# Data Mining HW 3

Patrick Chase

4/7/2021

## 1.

That approach would not account for lots of things, such as the differing styles of policing in cities, local policy, having a baseline comparison or counterfactual. In short, it may point at some relationship existing but the specific question of "increased police presence causes x to happen" can't really be based off of that analysis. Not only that, the question of generalizability is always an important one to ask. Maybe the impacts of police presence are heterogenous.

## 2.

Due to a policy decisions to increase police presence around Washington, D.C. because of higher threats of terrorist activity and not because of increased crimes, an opportunity for the researchers to conduct a natural experiment presented itself. The researchers compared crime rates on days with low terrorist threat levels to crime rates on days with high terrorist threat levels, controlling for day of the week and metro ridership. Table 2 presents these two simple models where column 1 regresses total daily crimes on alert levels controling for day of week and column two adds a measure of ridership to the the model. They found that on high alert days there were roughly 7 less crimes committed compared low alert days.

## 3.

They chose to control for Metro ridership because they believed it was possible that tourists avoided D.C. on days when it was publicized that the risk level was increased on a given day. Their hypothesis was that less tourists could lead to less crimes being commited. As a result, the researchers tried to control for this and compare days with similar threat levels AND similar ridership.

## 4.

The model is estimating the differing impacts within districts accounting for district level fixed effects. It seems that a disproportionate amount of the decrease in criminal activity occurs in District 1.

## 5. Green building model

### Overview

The goal of this analysis is to estimate the return to investing in green certification. Specifically, we want to estimate the average change in rental per square foot given green certification (LEED or EnergyStar).

## Data and Models

The data we will be using is a collection of 7894 commercial rental properties. In this dataset 685 rentals, or approximately 9% of the properties, are green certified. For this analysis we will only consider if a building has any green certification and not compare between LEED or EnergyStar. Observations with missing values have been omitted.

Lets begin with some linear regressions to get a feel of the relevant relationships. Model 1 regresses rent per square foot on green certification. Model 2 regresses rent per square foot on all variables, excluding LEED, EnergyStar, Rent, and leasing_rate. Rent and leasing_rate are being excluded because they are multiplied together to create our dependent variable.

Models 3 and 4 take a slightly more complex approach. Model 3 uses stepwise selection on Model 2, to arrive a model that relies on less variables. Model 4 uses a Random Forest model to predict the rent per square foot.

### Model 3

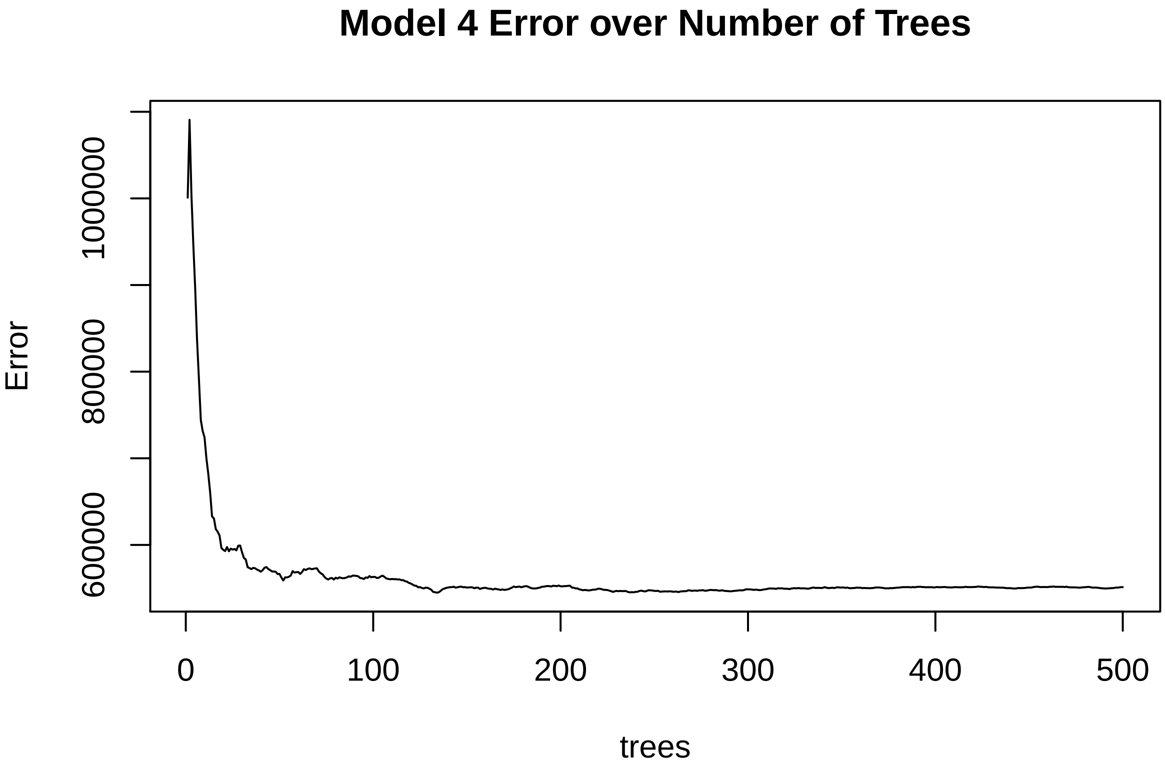
$revsqft = CSPropertyID + cluster + size + age + classa + classb + greenrating + net + amenities + hdtotal07 + ElectricityCosts + CityMarketRent$

