

1. It's a question of causation, and it's hard to draw definitive conclusions just by looking at the size of police forces. Cities with above-average crime rates may hire more police officers than average cities, which can lead to a false conclusion that police are ineffective at solving crime. On the other hand, having more police means that crime is more easily detected, which may lead some to conclude that crime is higher, when in fact it may be the same as in any other city with fewer police. In short, it is hard to say what impact increased policing has had on crime.

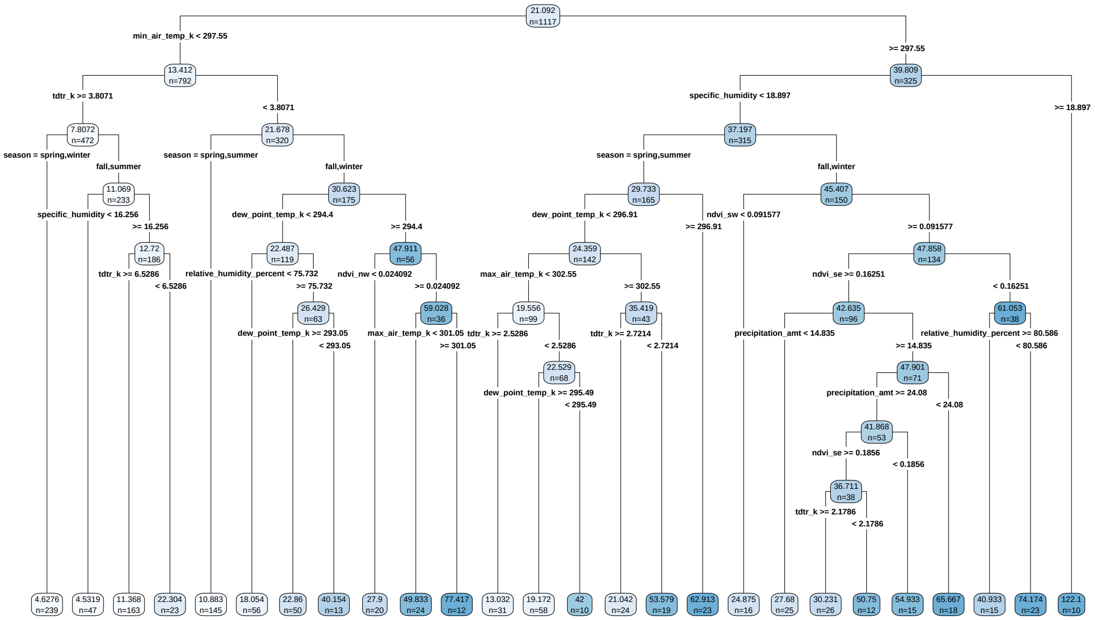
2. Basically, the researchers added an IV variable by using the terror alert system. The high terror alert means that no matter how many crimes are committed in a particular area, the police presence will increase. In their first return, they found that the high terror alert was expected to reduce the number of crimes by about seven. In the second return, which controls subway ridership, high terror alerts are expected to reduce crime by about six.

3. The researchers decided to control subway ridership to ensure that the low crime rates resulting from high terror rates were not just a matter of fewer people going out and walking the streets. The researchers tried to capture the effect of high terror rates on the number of people living in cities.

4. The first column includes a linear model using robust regression with three coefficients. One factor only takes into account the effect of a high terror rate within the First police District, the National Mall. That's because if terrorists were to attack Washington, D.C., they would likely focus on this area. The next factor is the effect of high alert on other police districts in Washington, DC. The third coefficient is the logarithm of subway ridership at noon. Basically, this regression suggests that heightened vigilance (and therefore increased police numbers) mainly reduced crime in the National Mall area, and the effect in the rest of the city was not as profound as in other areas, even if it still reduced crime by a small amount. However, the return still shows strong evidence that more officers can lower crime, and that's because the D.C. police force is likely to add the most officers in Ward 1 during periods of heightened alert.

