Chia Sheng Tu (ct33633)
Yu Hsiang Tseng (ty4874)

Q1. What causes what?

Question 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?

The relationship between police force size and its impact on crime is complex, making it difficult to draw definitive conclusions. Factors such as a city's crime rate and the ease of crime detection with a larger police force can lead to misleading conclusions. For instance, a higher crime rate may prompt a city to hire more police officers, which could be misinterpreted as a sign of ineffective policing. Conversely, more police may make it easier to detect crime, which could falsely suggest higher crime rates compared to a city with a smaller police force. Therefore, determining the causal effect of increased policing on crime rates remains challenging.

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

In essence, the researchers incorporated an independent IV variable into their study by utilizing the terror alert system. A high terror alert triggers an increase in police presence in an area, irrespective of the actual crime rate. In their initial regression analysis, the study revealed that a high terror alert is associated with a decrease of approximately 7 crimes. Subsequently, in their second regression analysis, after controlling for metro ridership, the high terror alert remained significant and predicted to lower the crime rate by approximately 6 crimes.

Question 3: Why did they have to control for Metro ridership? What was that trying to capture?

To ensure that the observed decrease in crime rate during a high terror alert was not solely due to a reduction in the number of people in public areas, the researchers included metro ridership as a control variable. Their aim was to isolate the impact of a high terror alert on the population density in the city.

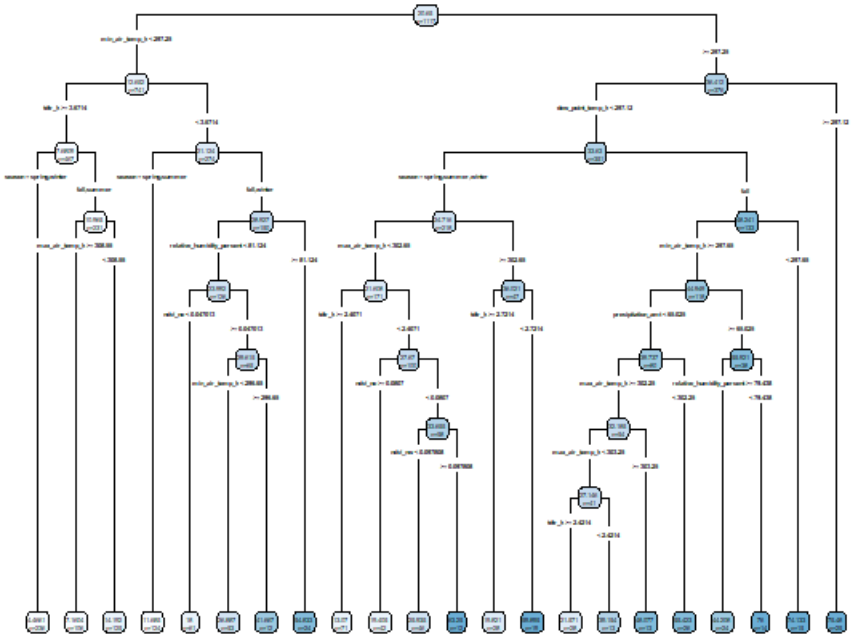
Question 4: Below I am showing you "Table 4" from the researchers' paper. Just focus on the first column of the table. Can you describe the model being estimated here? What is the conclusion?

The initial column of the regression table shows the results of a robust linear model

with three coefficients. One coefficient analyzes the impact of a high terror alert on the first police district area, which covers the National Mall. This area is particularly vulnerable to terrorist attacks, hence the focus of the analysis. The second coefficient measures the effect of a high alert on the remaining police district areas within DC. The third coefficient is the log of midday metro ridership. The regression analysis indicates that a high alert (and the accompanying increase in police presence) has a more significant impact on reducing crime in the National Mall area, whereas the effect in other parts of the city is less pronounced but still results in a modest decrease in crime. Overall, the regression provides strong evidence that increasing police presence lowers crime. This is particularly evident during a high alert, as the DC police force is likely to deploy more officers in district one.

