Homework 3

Cameron Wheatley

2023-03-27

1 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?

("Crime" refers to some measure of crime rate and "Police" measures the number of cops in a city.)

****If using a simple linear regression model with the "Crime" and "Police" data,**

the population regression function will contain the issue of selection bias as endogeneity will occur.

E(xu)=0 is violated in this case and OLS becomes inconsistent due to the changes in police force

being associated with both changes in crime and the error term in the estimate.**

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.

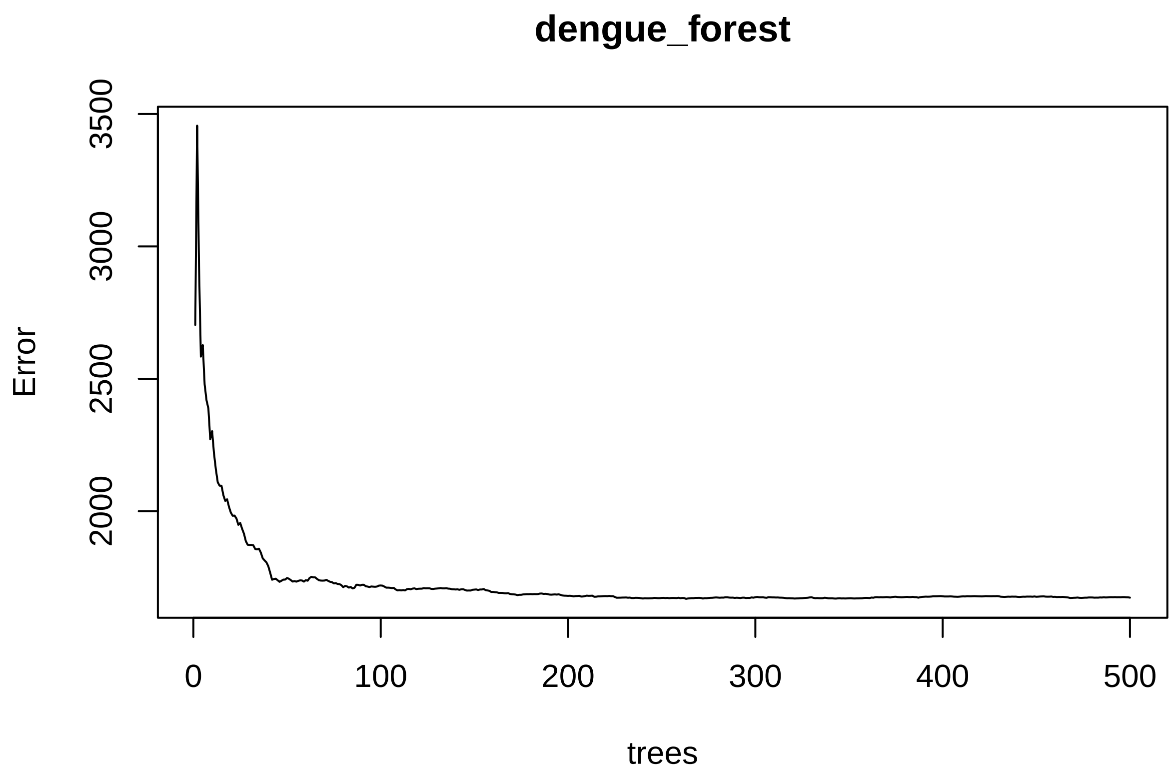
****Here, the dummy variable of "high-alert periods" gets rid of the endogeneity issue**

for police on crime as the alert level directly impacts the number of units sent to a