Problem 2 Tree Modeling: Dengue Cases

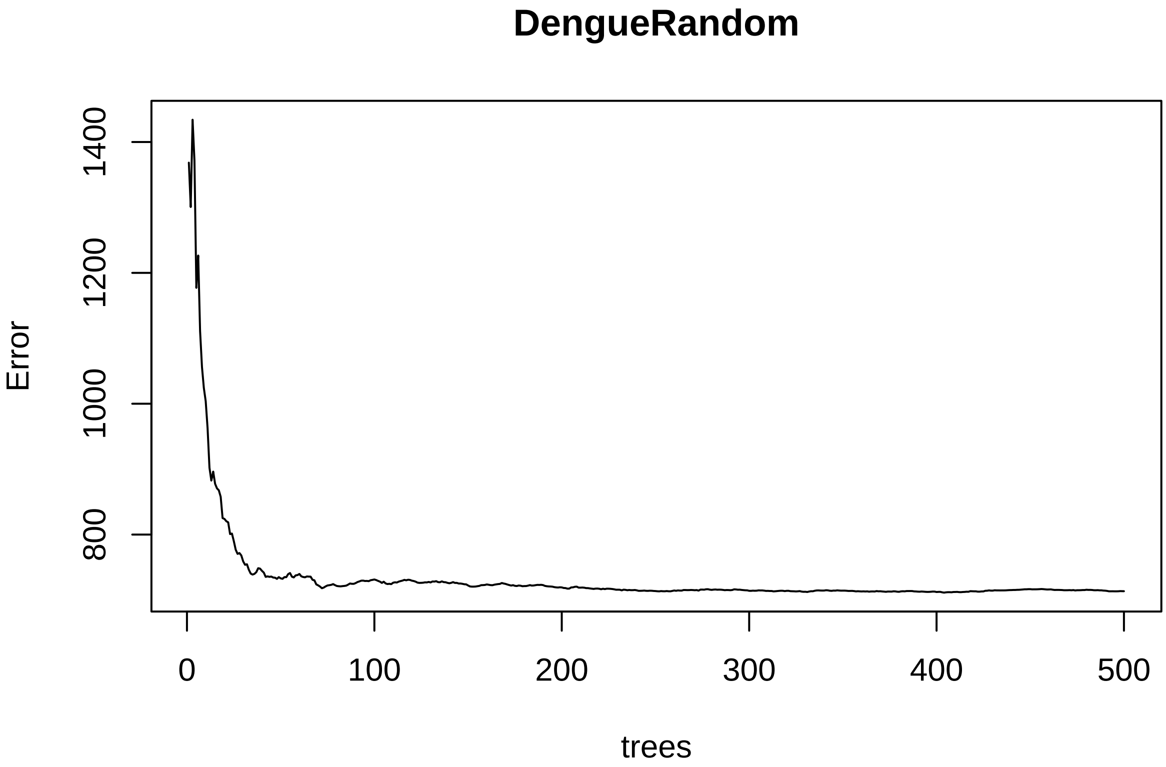
Part 1: CART



The model above shows the un-pruned CART Tree, we will proceed to prune and then calculate RMSE.

```
## 25.87936  RMSE for Pruned CART Model
```

