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).

The authors have seasonal dummies to measure the variables' effects based on the day of the week

(decreasing treatment window). Finally, the authors estimate a "Metro ridership" variable that accounts

for tourism and crime correlation.

Based on the results from Table 2, while high alert days decreases crime by approximately 7 crimes

per day (significant at the 5% level), increased metro ridership was associated with a small increase in

the number of crimes committed. When the authors included a logged midday metro ridership variable,

high alert levels were not being confounded with levels of tourism (high alert levels did not increase).

Therefore, change in the number of tourists on a given day does not explain a significant change in crime

level. This is due to the variable of midday metro ridership accounting for the fluctuations in crime caused

by tourism.**

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

****The authors are attempting to capture the amount of potential victims. By testing whether tourism**

is decreased on the high alert days, the amount of potential victims decreases causing less crime.**

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 model from Table 4 shows fixed effects grouped by district which demonstrates a particular**

clustered crime pattern for each district. By clustering by the day of the week, the dependent variable

of daily crime totals by district is now able to be unbiased and efficient. When a particular period

is in high alert, National Mall crime decreases by approximately 2.6 crimes daily in district 1. While

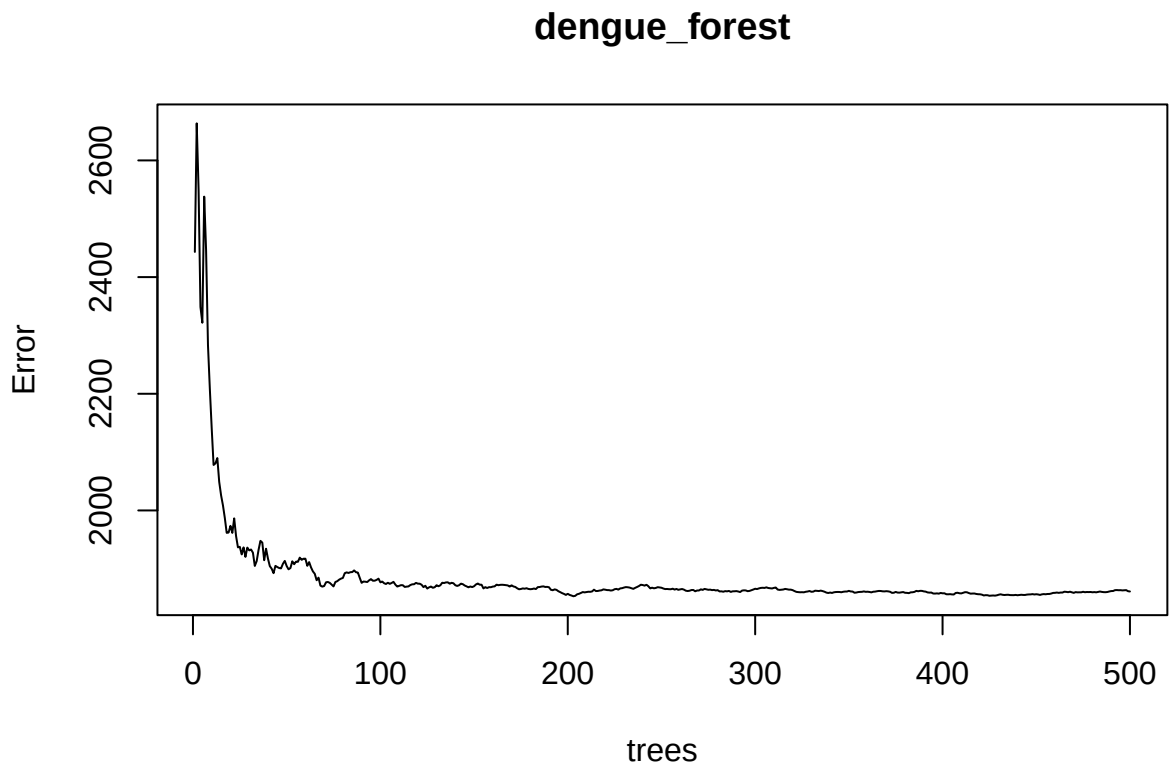
crime also decreases in other districts, the results were not as strongly significant (i.e. closer to zero).