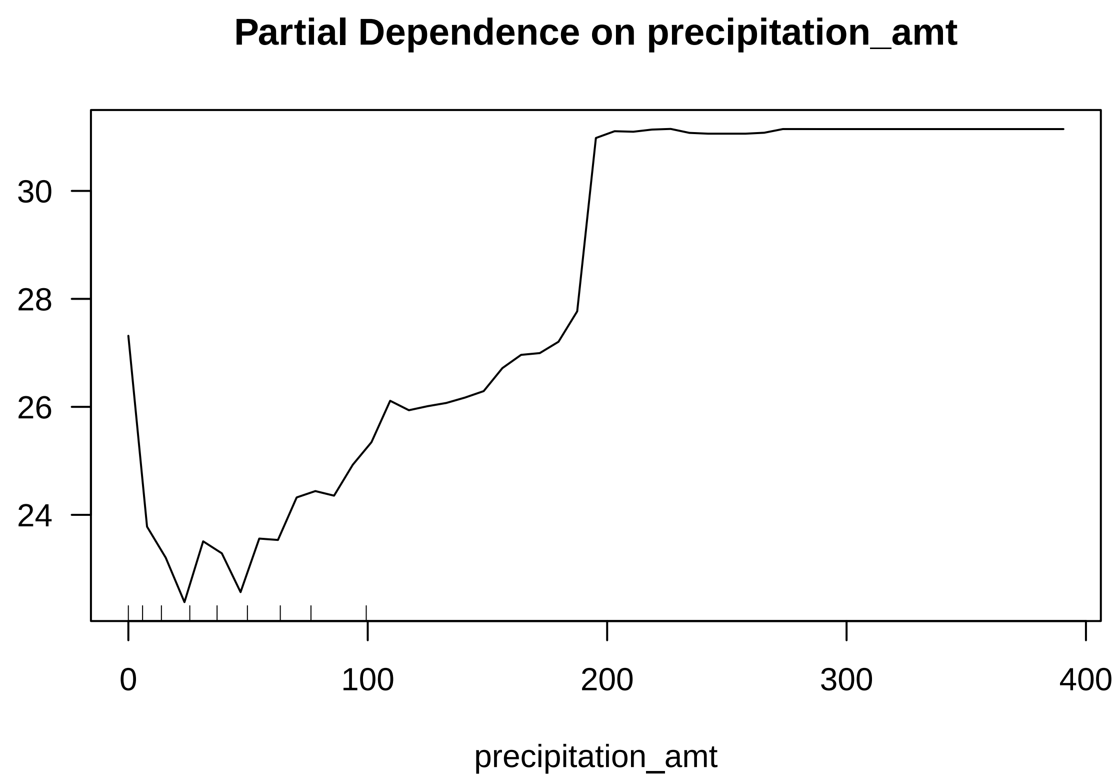
particular district. Furthermore, the authors choose the data that includes information of

repeted terror alerts which accounts for perfect first order autocorrelation (serial correlation).

```
plot(dengue_forest)
```




```
partialPlot(dengue_forest, as.data.frame(dengue_test), precipitation_amt, las=1)
```




```r
rmse(dengue_boost1, dengue_train_check)
```

```
## [1] 34.99243
```

```r
rmse(dengue_boost2, dengue_train_check)
```

```
## [1] 35.015
```

```r
rmse(dengue_boost3, dengue_train_check)
```