## Results

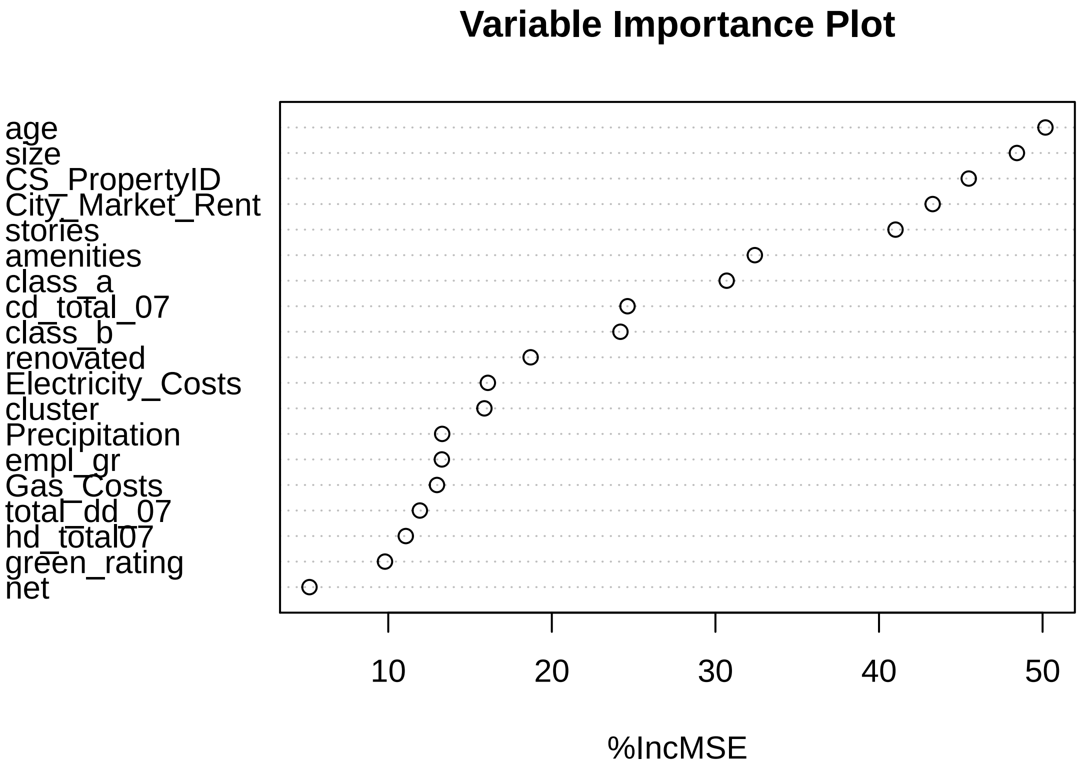
### RMSE Table

```
## # A tibble: 1 x 4
##   `Model 1` `Model 2` `Model 3` `Model 4`
##       <dbl>     <dbl>     <dbl>     <dbl>
## 1     1457.     1002.     1001.      697.
```