Therefore, taking only district 1 into account, number of crimes decreases by approximately 15% on high alert

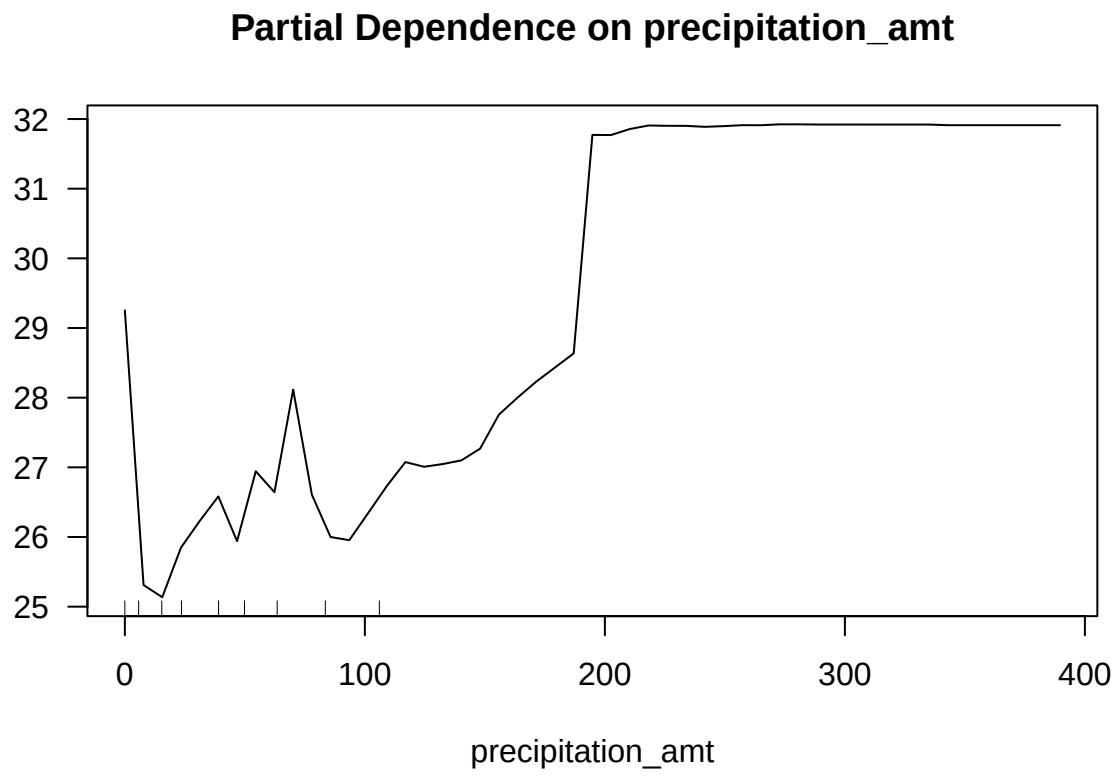
days.**

2 Tree modeling: dengue cases

```
plot(dengue_forest)
```




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




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

```
## [1] 36.70801
```

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

```
## [1] 36.67577
```

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

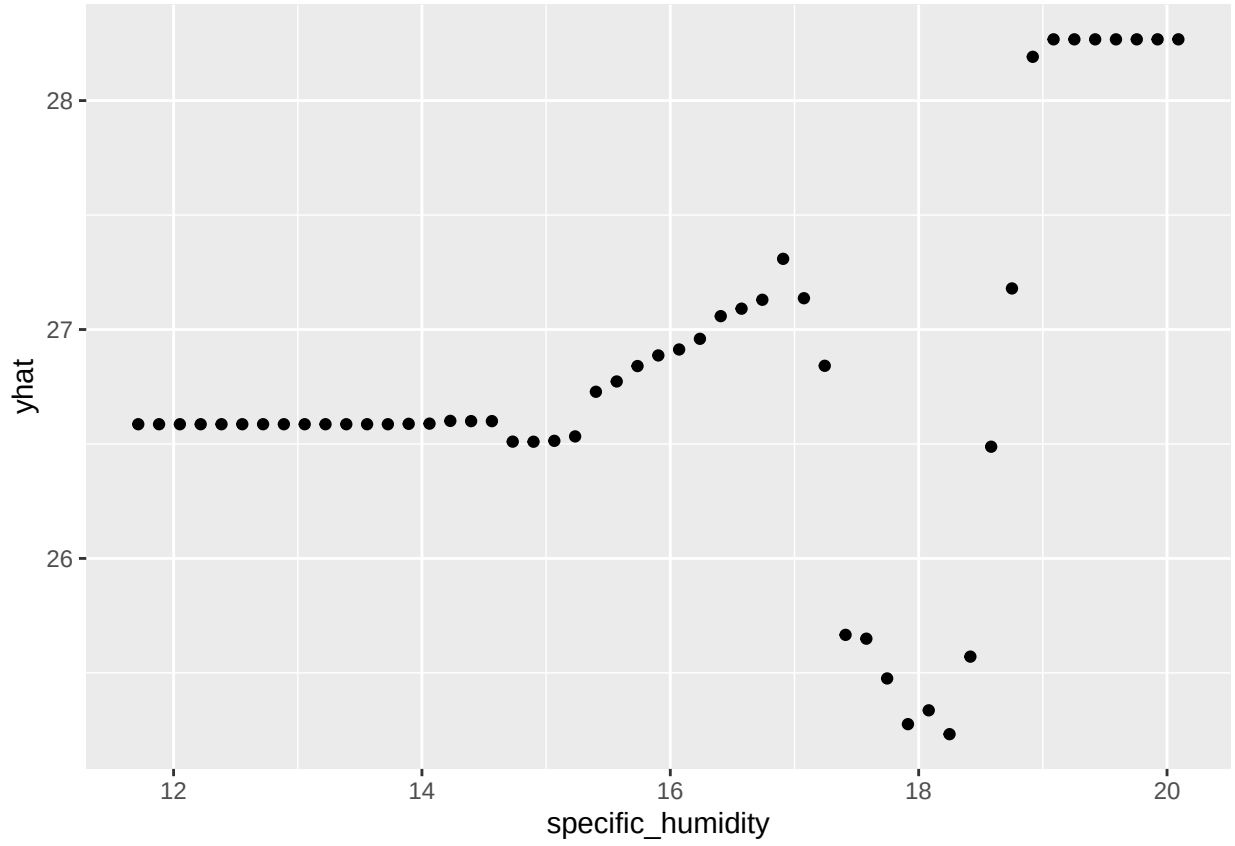
```
## [1] 34.06941
```

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

```
##    specific_humidity     yhat
## 1           11.71571 26.58632
## 2           11.88323 26.58632
## 3           12.05074 26.58632
## 4           12.21826 26.58632
## 5           12.38577 26.58632
## 6           12.55329 26.58632
## 7           12.72080 26.58632
```

```
## 8           12.88831 26.58632
## 9           13.05583 26.58632
## 10          13.22334 26.58632
## 11          13.39086 26.58632
## 12          13.55837 26.58632
## 13          13.72589 26.58632
## 14          13.89340 26.58797
## 15          14.06091 26.58885
## 16          14.22843 26.60083
## 17          14.39594 26.59975
## 18          14.56346 26.59975
## 19          14.73097 26.51020
## 20          14.89849 26.50979
## 21          15.06600 26.51365
## 22          15.23351 26.53304
## 23          15.40103 26.72804
## 24          15.56854 26.77305
## 25          15.73606 26.84033
## 26          15.90357 26.88684
## 27          16.07109 26.91275
## 28          16.23860 26.95938
## 29          16.40611 27.05836
## 30          16.57363 27.09096
## 31          16.74114 27.12990
## 32          16.90866 27.30876
## 33          17.07617 27.13672
## 34          17.24369 26.84154
## 35          17.41120 25.66617
## 36          17.57871 25.64947
## 37          17.74623 25.47578
## 38          17.91374 25.27606
## 39          18.08126 25.33648
## 40          18.24877 25.23236
## 41          18.41629 25.57105
## 42          18.58380 26.48802
## 43          18.75131 27.17946
## 44          18.91883 28.19128
## 45          19.08634 28.26752
## 46          19.25386 28.26752
## 47          19.42137 28.26752
## 48          19.58889 28.26752
## 49          19.75640 28.26752
## 50          19.92391 28.26752
## 51          20.09143 28.26752
```

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




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

```
## [1] 30.01402
```

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

```
## [1] 30.12404
```

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

```
## [1] 26.66075
```

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 containing one green-certified building and at least one other non-green-certified building within a quarter-mile radius of the cluster center.

size: the total square footage of available rental space in the building.

empl.gr: the year-on-year growth rate in employment in the building's geographic region.

stories: the height of the building in stories.

age: the age of the building in years.

renovated: whether the building has undergone substantial renovations during its lifetime.