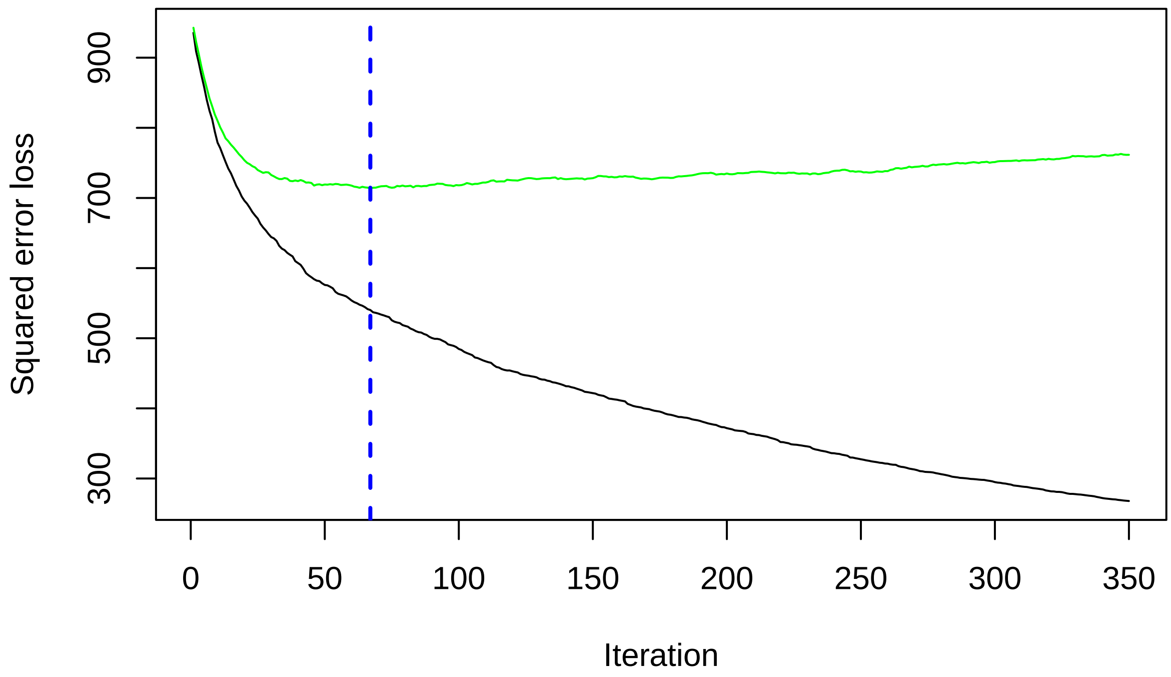
Part 2: Random Forest



This plot shows the out of bag MSE as a function of the number of trees used. Let's proceed to look at the RMSE compared to the testing set.

```
## 24.41752  RMSE for Random Forest
```

