Assignment 3

Maria Gilbert

4/7/2021

What causes what?

1. If you were to pull data from a few cities and run a simple regression of "crime" on "police", you will probably find that cities with more police also have high crime rates. However, to then come to the conclusion that having more police leads to a higher crime rate would be ignoring how cities that have a crime rate to begin with will often hire more police and/or increase funding to the police, with the goal being to eventually reduce the crime rate.

I was living in Minneapolis during last summer when George Floyd was murdered and then anything and everything about the way the city funded and managed the police force was being questioned (for good reason). Something that was interesting to learn was that although a lot of people agreed that having so many police officers and such high funding was unnecessary and did more harm than good, the city government had already made an agreement with the police union to increase funding and increase the number of officers annually for a number of years. This agreement was part of the city's contract with the police union, so it would be extremely difficult if not possible to change.

I bring this example up because although people's ideas of what causes and prevents crime are changing, many cities are probably locked into agreements that will cause increases to police funding and increases to the number of officers for years to come, even despite evidence that this will not actually decrease crime.

2. The researchers were able to isolate the effect of high levels of police on crime rates by finding examples of tourist attractions in Washington D.C. where there are a lot of police there for reasons other than crime. Rather than the reason for police being in these particular areas being crime, it was the possibility of terrorism, so the police were there to protect against possible terrorism. They ended up finding that having more police presence does lead to less crime.

3. The researchers had to control for public transportation ridership levels to account for the possibility of the reason for less crime being the presence of fewer potential crime victims. They had to make sure that the days when increased police were out and about, and crime rates were lower, were not the same days as when hardly anyone was riding the public transportation around that part of the city.

4. It looks to me like the researchers did a regression between an indicator variable that equaled one when it was a “high alert” day with data in the area with all the touristy stuff in D.C., and zero otherwise, another indicator variable that equaled one when it was a “high alert” day with data in all other areas of D.C., and zero otherwise, and the natural log of midday public transportation ridership.

Predictive model building: green certification

1. Introduction

Commercial real estate is a huge business, supporting millions of jobs and contributing over $1 trillion to the U.S. GDP annually. It is very important for real estate investors and developers to be able to estimate how much money they can earn leasing out space in the buildings they have invested in. Additionally, while environmental certifications such as LEED and Energystar can show that a building is relatively environmentally friendly, often the construction of buildings that meet these standards and renovations of older buildings to make them reach these standards are expensive. The goal of my analysis will be to quantify how much more rent per square foot per year can be expected for buildings with a LEED or Energystar certification, so that investors can decide if qualifying for and obtaining such a certification will pay off.