class.a, class.b: indicators for two classes of building quality (the third is Class C). These are relative classifications within a specific market. Class A buildings are generally the highest-quality properties in a given market. Class B buildings are a notch down, but still of reasonable quality. Class C buildings are the least desirable properties in a given market.

green.rating: an indicator for whether the building is either LEED- or EnergyStar-certified.

net: an indicator as to whether the rent is quoted on a "net contract" basis. Tenants with net-rental contracts pay their own utility costs, which are otherwise included in the quoted rental price.

amenities: an indicator of whether at least one of the following amenities is available on-site: bank, convenience store, dry cleaner, restaurant, retail shops, fitness center.

cd.total.07: number of cooling degree days in the building's region in 2007. A degree day is a measure of demand for energy; higher values mean greater demand. Cooling degree days are measured relative to a baseline outdoor temperature, below which a building needs no cooling.

hd.total07: number of heating degree days in the building's region in 2007. Heating degree days are also measured relative to a baseline outdoor temperature, above which a building needs no heating.

total.dd.07: the total number of degree days (either heating or cooling) in the building's region in 2007.

Precipitation: annual precipitation in inches in the building's geographic region.

Gas.Costs: a measure of how much natural gas costs in the building's geographic region.

Electricity.Costs: a measure of how much electricity costs in the building's geographic region.

City_Market_Rent: a measure of average rent per square-foot per calendar year in the building's local market.**

Results

```
rmse_lm
```

```
##  result
## 1061.96
```

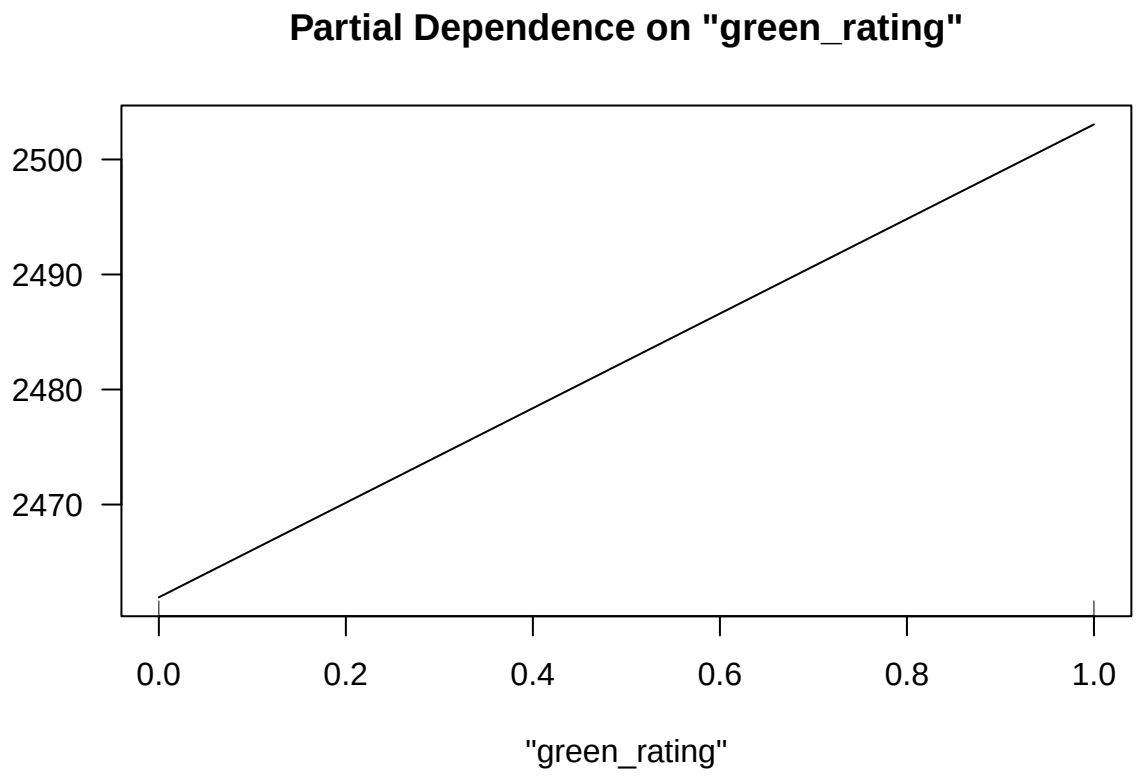
```
rmse_forest_green
```

```
## [1] 751.55
```