Part 3: Gradient Boosted Trees



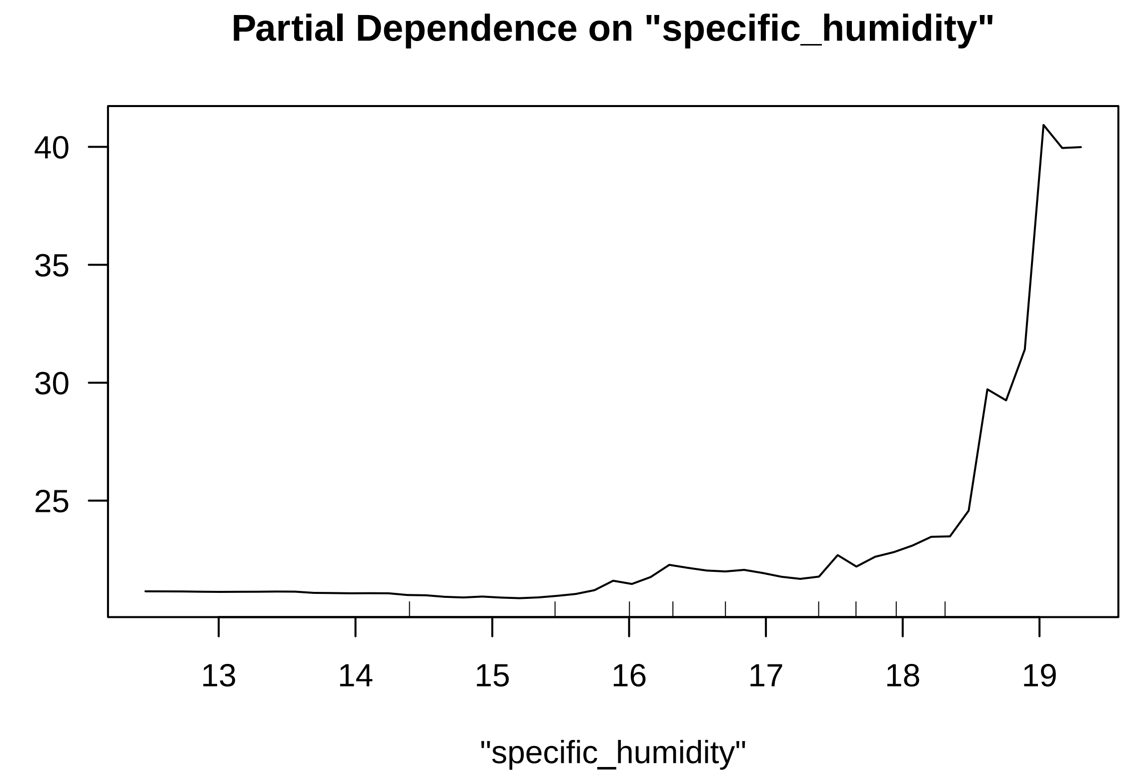
```
## [1] 67
```

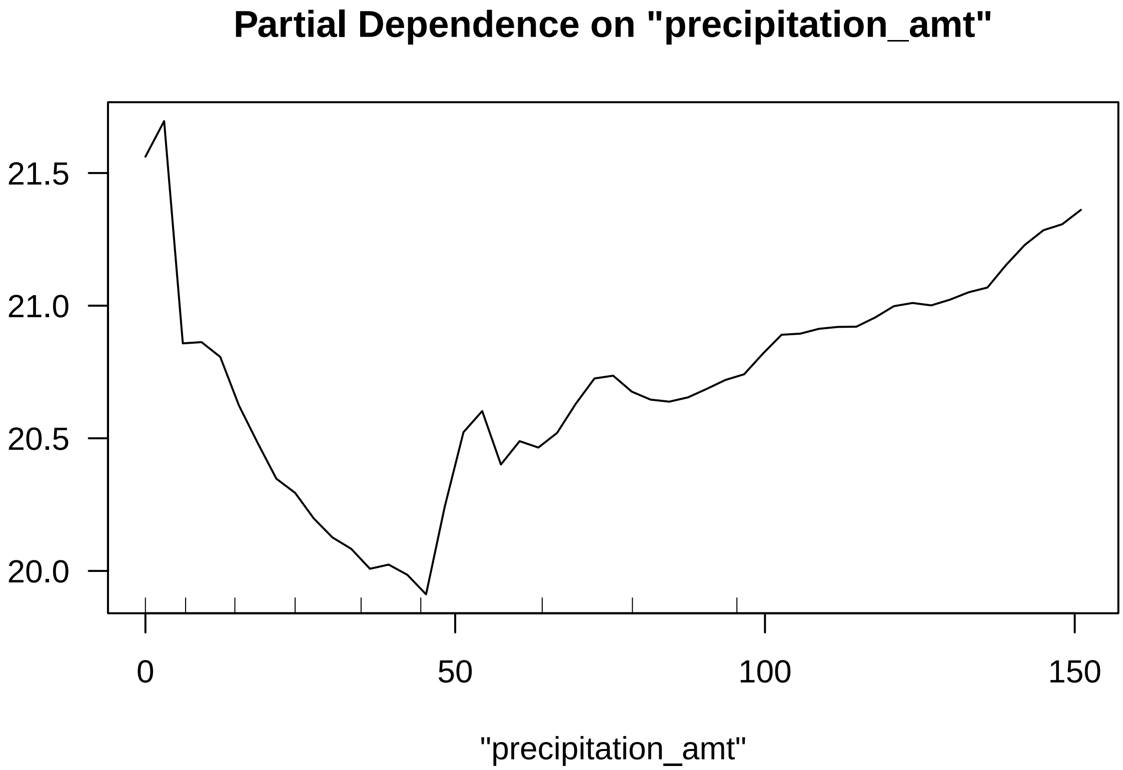
This plot shows the error curve of the Gradient Boosted Model, with the optimal number of trees listed as output. Let's now check the RMSE for the Gradient Boosted Trees Model.

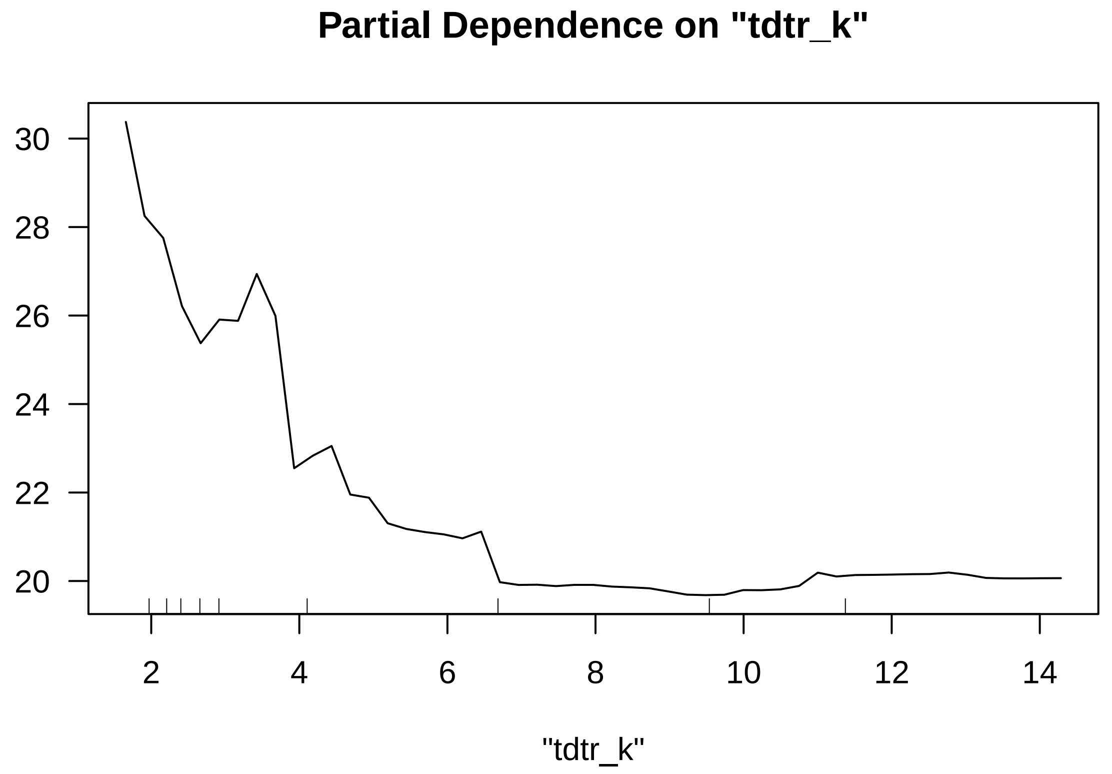
```
## 23.5897  RMSE for Gradient Boosted Trees
```