Q2 Tree Modeling: Dengue Cases

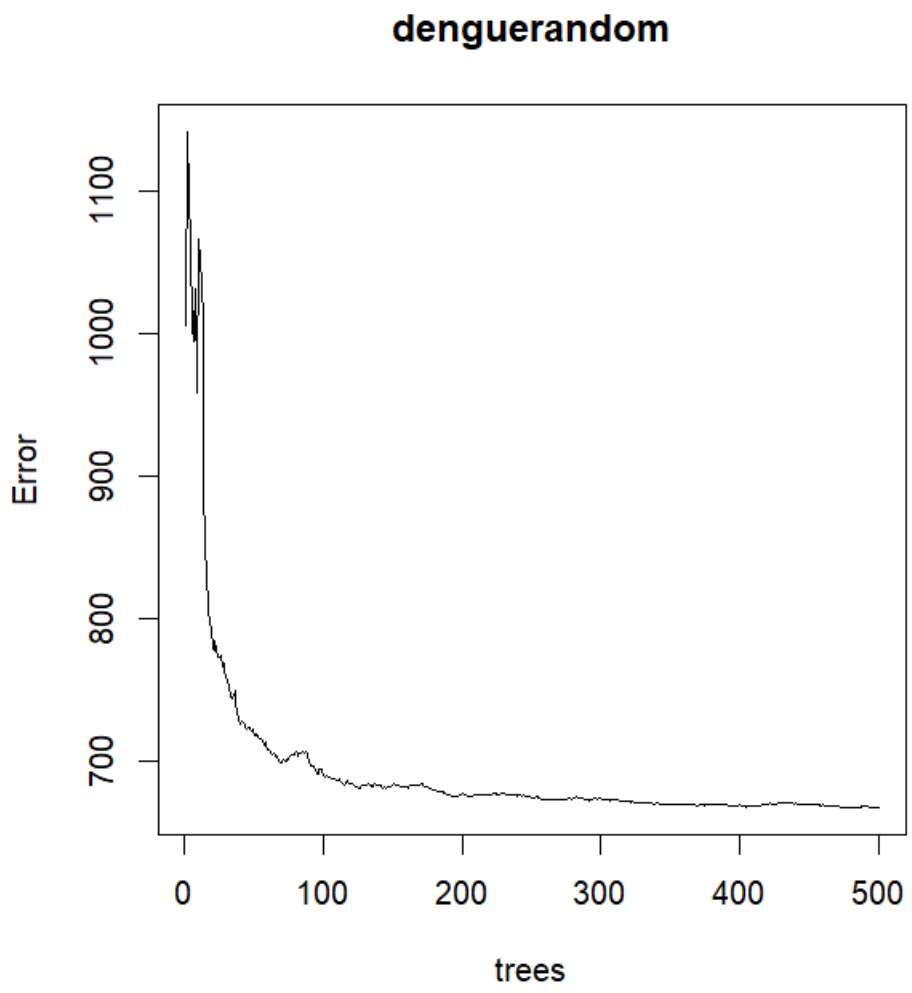
Part 1: Cart



After displaying the un-pruned CART Tree in the above model, our next step will be to prune it and subsequently compute the RMSE.