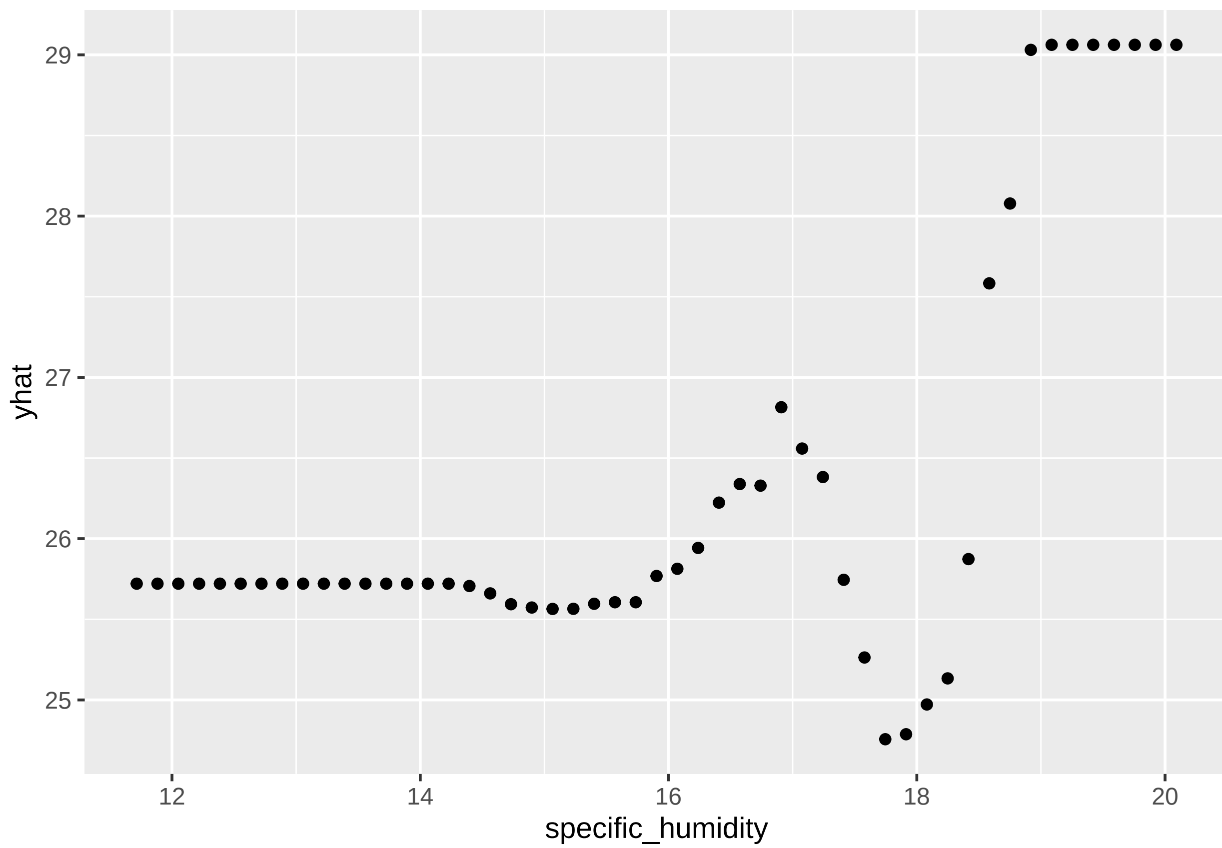
```
## [1] 32.30899
```

```r
p1 = pdp::partial(dengue_boost3, pred.var = 'specific_humidity', n.trees=1000)
p1
```

```
##    specific_humidity     yhat
## 1           11.71571 25.72102
## 2           11.88323 25.72102
## 3           12.05074 25.72102
## 4           12.21826 25.72102
## 5           12.38577 25.72102
## 6           12.55329 25.72102
## 7           12.72080 25.72102
```

```
## 8          12.88831 25.72102
## 9          13.05583 25.72102
## 10         13.22334 25.72102
## 11         13.39086 25.72102
## 12         13.55837 25.72102
## 13         13.72589 25.72102
## 14         13.89340 25.72102
## 15         14.06091 25.72102
## 16         14.22843 25.72102
## 17         14.39594 25.70640
## 18         14.56346 25.66059
## 19         14.73097 25.59360
## 20         14.89849 25.57303
## 21         15.06600 25.56434
## 22         15.23351 25.56499
## 23         15.40103 25.59610
## 24         15.56854 25.60569
## 25         15.73606 25.60569
## 26         15.90357 25.76857
## 27         16.07109 25.81311
## 28         16.23860 25.94261
## 29         16.40611 26.22346
## 30         16.57363 26.33870
## 31         16.74114 26.32881
## 32         16.90866 26.81497
## 33         17.07617 26.55885
## 34         17.24369 26.38217
## 35         17.41120 25.74497
## 36         17.57871 25.26324
## 37         17.74623 24.75641
## 38         17.91374 24.78719
## 39         18.08126 24.97183
## 40         18.24877 25.13356
## 41         18.41629 25.87337
## 42         18.58380 27.58307
## 43         18.75131 28.07815
## 44         18.91883 29.03075
## 45         19.08634 29.06218
## 46         19.25386 29.06218
## 47         19.42137 29.06218
## 48         19.58889 29.06218
## 49         19.75640 29.06218
## 50         19.92391 29.06218
## 51         20.09143 29.06218
```

```r
ggplot(p1) + geom_point(mapping=aes(x=specific_humidity, y=yhat))
```




```
modelr::rmse(pruned_dengue, dengue_test)
```

```
## [1] 39.19054
```

```
modelr::rmse(dengue_forest, dengue_test)
```

```
## [1] 36.06903
```

```
modelr::rmse(dengue_boost1, dengue_test)
```

```
## [1] 35.38469
```

The results suggest that the random forest has (slightly better than boosting) the best performance on the testing data.

3 Predictive model building: green certification

3.1 Overview

**Landlords are worried about revenue by square feet per year. Given that their leasing revenue

depends on many factors/parameters that are apart of a tenants' living environment, people

might pay more money to a landlord that has a green certification. Thus, conducting research on

a potential relationship between rent income and green certification could be worthwhile. Thus, we

will find the best best model possible that predicts revenue per square foot in order to measure the

estimated change in rental income when taking green certification into account.**

3.2 Data and research design

3.2.1 Data

**There are 7,894 data points from the raw data. When filtering the data,

"greenbuildings" now has 7,820 observations.**

3.2.2 Predictive variable and features

**Yearly revenue per square foot becomes the predictive variable which is the product of rent, leasing_rate...

holding all other covariates fixed.