2. Data and model

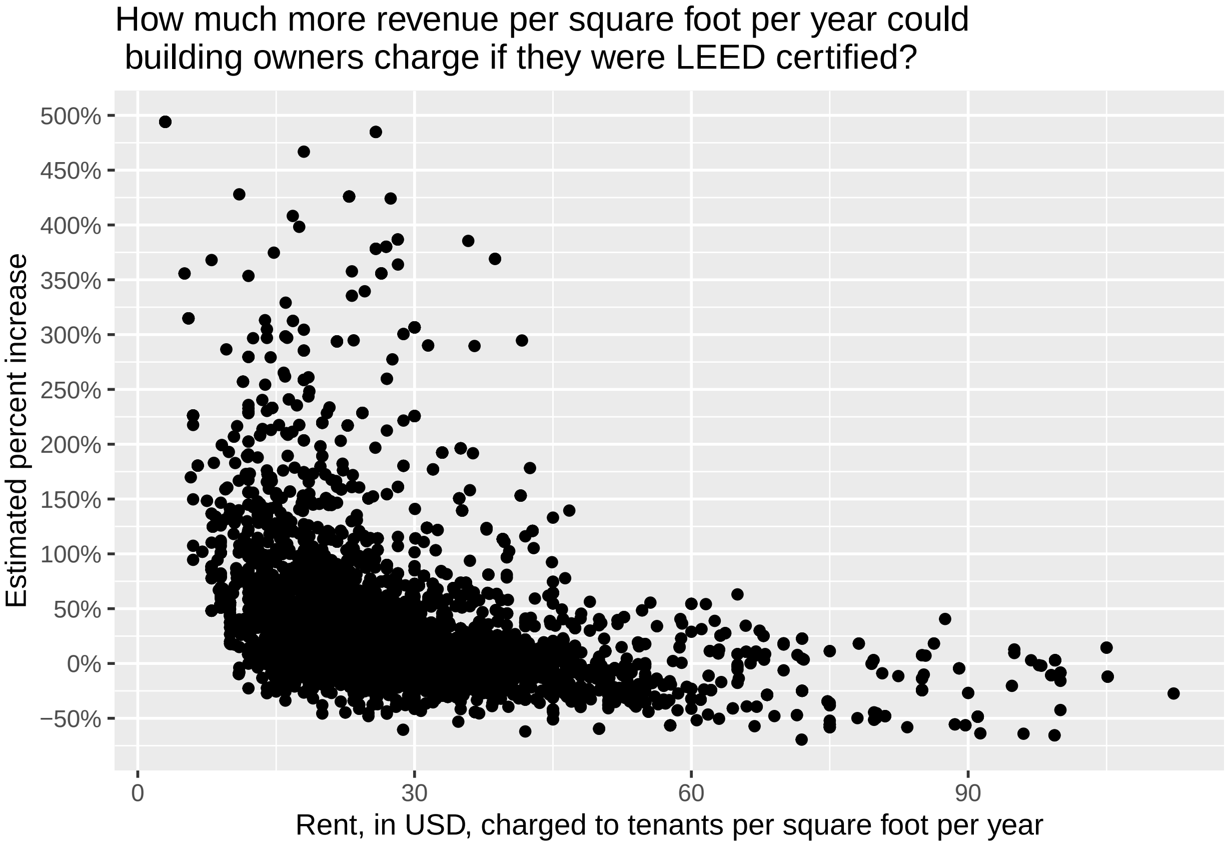
In order to model the effect of green building certifications on rent per square foot per calendar year, I fit a boosted model using a square footage, annual employment growth rate in the region, number of stories, age of the building, whether the building is renovated or not, building quality class (A being the best, C being the worst), whether rent for the building tenants includes utilities or not, the presence of amenities in the building, the number of heating and cooling days where the building needed to be heated or cooled inside, annual precipitation in the region, gas costs, and electricity costs, to predict revenue per square foot per year (the product of rent and leasing rate). The data I used was from a dataset of 7,894 commercial real estate properties across the United States. 54 of them are LEED certified, 638 of them are Energystar certified, 7 are both (these are also included in my figures for how many have either separate certification), and 7209 have neither LEED nor Energystar certification.

The other variables in the data include a unique identifier, a cluster identifier (each cluster includes at least one building with a green certification and at least building without one within a 1/4 mile radius of each other), square footage, annual employment growth rate in the region, rent amount per square foot each year, leasing rate, stories, age of the building, whether the building is renovated or not, building quality class (A being the best, C being the worst), whether rent for the building tenants includes utilities or not, the presence of amenities in the building, the number of heating, cooling, and total days where the building needed to be heated or cooled inside, annual precipitation in the region, gas costs, electricity costs, and market rent in the local area.