33.70785 RMSE for Pruned CART Model

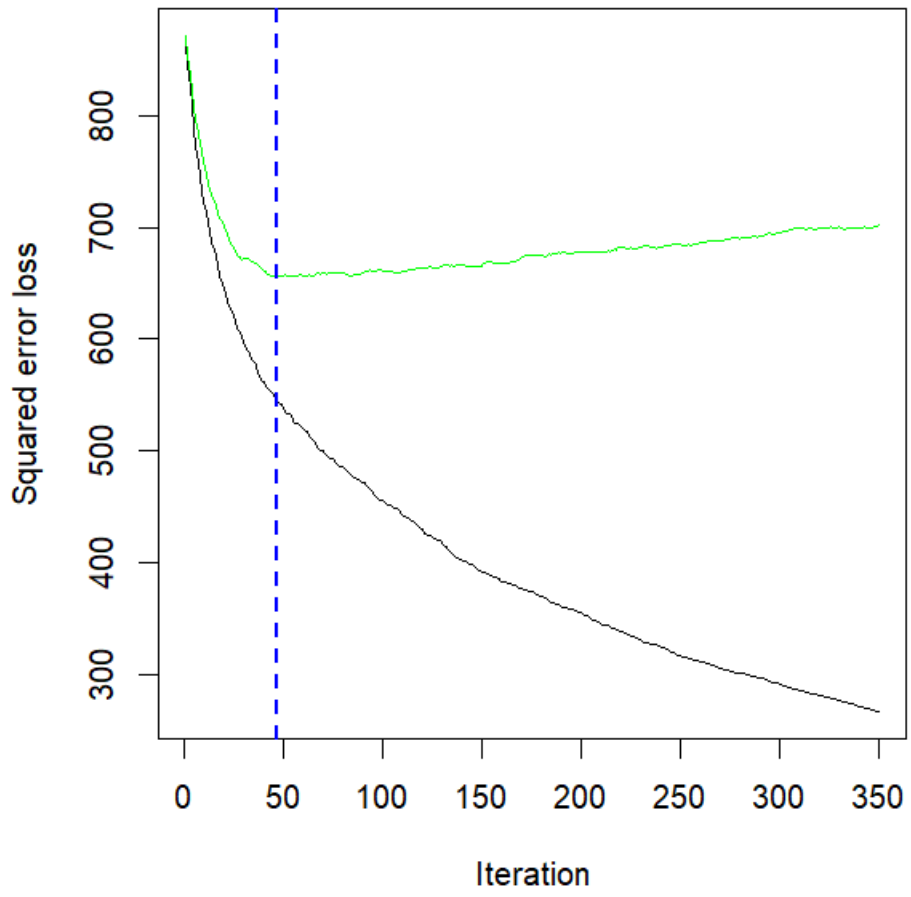
Part 2: Random Forest



The graph displayed illustrates the relationship between the out-of-bag MSE and the number of trees utilized. Next, we will examine the RMSE in comparison to the testing set.

```
29.35772    RMSE  for  Random  Forest
```