The features of our model...

cluster: an identifier for the building cluster, with each cluster

```r
rmse_lm
```

```
##  result
## 1021.36
```

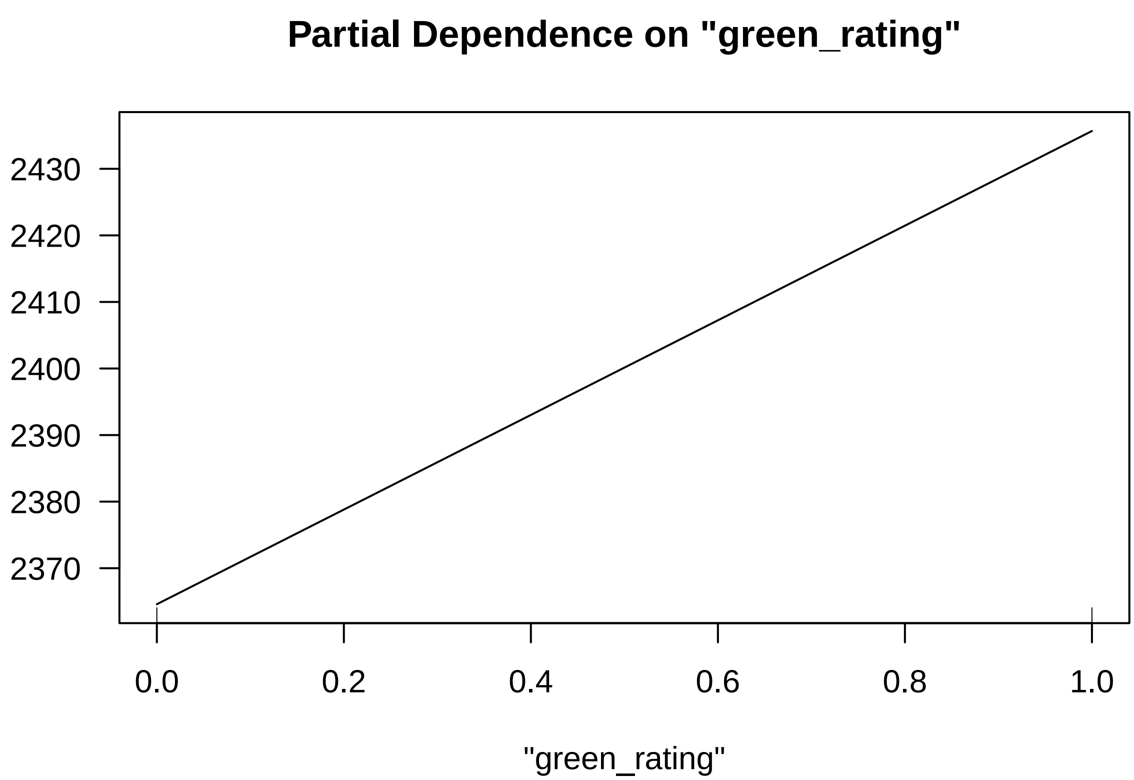
```r
rmse_forest_green
```

```
## [1] 718.01
```

```r
rmse_boost_green
```