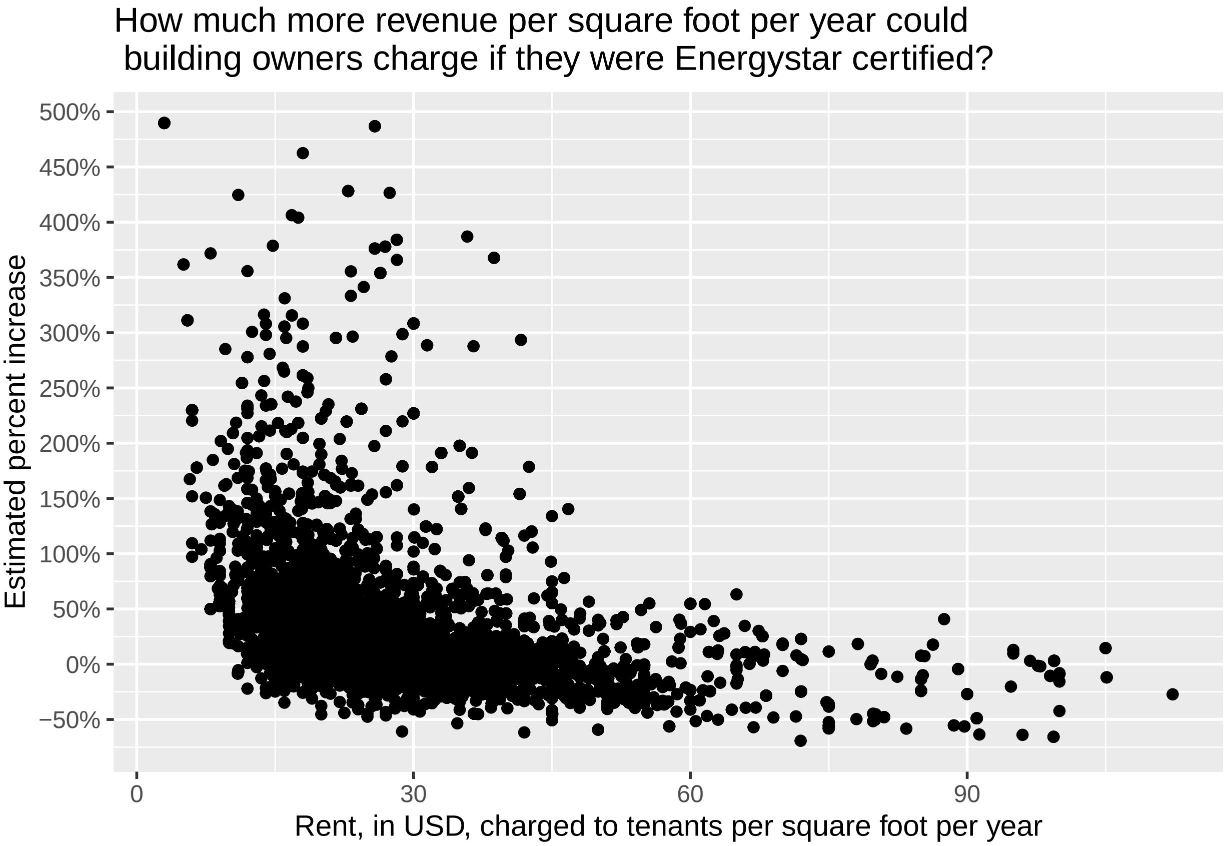
I also tested a step model and a random forest model and found that the boosted model had the lowest root mean squared error, though it was no more than 20% lower than either of the other models.