Looking at the RMSE results from the three models, it appears that random forest would be the best choice for this particular set of data. The next section shows the partial dependency plots for the Random Forest Model.

Part 4: Partial Dependency Plots







Wrap Up:

Looking at the PD plots, most seem to make sense in the context of the science of mosquito breeding. Mosquitos require standing water in order to make baby mosquitos, it makes sense that as precipitation increases, the number of mosquitos increases, the increased number of mosquitos leads to more cases of Dengue. The same seems to be true of humidity. Humidity is a measure of how much evaporated moisture there is in the air, higher humidity would seem to indicate that there is a higher amount of water on the ground, and thus the amount of mosquito breeding grounds. Our wild card PD plot looks at the Average Diurnal Temperature Range. It shows that as DTR increases, the amount of predicted Dengue cases decreases. This makes sense as well, it's possible that temperature shocks kill mosquitos which leads to less Dengue cases.

Problem 3: Green Certification

Introduction

This question asks us to quantify the effect of green certifications on revenue per square foot in buildings with such a certification. Green certifications clearly have an environmental impact, but do they also make buildings more attractive to potential renters, and people pay attention when buidlings recieve a green certification? We attempt to answer these concerns below.

Analysis

Data Cleaning Since we focused on the revenue of the property, we generated a new variable `revenue` as the product of `Rent` and `leasing_rate`. We created green_rating to see the overall impact of green certificate instead of using both `LEED` and `EnergyStar` (`green_rating` is a collapsed version of these two ratings). Other important factors that will affect energy usage are the cooling degree days and the heating degree days, in our model we used `total_dd_07` to represent this factor.

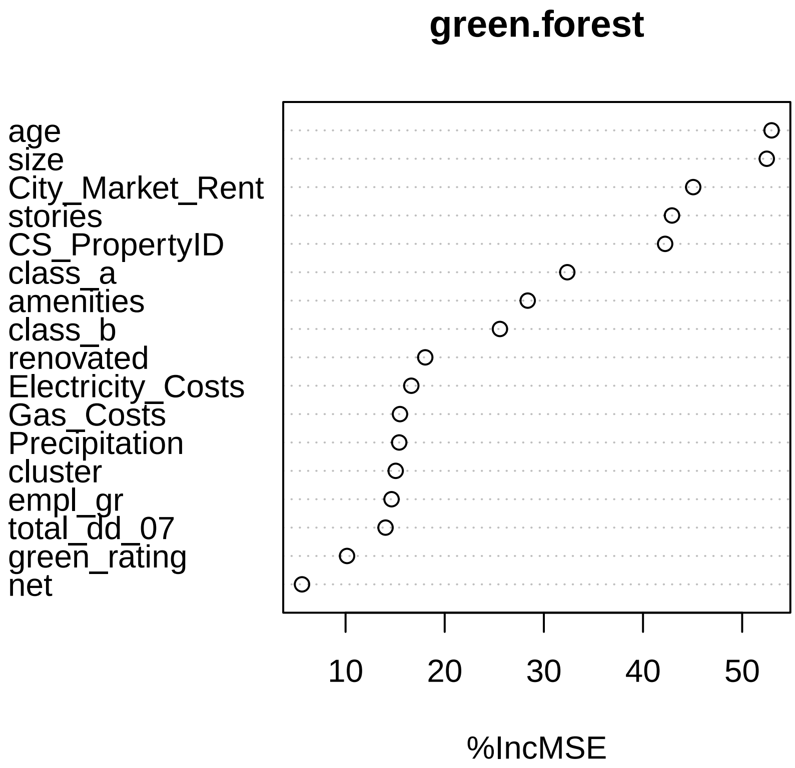
Modeling We tried three different models to see which one will yield the best out-of-sample RMSE, they are CART, random forest, and gradient-boosted model. We split the data set into training and testing sets, and we trained all of the three models on training data using all of the variables.