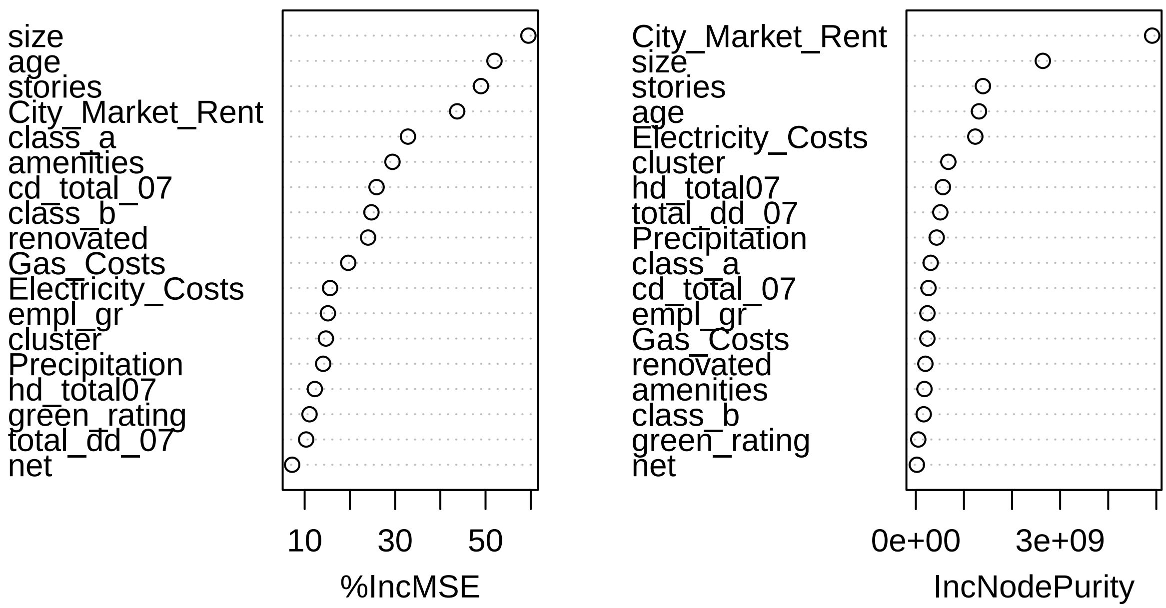
```
## [1] 919.55
```

```r
partialPlot(green_forest, as.data.frame(green_test), 'green_rating', las = 1)
```