Part 3: Gradient Boosted Trees



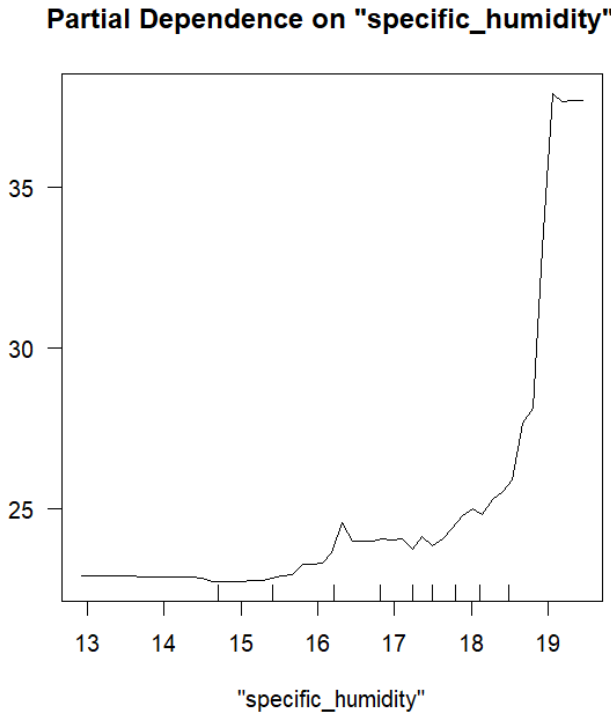

```
[1] 47
```

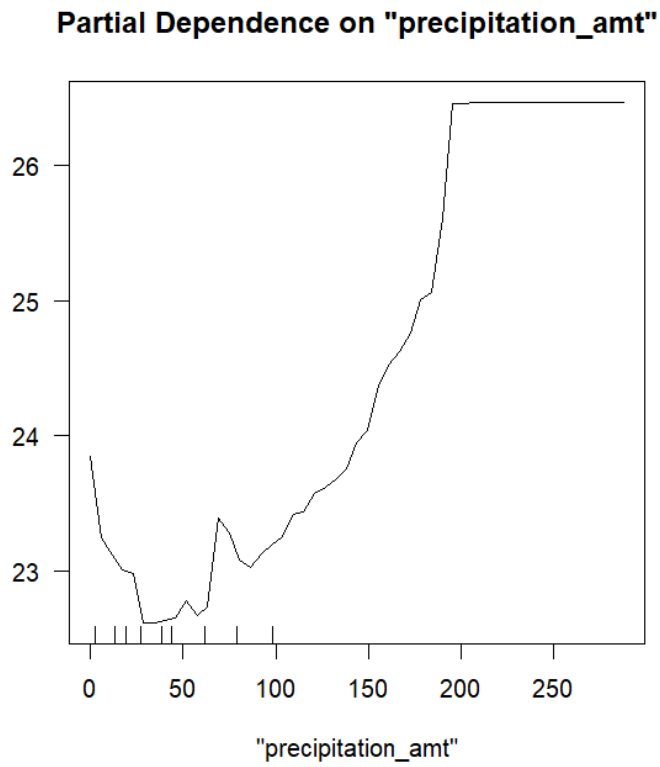
The depicted graph exhibits the error curve of the Gradient Boosted Model, with the optimal number of trees specified as the output. Moving forward, we will verify the RMSE of the Gradient Boosted Trees Model.

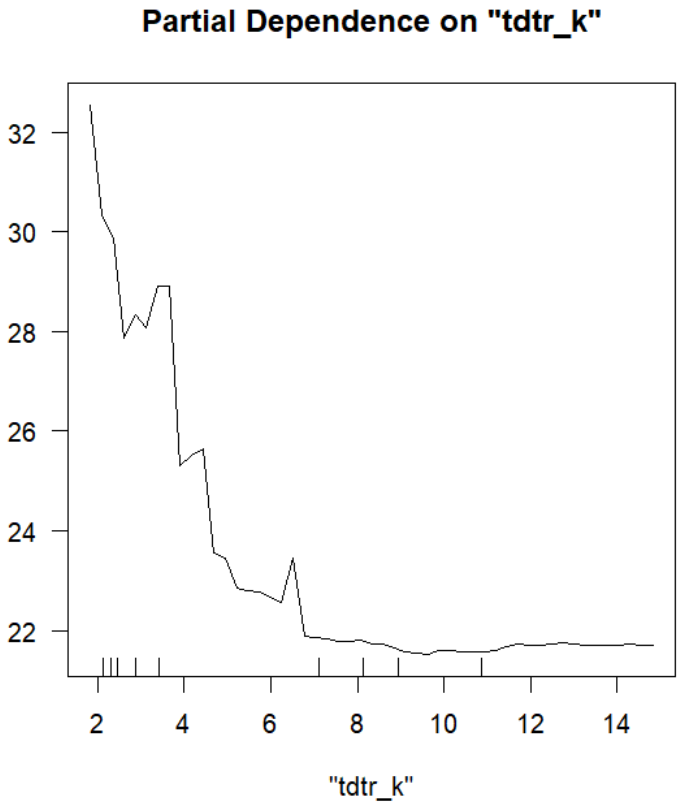
```
30.41227  RMSE for Gradient Boosted Trees
```

Based on the RMSE outcomes of the three models, it seems that the random forest would be the best option for this specific data set. The subsequent section will exhibit the partial dependency plots for the Random Forest Model.

Part 4: Partial Dependency Plots







Upon examining the PD plots, most of them seem to align with the principles of mosquito breeding. Since mosquitos require standing water to breed, it's logical that as precipitation levels increase, the mosquito population also increases, ultimately resulting in more cases of Dengue. Similarly, humidity, which is a measure of evaporated moisture in the air, also appears to be correlated with the number of mosquito breeding grounds and hence, the incidence of Dengue. However, the Average Diurnal Temperature Range PD plot seems to be a wildcard. It indicates that as DTR rises, the predicted number of Dengue cases decreases. This trend can be explained by the possibility that sudden temperature shocks may kill mosquitos, thereby leading to a reduction in the incidence of Dengue cases.