To tune the gradient-boosted model, we first created a grid that specifies `shrinkage`, `interaction.depth`, `n.minobsinnode`, and `bag.fraction`. then ran the gbm across all different combination of these parameters. After narrowing the tuning grid 2 times, the average RMSE across 10 folds is still higher than random forest. The final comparison are shown below.

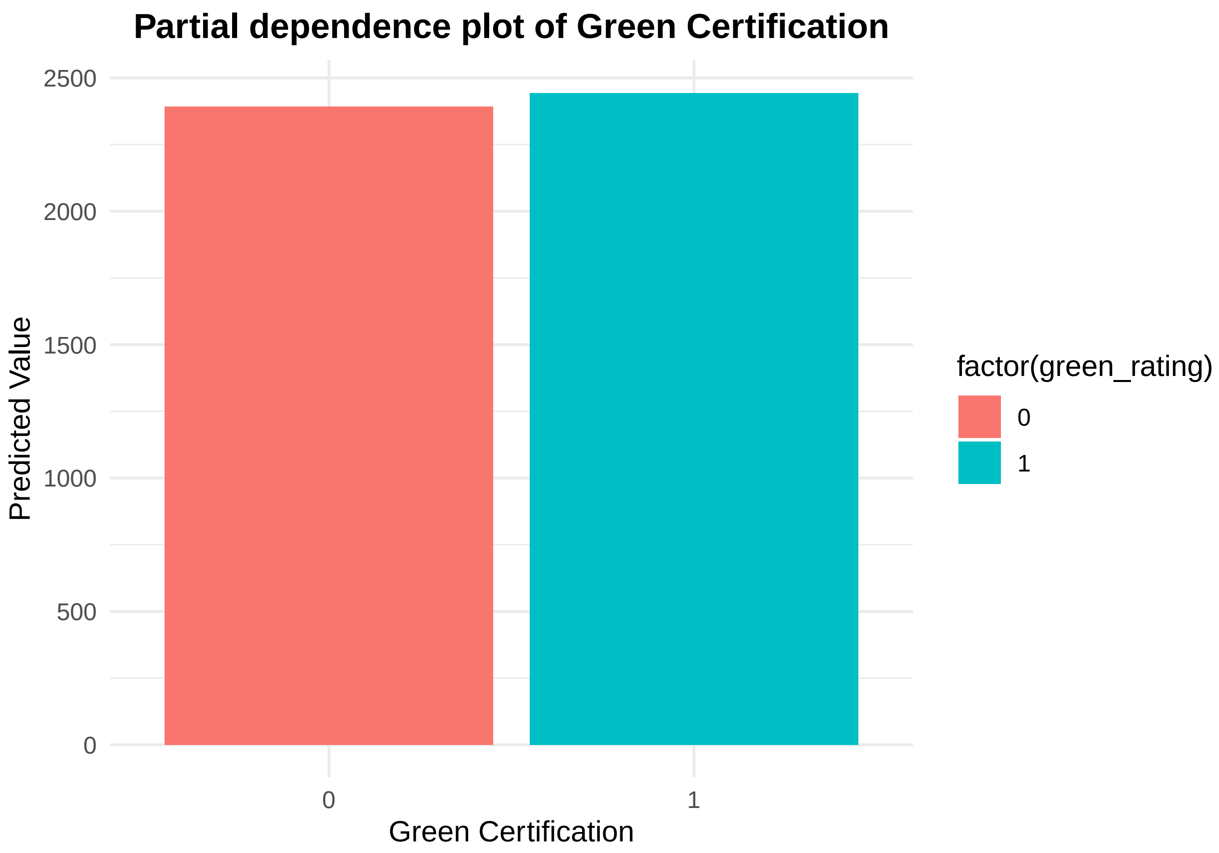
Model performacne The out-of-sample RMSE of the models: CART: 772.38 Random Forest: 609.07 Gradient-boosted: 683.56

So we decided to use random forest as our best predictive model.

Plots This is the variable importance plot for our random forest model. As you can see, size, market rent, and age seem to be the biggest factors in predicting. Our green rating variable is actually show to have the lowest importance out of all of the parameters.



As you can see, it appears that green rating is only estimated to increase revenue by fifty dollars.