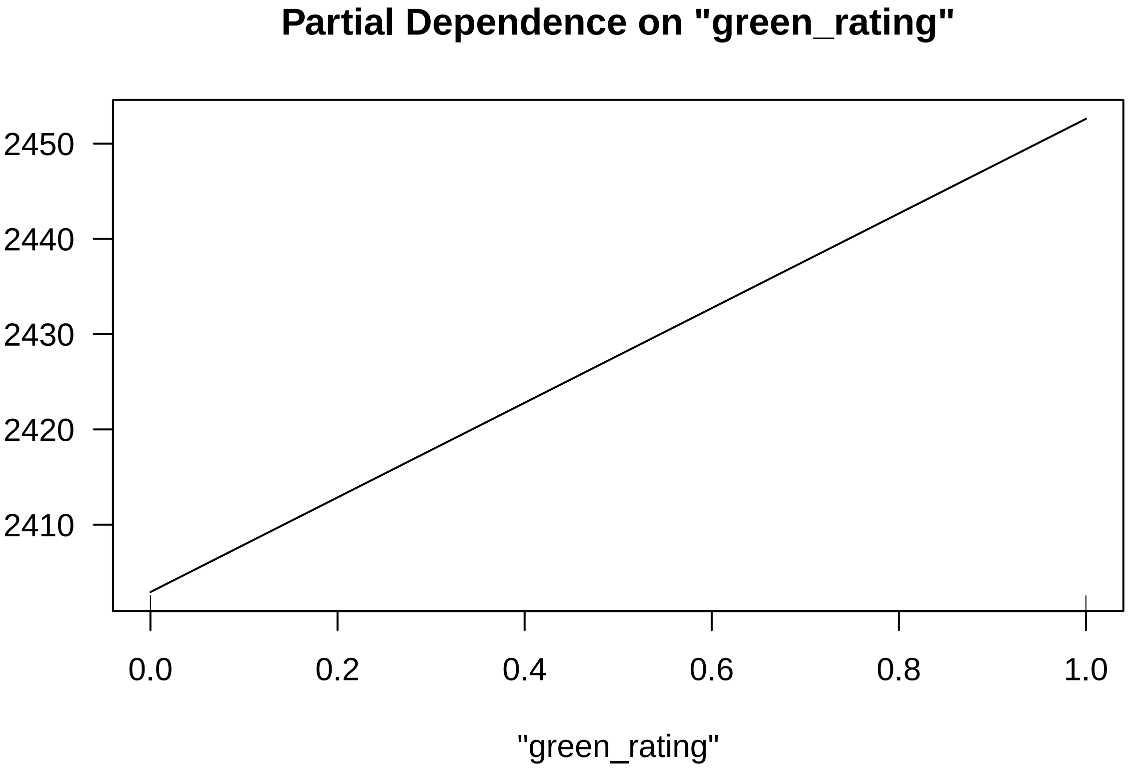
The above table shows that Model 4 has the best performance of our candidate models. As a result, we will rely on it's prediction to estimate the impact of green ratings on rent per square foot.



This plot is showing that error is minimized at approximately 400 trees.



The variable importance plot gives an indication of how useful each variable is for prediction. Interestingly enough, green_rating does not appear to be highly impactful on revenue per square foot in this dataset. That said the partial importance plot shows the average increase in revenue given green certification. On average, buildings with a green certification generate $60 per square foot more in rent then those without a certification.



### Conclusion

On average, buildings with a green certification generate $60 per square foot more in rent then those without a certification. That said, green_rating does not appear to be as impactful as other variables in our model. The age and size of the property seem to have a higher predicting power for the revenue per square foot.

## 6. California Housing Model

### Overview

Below is a predictive model for the median house value for the state of California utilizing a random forest.