Q3. Green Certification

Introduction

This study aims to assess the impact of green certifications on revenue per square foot in certified buildings. Although green certifications are known to have environmental benefits, it remains unclear whether they attract potential renters and whether people pay attention to these certifications. In this study, we investigate these concerns.

Analysis

Data Cleaning

To analyze the revenue of the property, we created a new variable, revenue, which is the product of rent and leasing rates. We also created a variable, green_rating, to represent the overall impact of green certification, collapsing LEED, and EnergyStar ratings. We included cooling degree days and heating degree days as important factors affecting energy usage, represented by the variable total_dd_07 in our model.

Modeling

We compared the performance of three different models (CART, random forest, and gradient-boosted model) to determine the best out-of-sample RMSE. We divided the data set into training and testing sets and trained all three models on the training data, using all variables.

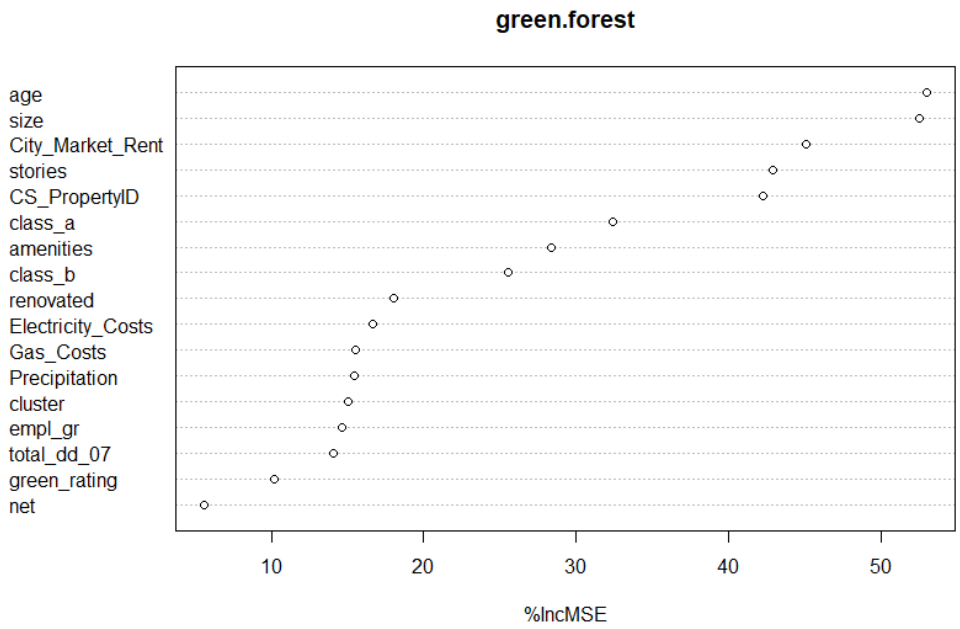
To optimize the performance of the gradient-boosted model, we created a tuning grid specifying shrinkage, interaction depth, n.minobsinnode, and bag.fraction parameters. We ran the gbm across all different combinations of these parameters and narrowed down the tuning grid twice. However, even after tuning, the average RMSE across 10 folds was still higher than that of the random forest model. The final comparison of the models is presented below.