Wrap up

After testing three models, we decided to employ random forest. Using our predictive modeling, we found that green certification doesn't really lead to a dramatic increase in revenue. According to our partial dependence plot, a green certification is only expected to increase yearly rent revenue (per square foot) by fifty dollars.

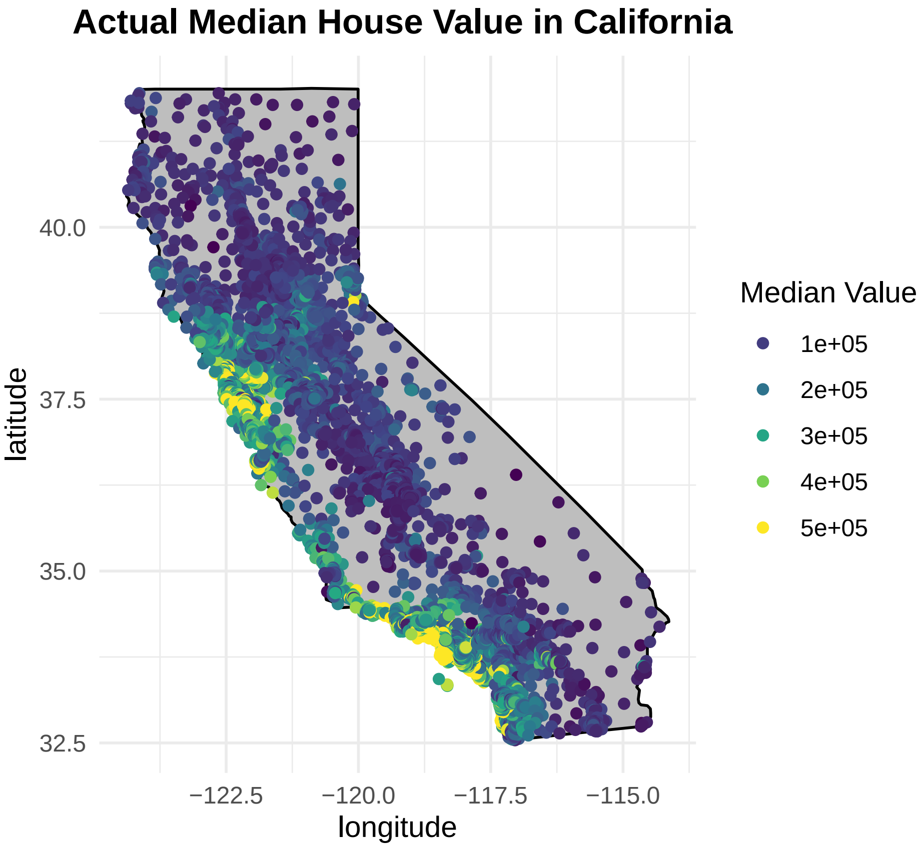
Problem 4: California Housing

Modeling Again in this question we will compare CART, random forest, and gbm to get the best predictive model. After tuning the gbm model, we get a out-of-sample RMSE that is smaller than random forest. So the best predictive model we will use here would be gbm.

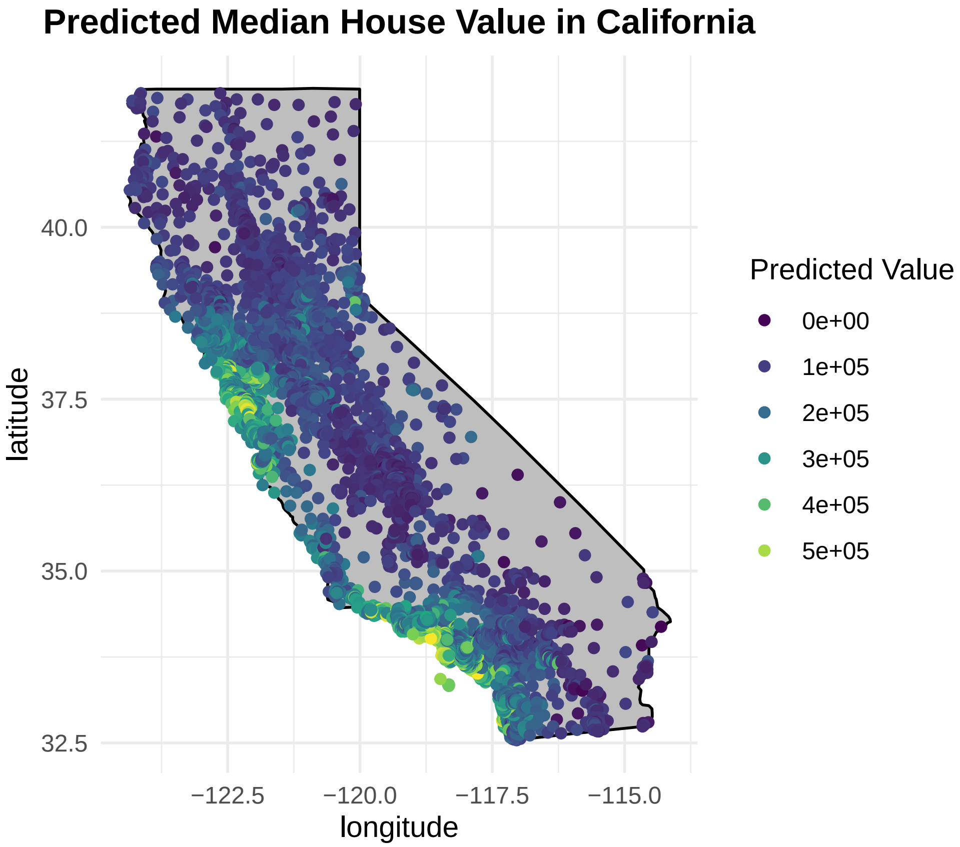
Now, we use the gbm model to produce a plot of prediction and a plot of model's residuals.