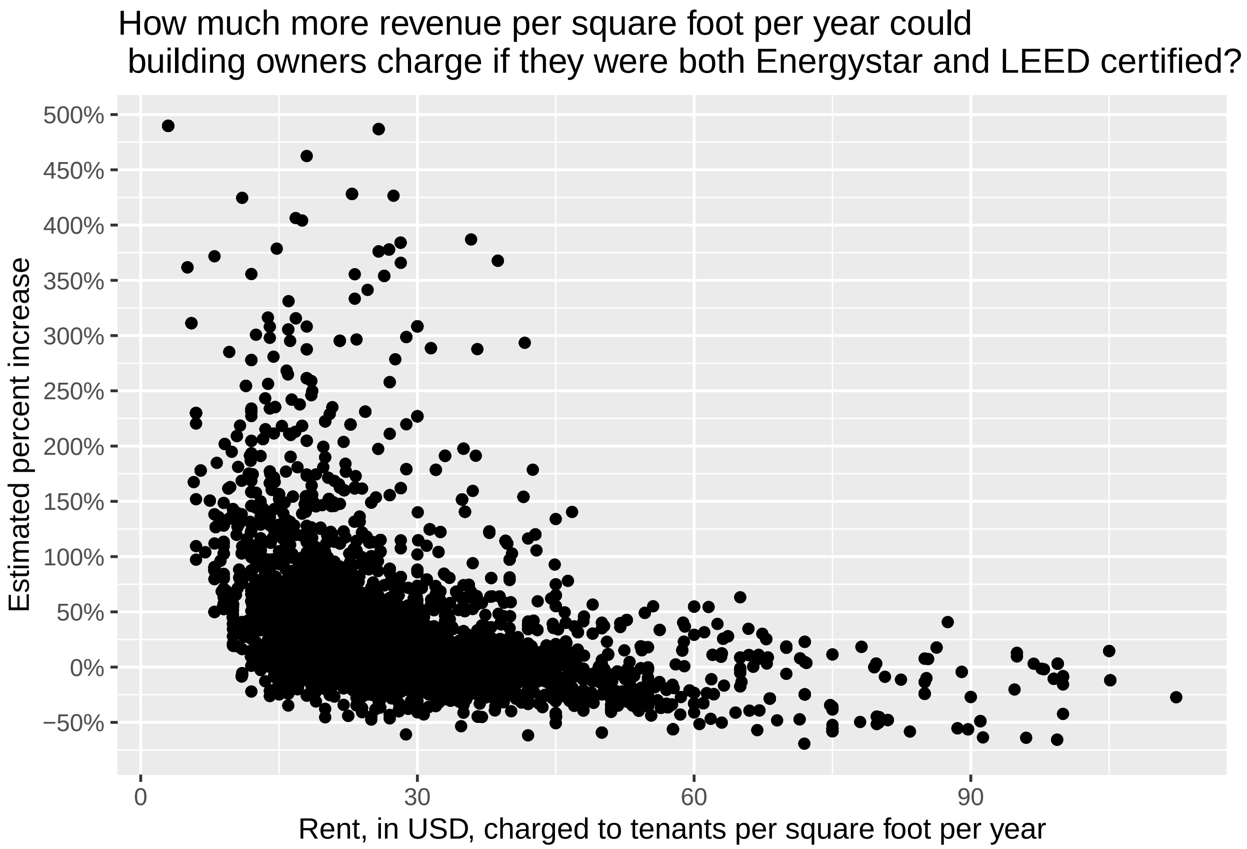
3. Results

In order to show how having a LEED, Energystar, or both certifications affects estimated revenue per square foot per year, I decided to create a new dataset of just the buildings with no certification, use my model to estimate what the revenue per square foot per year would be if they had a certification, and then find the percent difference between that and the actual revenue per square foot per year.



For all three graphs, I excluded outliers where either the rent per square foot per year was greater than $125, since there were very few data points beyond that, as well as those where the model would result in a greater than 500% increase to the actual revenue per square foot per year.





For all three graphs, I excluded outliers where either the rent per square foot per year was greater than $125, since there were very few data points beyond that, as well as those where the model would result in a greater than 500% increase to the actual revenue per square foot per year.

Next, I calculated the median percent increase that adding either or both certification would have on the revenue per square foot per year in my model. For LEED only,

```
## [1] "20%"
```

For Energystar only,

```
## [1] "20%"
```

And for both certifications,

```
## [1] "20%"
```

As you can see, adding an Energystar certification has about the same impact as adding a LEED certification, so it is a case of decreasing returns as you add more certifications. Of course, this is disregarding the potential increased construction and/or renovation costs to needed qualify for either or both certification.