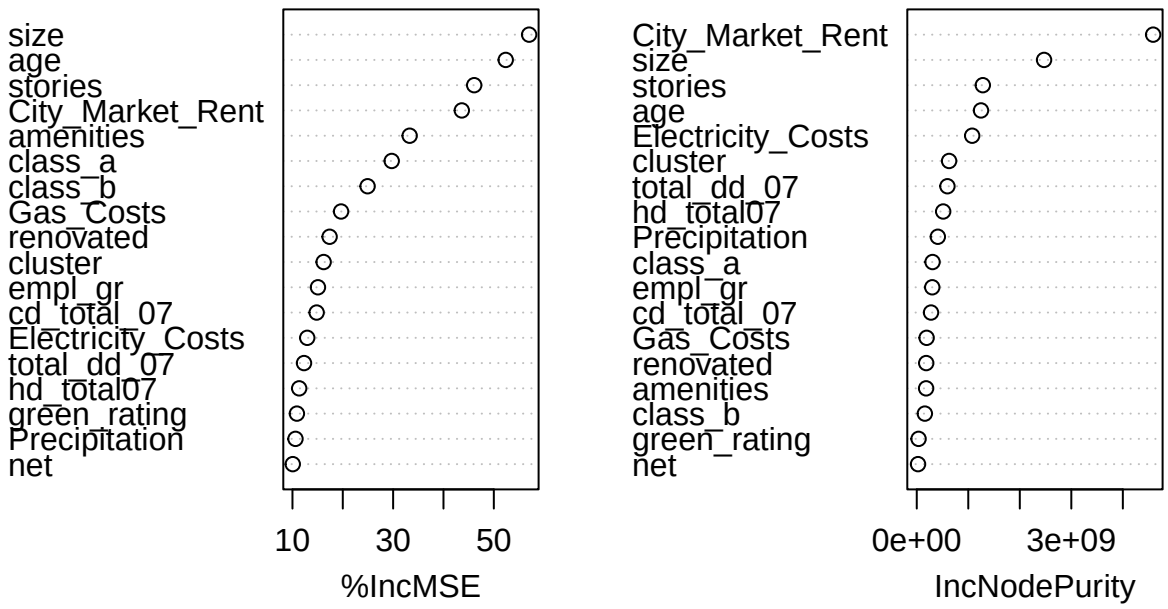
```
rmse_boost_green
```

```
## [1] 960.38
```

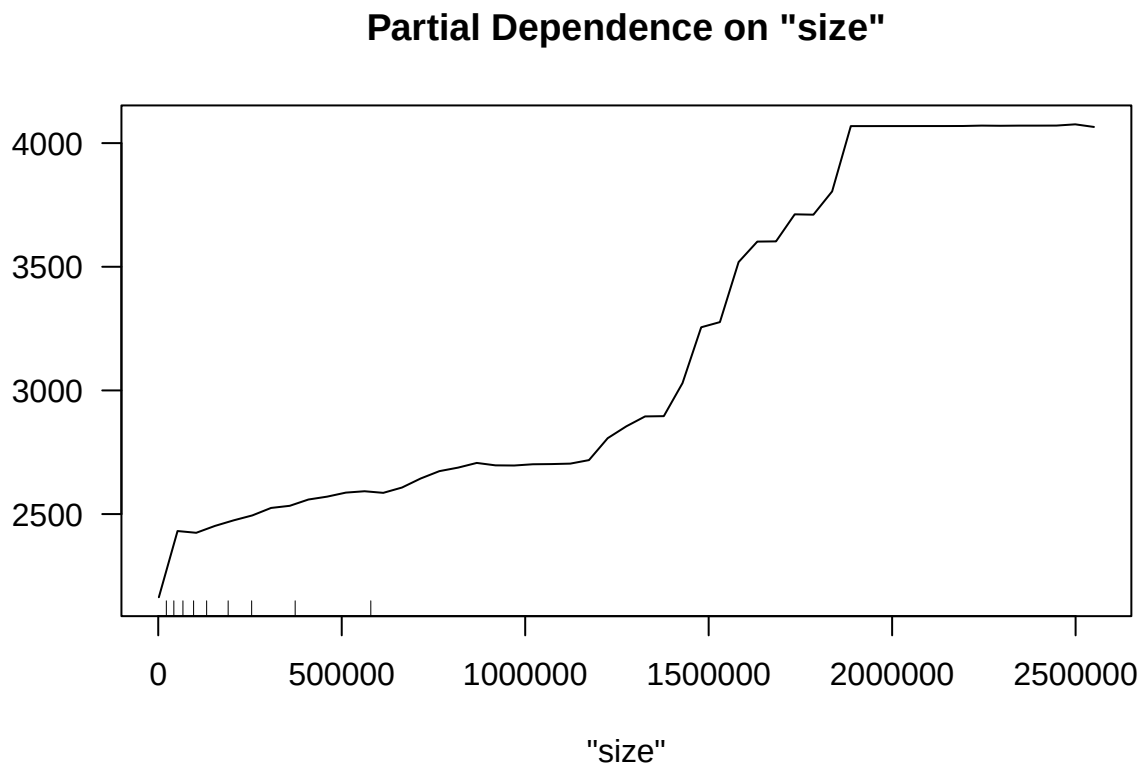
```
partialPlot(green_forest, as.data.frame(green_test), 'green_rating', las = 1)
```