Plots To produce the three plots, first we set up the base plot of California using the data from the package `maps`.

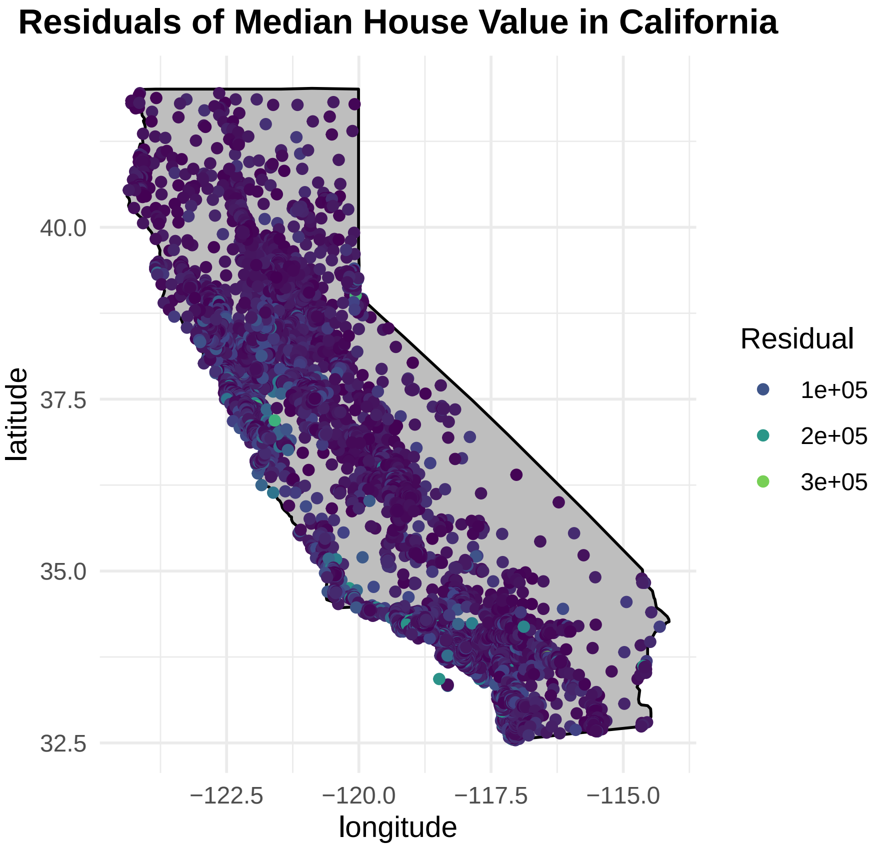
The plot displays the median house value in California with the colors becoming brighter as the house value increases. The plot clearly displays the higher home values of the Los Angeles and San Francisco Bay Area (including coastal suburbs), having the most concentrated collection of homes with high values.



This plot displays the predicted median house value from our model. As you can see, compared to the actual data, our model appears to do a good job at predicting house value.



This plot shows the absolute value of the residuals between actual and predicted values. It appears that most of the errors from our model are small.


Residuals of Median House Value in California
latitude
40.0
37.5
35.0
32.5
−122.5
−120.0
−117.5
−115.0
longitude
Residual
● 1e+05
● 2e+05
● 3e+05