In order to demonstrate that these percent increases are not due to a bias on the part of my model, I also calculated the median percent difference between my model's estimated values and the actual values of revenue per square foot per year for just the green buildings in the data set.

```
## [1] "-1%"
```

And finally, I also calculated this median percent difference between my model's estimated values without adding either a LEED or an Energystar certification to the non-green buildings in the data set.

```
## [1] "3%"
```

4. Conclusion

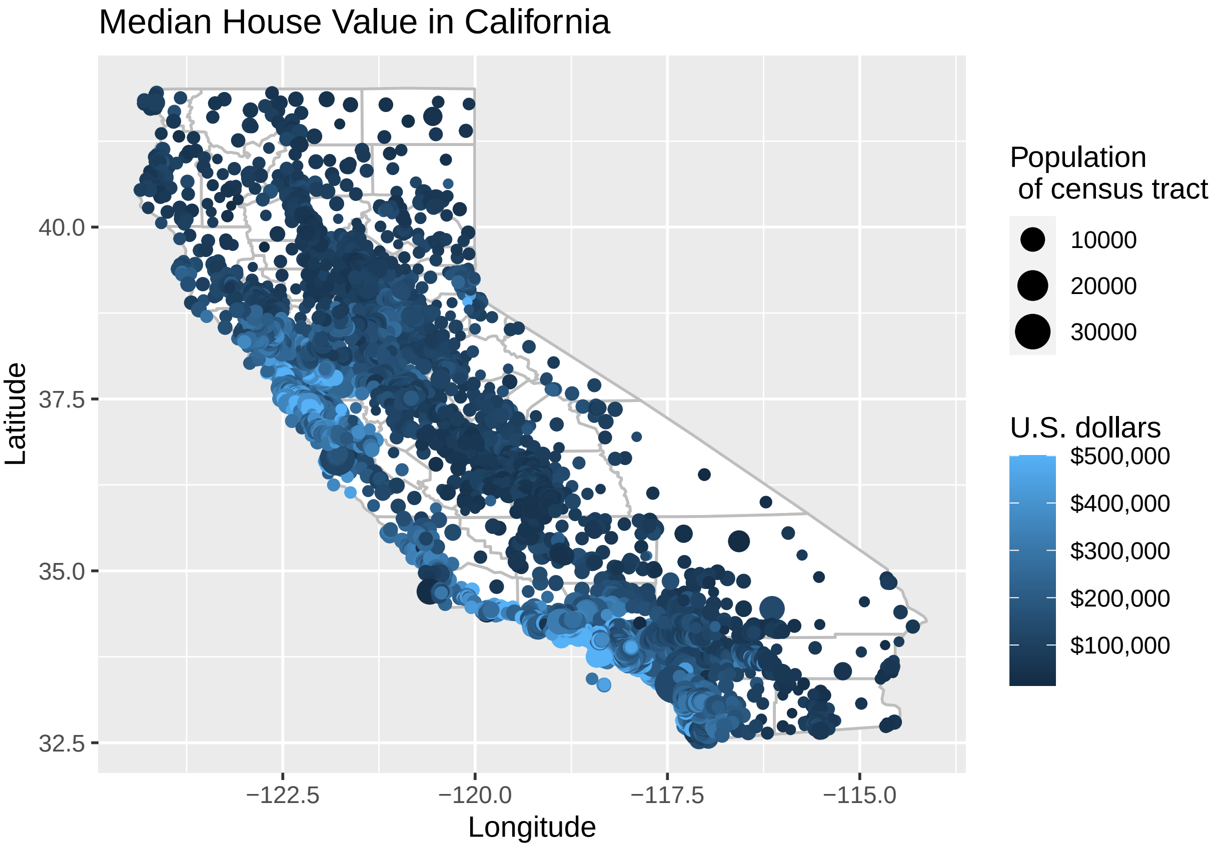
My model shows that having a LEED or Energystar certification (or both), with all other variables held constant, can lead to modestly higher revenue per square foot per year for commercial real estate investors. It is important for investors to know this so that they will be more likely to decide to build or renovate their properties to meet the proper green specifications, which will be good for the environment.