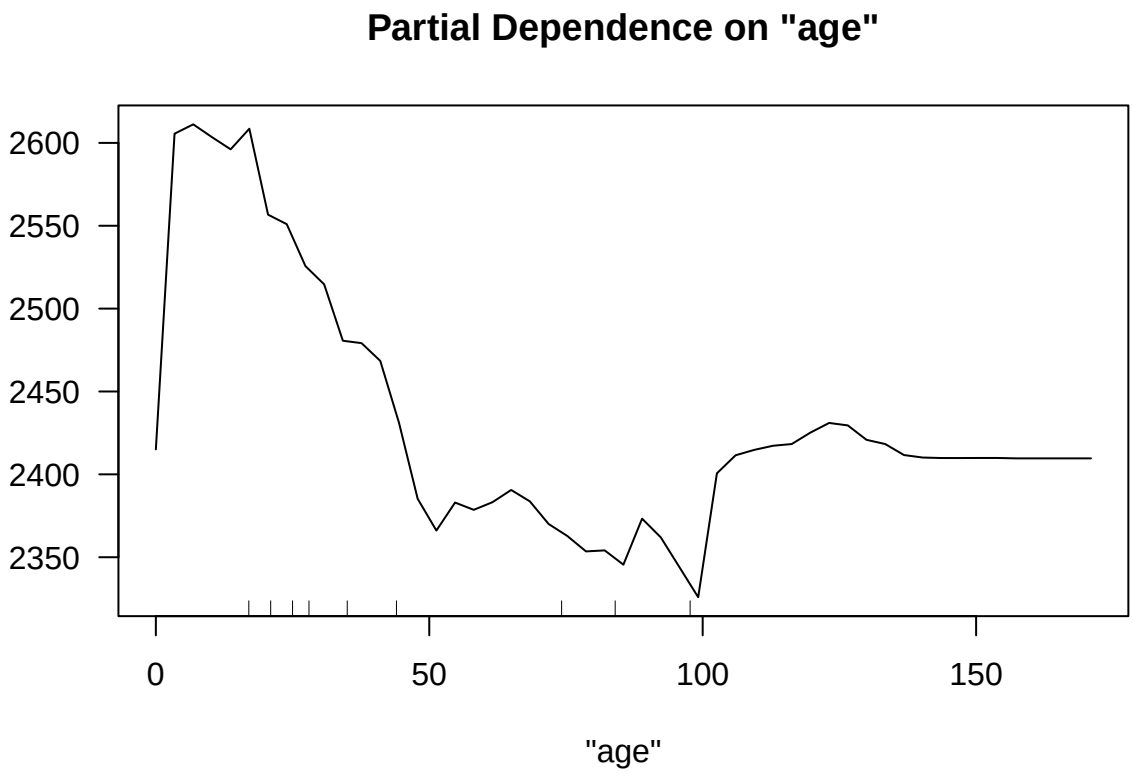

```
varImpPlot(green_forest)
```




```
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"
2600
2550
2500
2450
2400
2350
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.

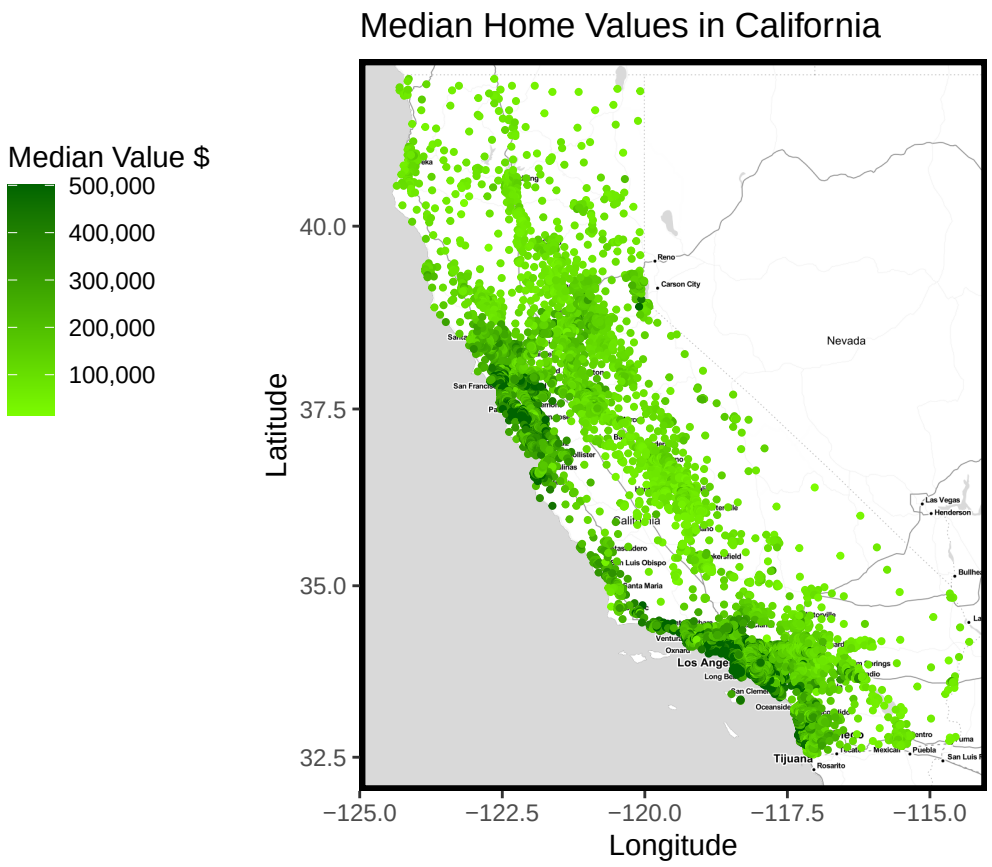
```
rmse_lm_houses
```