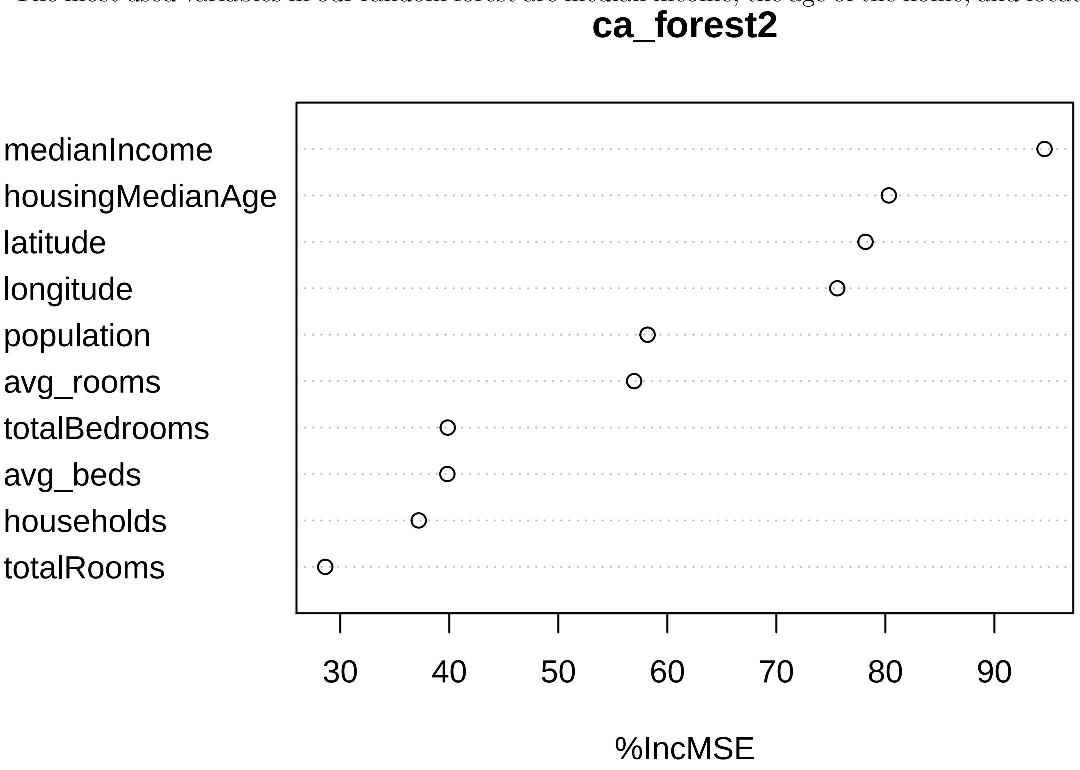
### Data and Model

Our data is census tract level with over 20000 observations. I began with three simple models using linear regression, stepwise selection, and random forest. Out of of those three, the random forest model performed the best. I then decided to run another random forest model, but only using the variables identified through stepwise selection. This was my best performing model and my choice for moving forward.

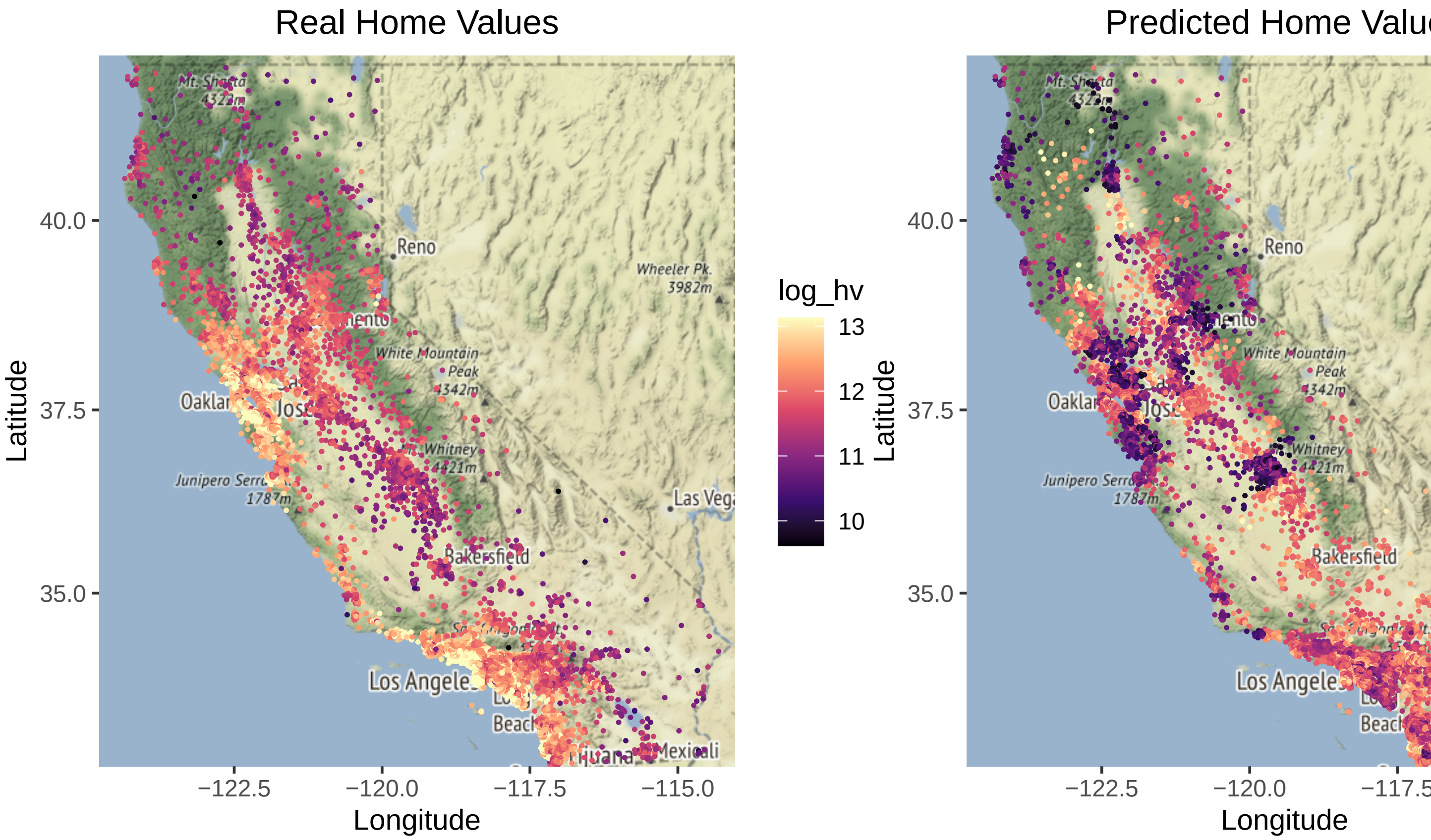
```
## # A tibble: 1 x 4
##   `Random Forest2` `Random Forest` `Regression Model` `Stepwise Selection`
##              <dbl>           <dbl>              <dbl>                <dbl>
## 1            0.248           0.247              0.342                0.342
```