Predictive model building: California housing

1. Introduction

With housing being increasingly expensive, people are having to decide where they can afford to move to or live based on how much it is going to cost to buy or rent a house there. Those who are more fortunate are lucky enough to be deciding where they might want to invest in real estate to earn money on that property. Both groups need to know how much they can expect a house to be worth in a geographical area. The goal of my analysis will be to make a model predicting the median house value in an area based on census tract data.

Here is a graphic showing the actual median market house values in the data, sized by population.



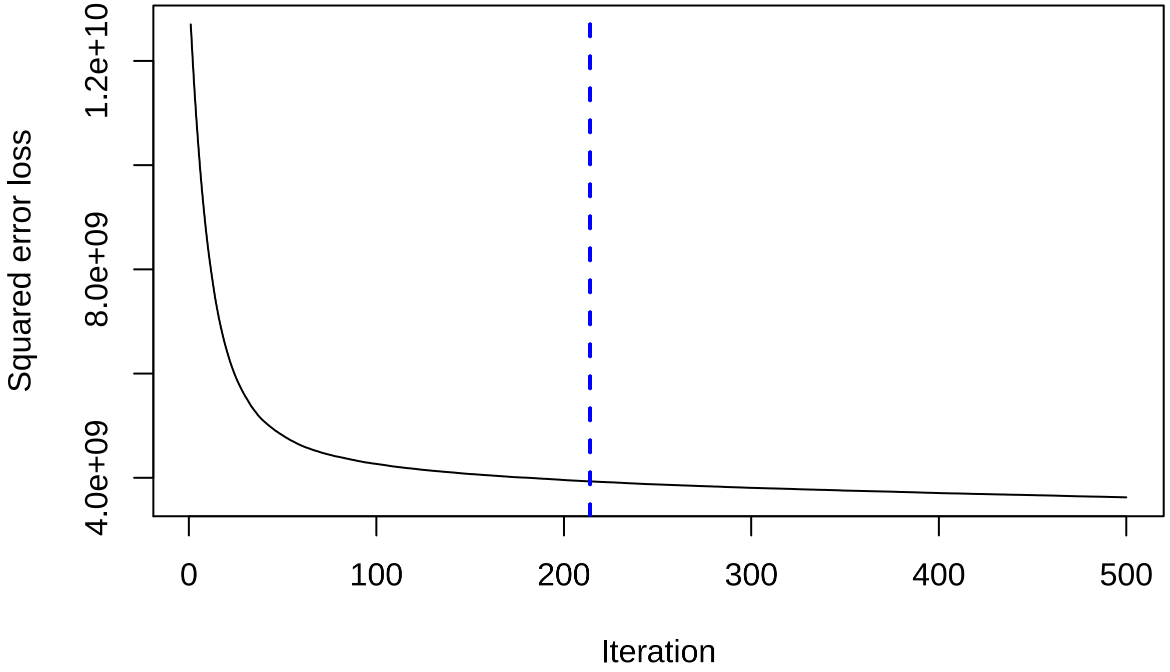
2. Data and model

In order to model median house value, I fit a generalized boosted regression model with Gaussian error on the median age of all houses within the tract, the total human population in the tract, number of households within the tract, total number of rooms and bedrooms within all homes in the tract, median annual income for the households within the tract, the average number of people per household, the average number of rooms per household, the average number of bedrooms per household, the average number of people per bedroom, and the median per capita income, to predict median house value within a census tract. The data I used was from California census tracts and also included variables for longitudinal and latitudinal coordinates of the tract, as well as the actual median market house value in the tract.