Model performacne

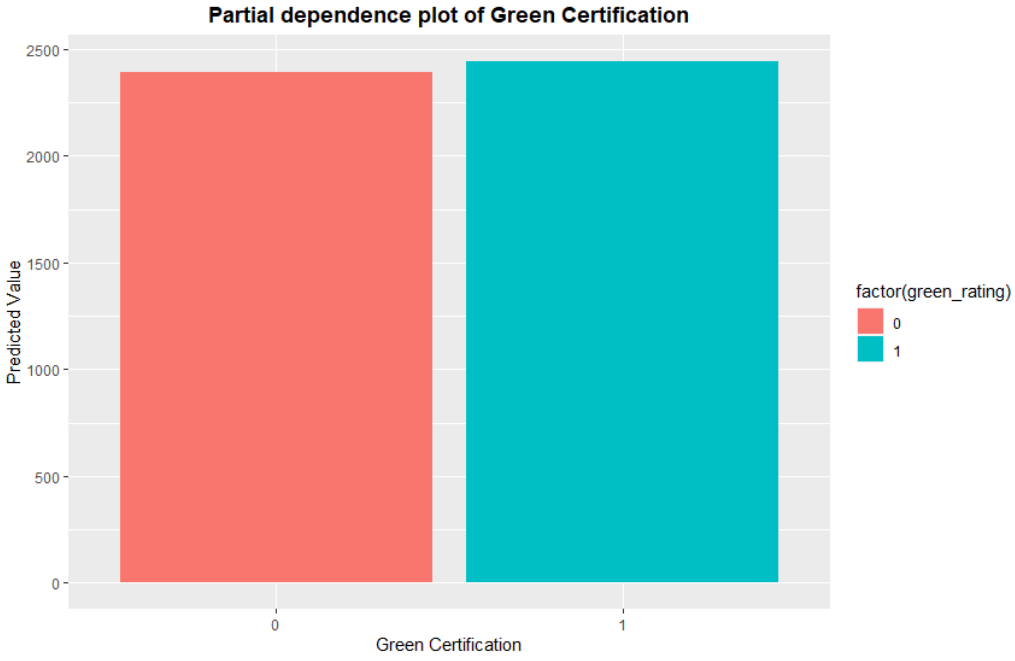
After comparing the out-of-sample RMSE of the three models, we found that the random forest model had the lowest value of 609.07. Therefore, we decided to use the random forest model as our best predictive model.

Plots

The variable importance plot for our random forest model reveals that the size, market rent, and age are the most important factors in predicting the revenue per square foot. Surprisingly, our green rating variable has the lowest importance compared to all other parameters.



The analysis suggests that a green certification may only lead to a modest increase in yearly rent revenue (per square foot) by around fifty dollars.



Wrap up

Based on our testing of three models, we chose to utilize the random forest model. Our predictive modeling showed that a green certification may not result in a significant increase in revenue. The partial dependence plot indicates that a green certification is estimated to only increase yearly rent revenue (per square foot) by approximately fifty dollars.

Q4. California Housing

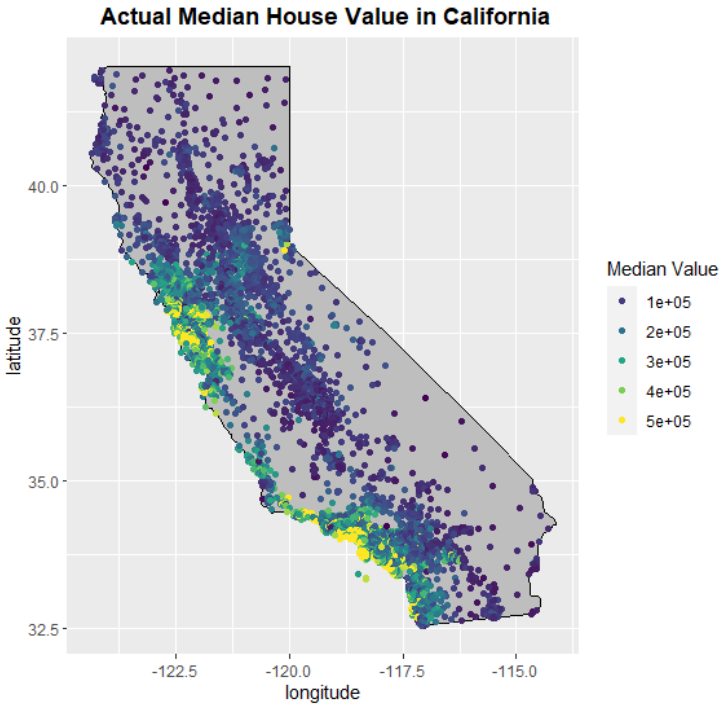
Modeling

In this question, we have compared CART, random forest, and gbm models to determine the best predictive model. After tuning the gbm model, we obtained an out-of-sample RMSE that is smaller than that of the random forest model. Therefore, we will use the gbm model as the best predictive model in this case.