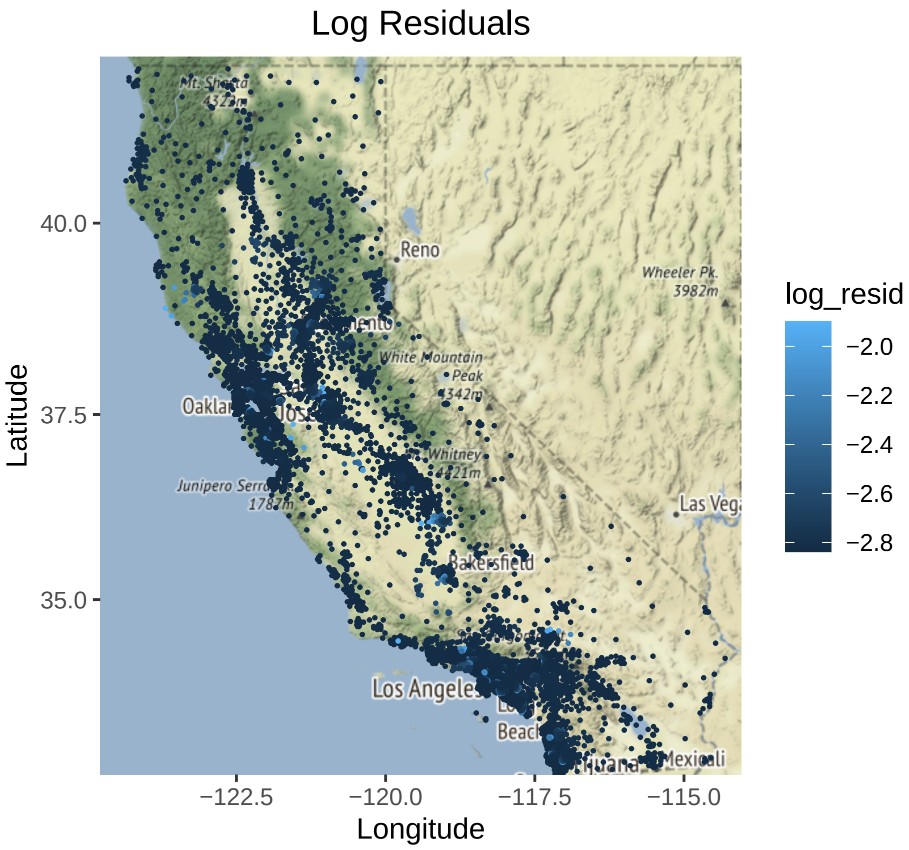
### Results

This data bears out the conventional wisdom about a large proportion of the factors that largely impact local home value. The most used variables in our random forest are median income, the age of the home, and location (lat, long).



The locational impacts on home value can be seen on the map titled “Real Home Values”. As you move east to west towards the pacific ocean, home values in the state California tend to increase. This makes sense given that a high income earners are concentrated in coastal cities and the desirability of living on or near the ocean.





## Conclusion

Home values generally increase as you move towards the Pacific ocean.