In order to find the model with the best performance, I minimized out-of-sample root mean square error (RMSE). I tested a random forest model, linear model, and log-linear model, all using the same variables as the final boosted model. The boosted model and random forest model performed about equally, and both significantly outperformed the linear and log-linear models. The RMSE of the random forest model was:

```
## [1] 63832.49
```

I found that the best number of trees for my boosted model was around 220 or more.



```
## [1] 214
## attr(,"smoother")
## Call:
## loess(formula = object$oobag.improve ~ x, enp.target = min(max(4,
##     length(x)/10), 50))
##
## Number of Observations: 500
## Equivalent Number of Parameters: 39.85
## Residual Standard Error: 3273000
```

The RMSE of my final boosted model was:

```
## [1] 64299.08
```

The RMSE of my linear model was:

```
## [1] 72969.56
```

And the RMSE of my log-linear model was:

```
## [1] 72969.56
```

The relative performance improvement of the final boosted model over the random forest, linear, and log-linear models were, respectively:

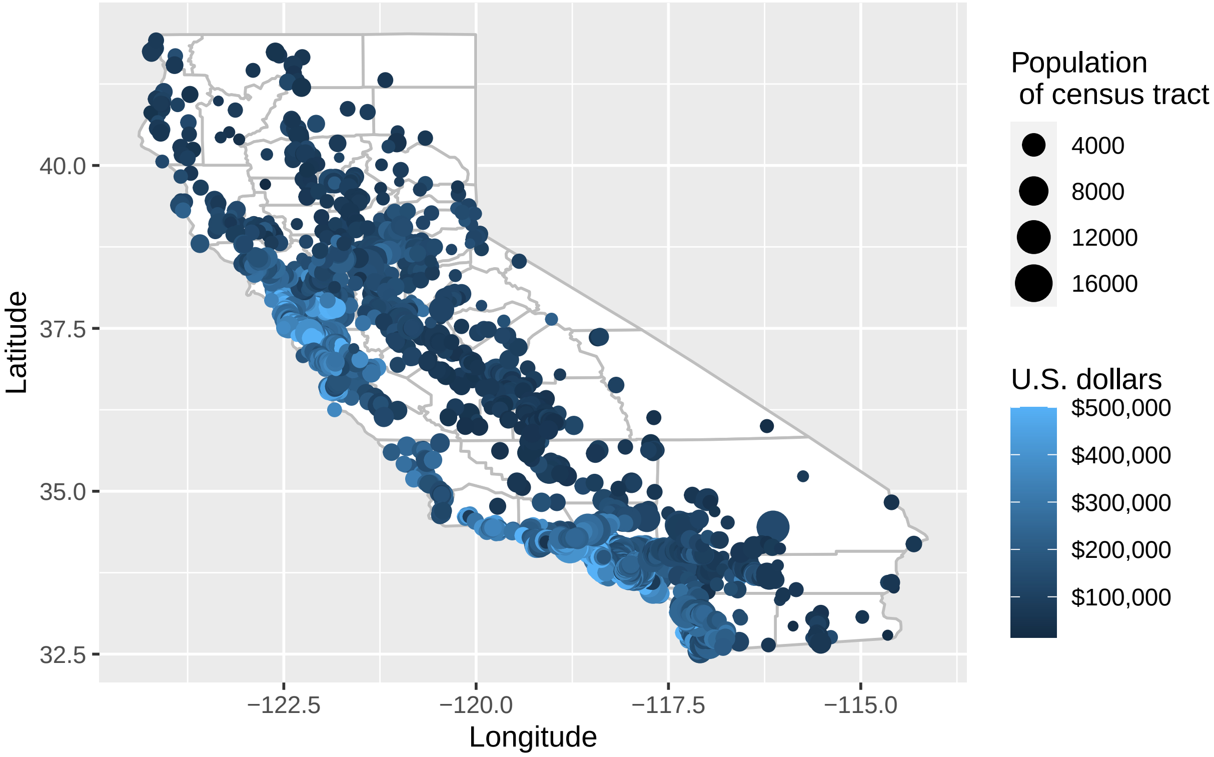
```
## [1] -0.007256572
## [1] 0.1348462
## [1] 0.1348462
```

As you can see, the performances of the boosted model and random forest model are very close to equal, while they outperform the linear and log-linear by a significant amount.