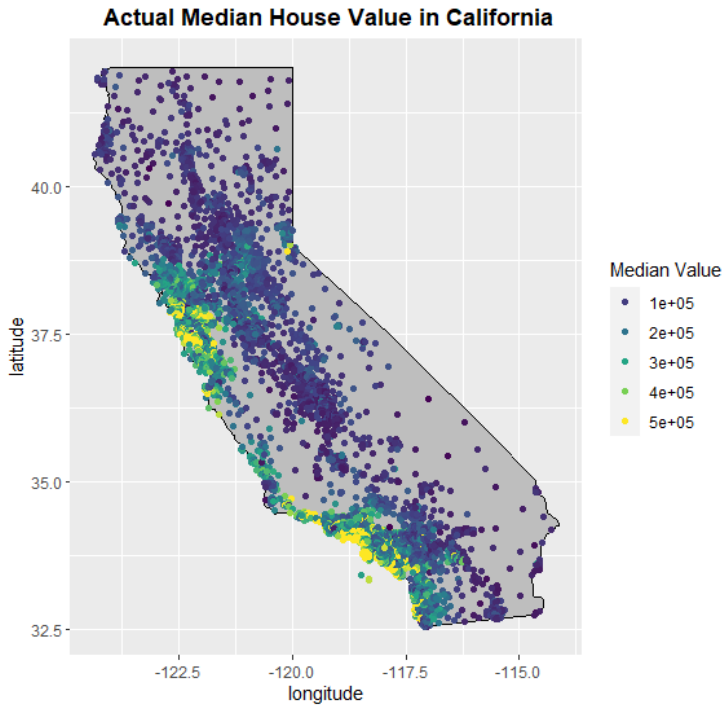
We have utilized the gbm model to create a plot of predictions and a plot of the model's residuals.

Plots

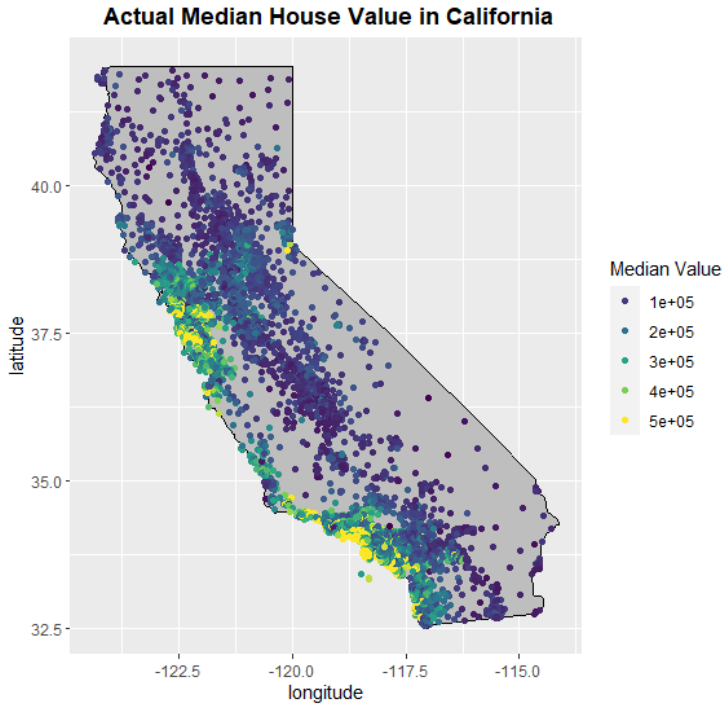
To generate the three plots, we initially set up the base plot of California by utilizing the data from the maps package.



The plotted data represents the median house values in California, with brighter colors indicating higher house values. The plot shows that areas such as Los Angeles and the San Francisco Bay Area, including coastal suburbs, have the most concentrated collection of homes with high values.



The plotted data represents the predicted median house values from our model. Upon comparison with the actual data, it appears that our model does a good job at predicting house values.



The plot illustrates the magnitude of the residuals between the actual and predicted values. The majority of the residuals appear to be relatively small, indicating that the model has a good overall fit.