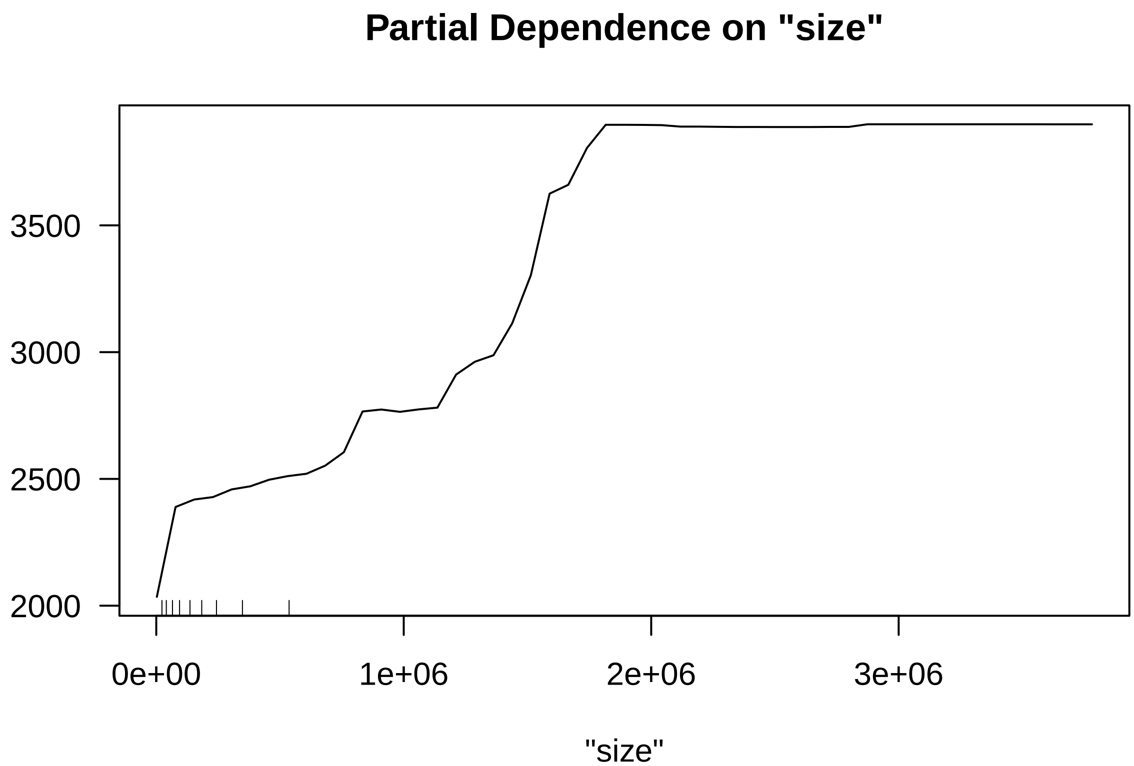



```r
varImpPlot(green_forest)
```

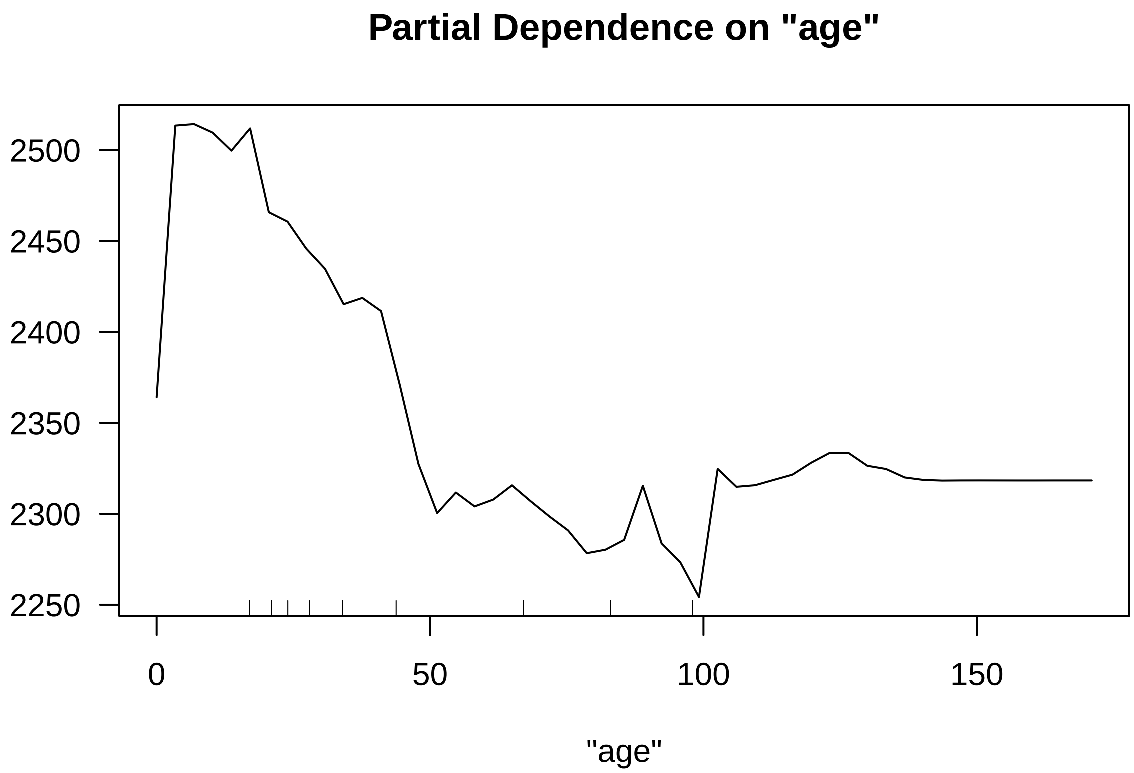
green_forest




```
partialPlot(green_forest, as.data.frame(green_test), 'size', las = 1)
```




```
partialPlot(green_forest, as.data.frame(green_test), 'age', las = 1)
```


Partial Dependence on "age"
2500
2450
2400
2350
2300
2250
0
50
100
150
"age"

When comparing the linear model, random forest, boosting models, it was the random forest model that gave the most accurate predictions. The results show that green_rating doesn't have a significant impact on the model. However, parameters such as size and age were, on the other hand, significant.

Therefore, building and having a green certification did not have an impact on rental income per square foot.

4 Predictive model building: California housing

By dividing the variables total rooms and total bedrooms by the number of households, we were able to obtain the mean of rooms and bedrooms per household in each tract. We also obtained the variable mean house size for our model. Finally, by including all variables except for total rooms and total bedrooms, we calculated the average RMSE from both a linear model and random forest to obtain the best accuracy possible.