```
##  result
## "70253"
```

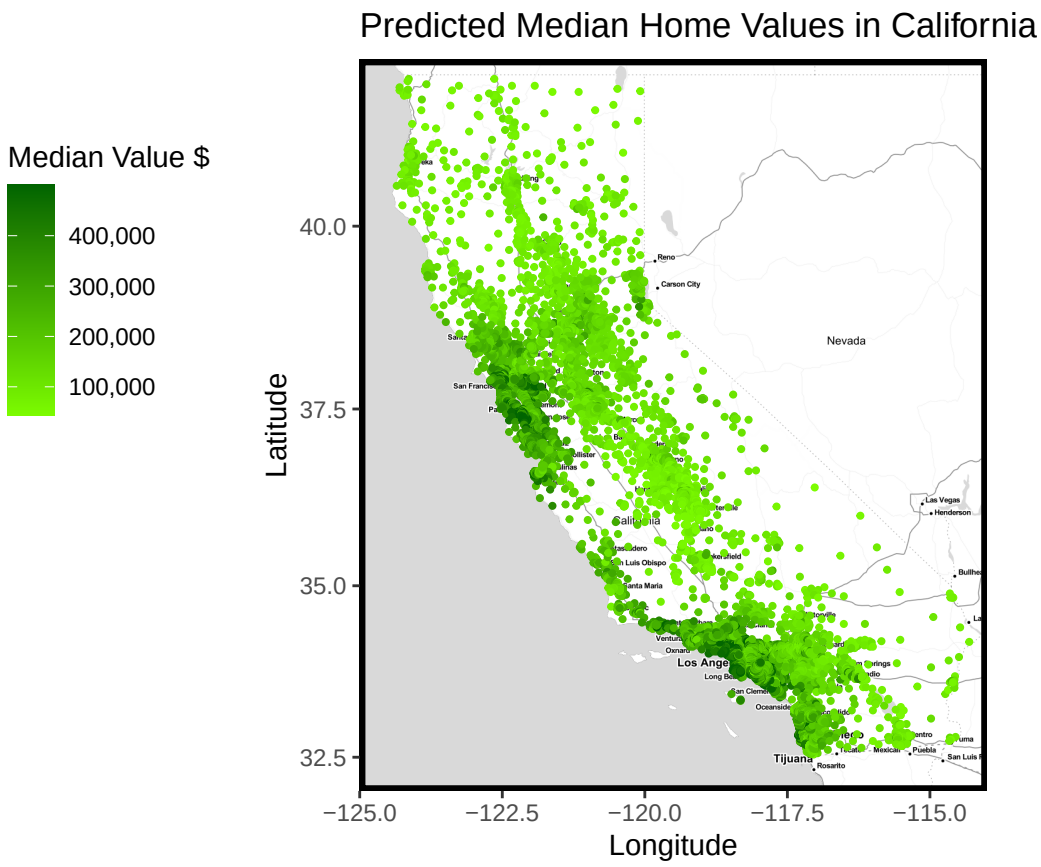
```
rmse_forest_houses
```