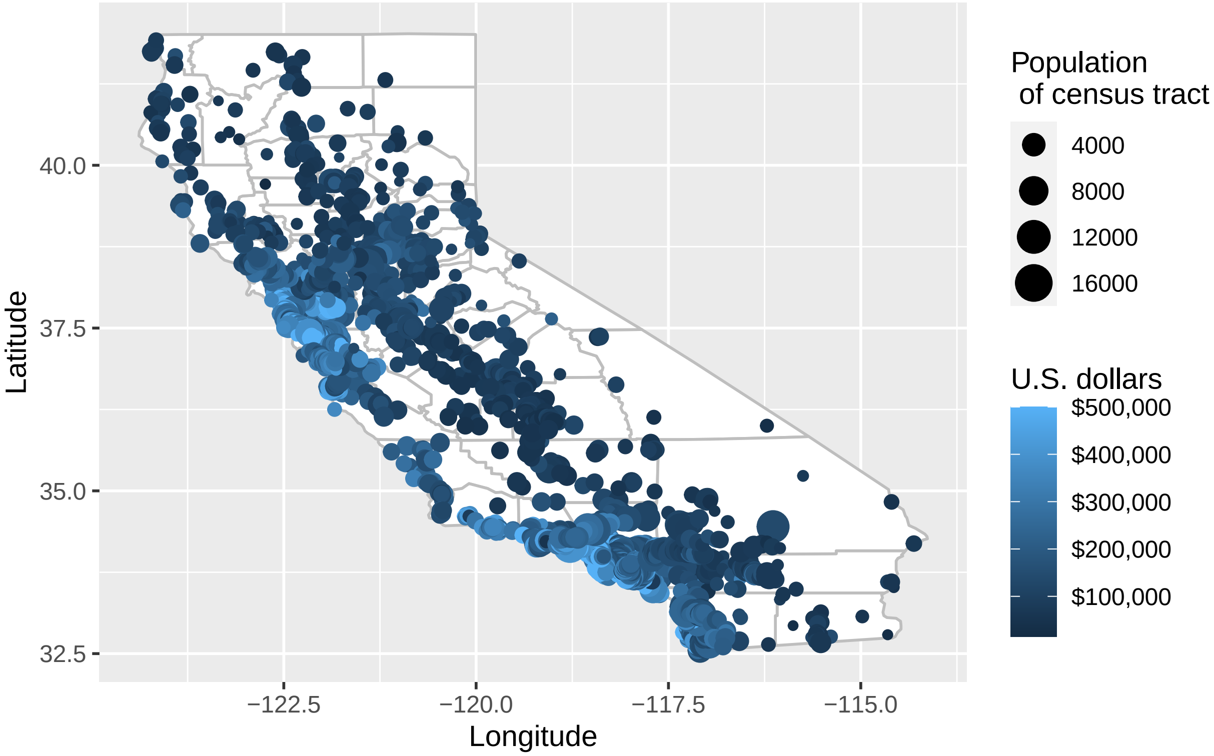
3. Results

Here I will show the same graphic as before, but comparing the actual house value within the testing set with that calculated by the model.

Median House Value in randomly selected
California census tracts California



Boosted model of Median House Value in randomly selected
California census tracts California



4. Conclusion

Using just a limited number of variables available on census tract data, it is possible to quite accurately estimate the median house value within an area using a boosted model or a random forest model. It would be very interesting to see how well the same model would perform in other states besides California.