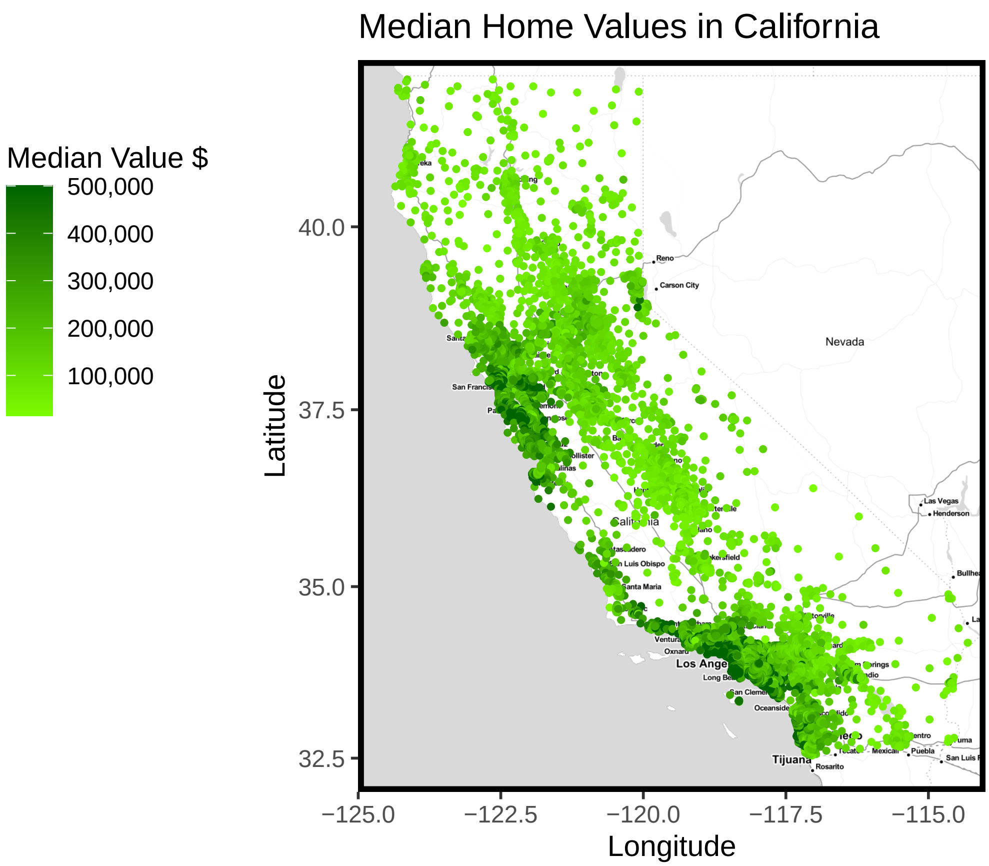
```
rmse_lm_houses
```

```
##  result
## "69291"
```

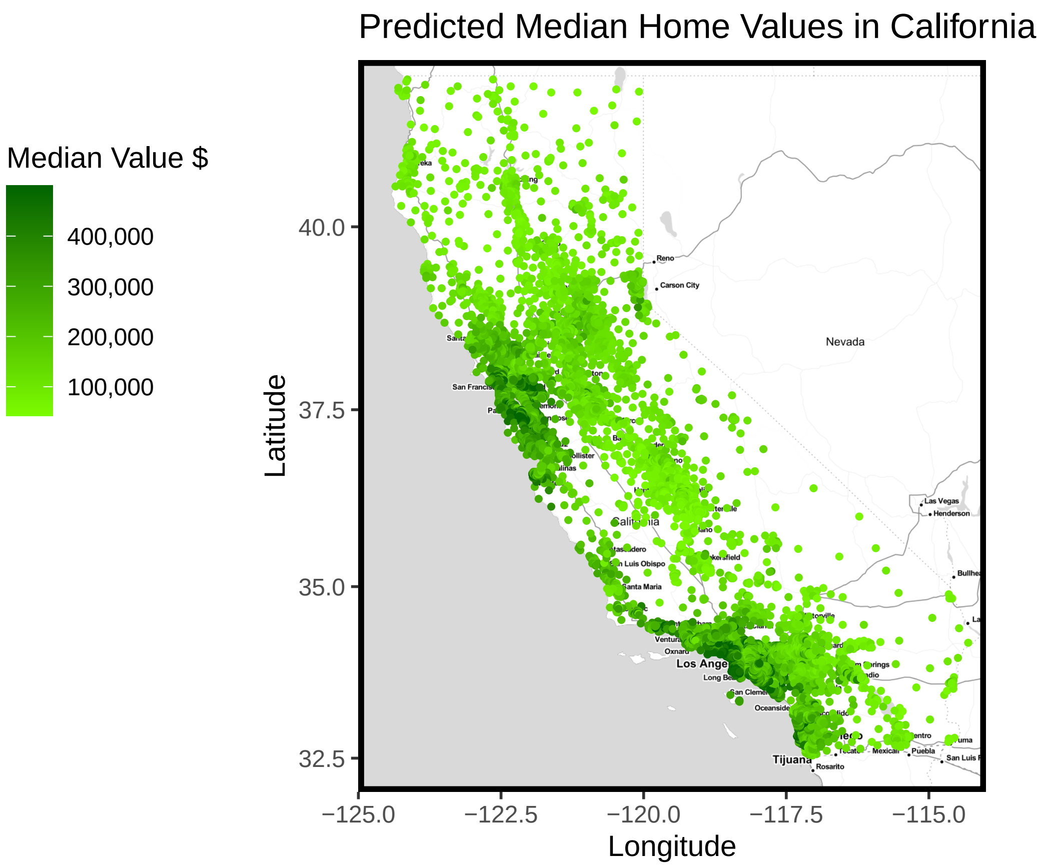
```
rmse_forest_houses
```

```
## [1] "48076.14"
```