```
## [1] "50322.52"
```

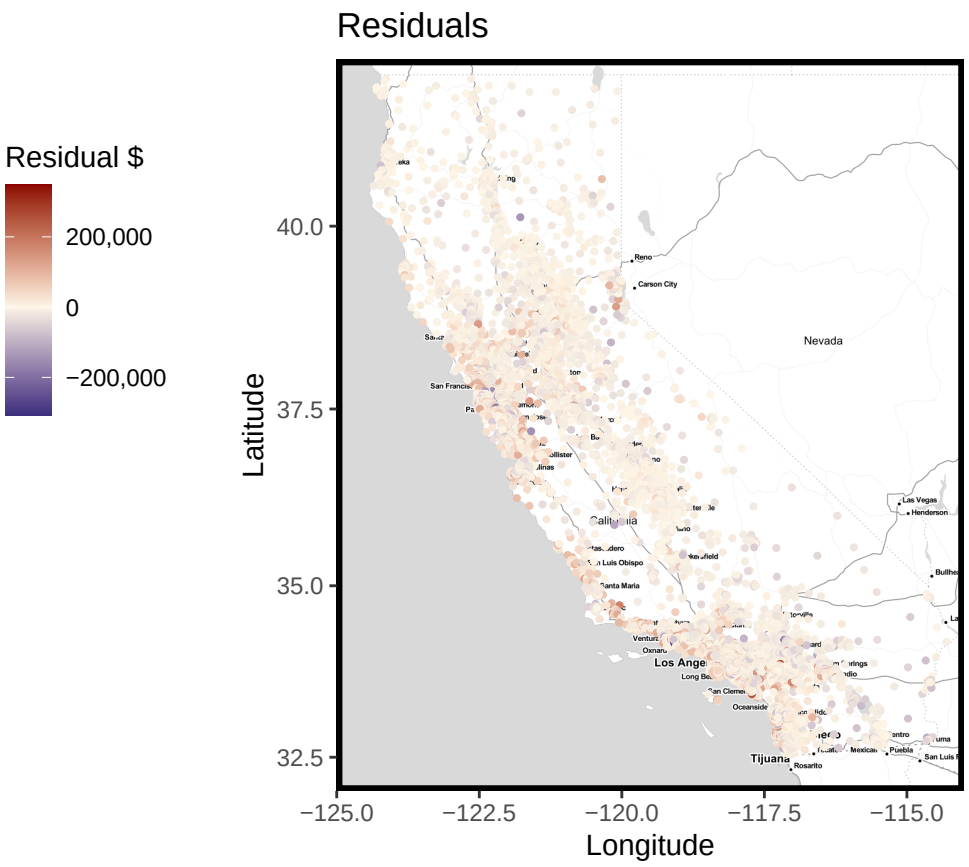
```
data_map
```


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
Las Vegas
Henderson
Los Angeles
Tijuana

```
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
Las Vegas
Henderson
Los Ange
Tijuana

```
resid_map
```


Residuals
Residual $
200,000
0
−200,000
Latitude
40.0
37.5
35.0
32.5
−125.0
−122.5
−120.0
−117.5
−115.0
Longitude
Nevada
Reno
Carson City
Las Vegas
Henderson
Tijuana
Rosarito
Puebla