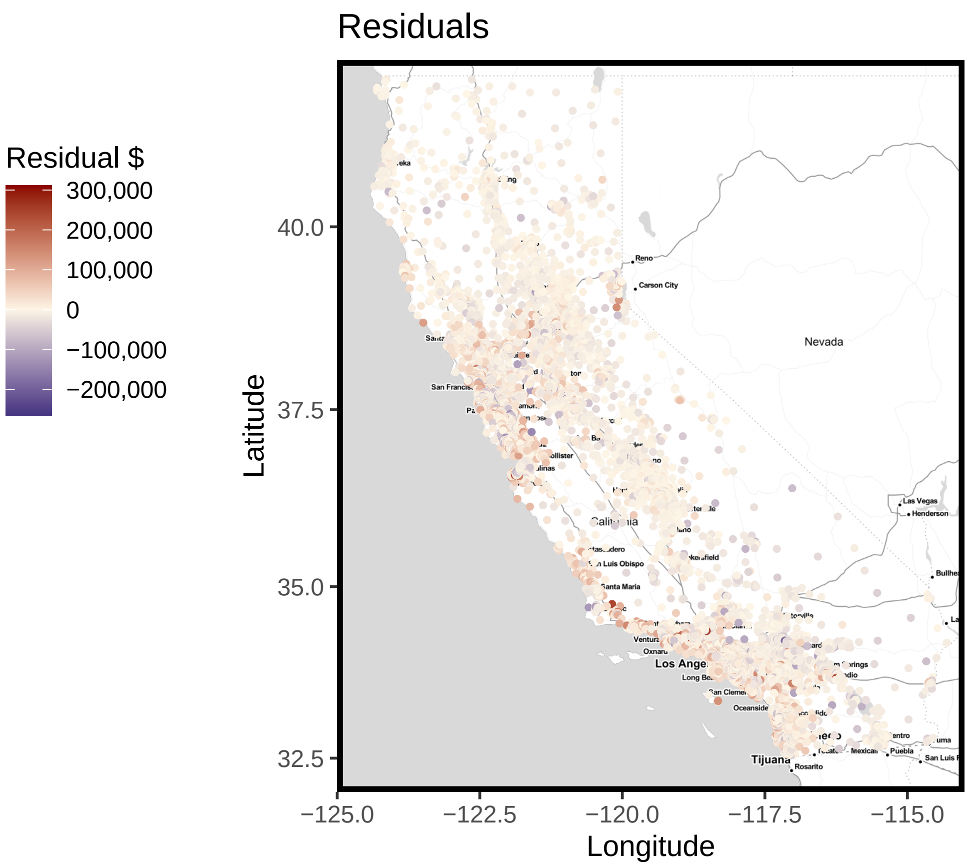
```
data_map
```


Median Home Values in California
Median Value $
500,000
400,000
300,000
200,000
100,000
Latitude
40.0
37.5
35.0
32.5
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Nevada
Reno
Carson City
Los Angeles
Tijuana
Las Vegas
Henderson

```
pred_map
```


Predicted Median Home Values in California
Median Value $
400,000
300,000
200,000
100,000
Latitude
40.0
37.5
35.0
32.5
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Nevada
Reno
Carson City
San Francisco
Los Angeles
Tijuana
Rosarito
Las Vegas
Henderson
Santa Maria
Oxnard
Ventura
Oceanside

```
resid_map
```



Residuals
Residual $
300,000
200,000
100,000
0
−100,000
−200,000
Latitude
40.0
37.5
35.0
32.5
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Reno
Carson City
Nevada
San Francisco
Las Vegas
Henderson
San Luis Obispo
Santa Maria
Oxnard
Los Ange
Oceanside
Tijuana
Rosarito