PS3

Yuri Lee, Seunghoon Choi

2021 4 7

1. What causes what?

1) The city with high crime rate tends to hire more policemen. The number of policemen explains the crime rate, and the crime rate explains the number of policemen(endogeneity). The result of the regression of "Crime" on "Police" may show that more policemen, higher crime rate.

2) The researchers from UPenn used the variables which are not related to the crime rate. They used the terrorism alert system as an instrumental variable. When the terror alert level goes high, more policemen are thrown in. So, the researchers can find the effect of the additional policemen on the crime. The first column of table 2 shows the effect of police on crime by using "High Alert" variable. This says that additional input of policemen is associated with decrease of the crime rate by 7.316(p-value is significant at 5% level).

3) The terror alert can decrease the number of people who might be potential targets of crime, which makes crime rate go down as well. Additional policemen and decreased number of people(such as tourists) can decrease crime simultaneously. So, the researchers split effects of the number of potential victims on crime by using the metro ridership. The second column of table 2 shows the effect of police and victim on crime. The effect of additional policemen on the lower crime is around 6(less than Q.2). Since fewer potential victims also are related to less crimes.

4) Table 4 shows the interaction variables between high alert and districts. The interaction variables show more precise effect of police on crime. When high alert was announced in the district 1, and additional police force was input there, the daily total number of crimes decreased by 2.6 with a significant level of p-value. However, if high alert was announced in other area and police force was dispatched there, there was no significant level of effects on the crime in the district 1. The results show that only the area where had additional police force experienced crime decrease, which means additional police force caused decrease of crime.

2. Predictive model building: green certification

1) Overview

Our goal is to build the best predictive model for revenue(per square foot per calendar year). In order to get best model, we will use 'Step wise', 'CART', 'Random Forests', Boosting' models, and compare the RMSE of these models. And by using this model, we will quantify the average change in rental income associated with green certification. We chose 'green_rating' as "green certified" category, rather than 'LEED' and 'Energystar'

2) Data and model

2-1) data cleansing

In data, 74 rows have omitted in 'empl_gr' variable, so we omit them

2-2) mutation of 'revenue', using 17 variables

The revenue per square foot per year is the product of two terms: rent and leasing_rate. So, we apply 'revenue = Rent * (leasing_rate/100)' to models

2-3) fit models

First, data split to training and testing

Second, we built 'Step wise', 'CART', 'Random Forests', Boosting' models. And we used 17 variables, except 3 variables(LEED, Energystar, cluster)

3) Results

3-1) compare the RMSE of 4 models

The RMSE of 'Random Forests' model is lowest, so we select 'Random Forests' model.

a) Step wise

```
## [1] 9.746803
```

b) CART

```
## [1] 9.441961
```

c) Random Forests

```
## [1] 6.850442
```

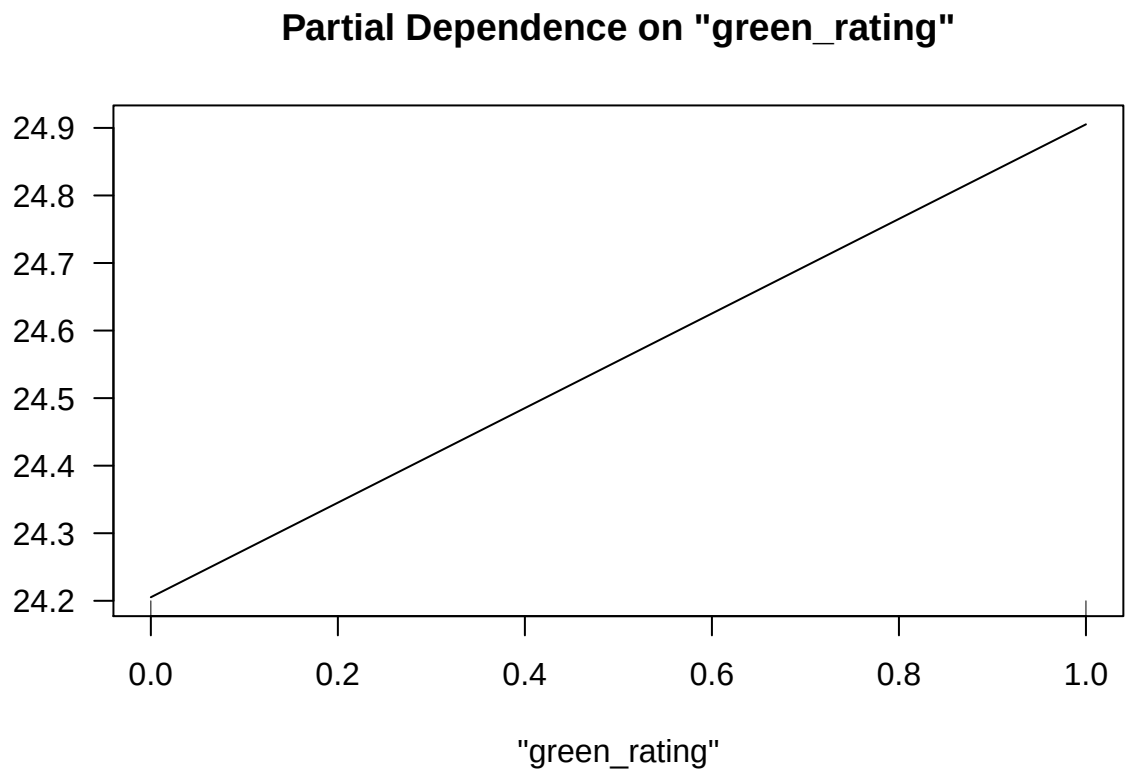
d) Boosting

```
## [1] 8.802109
```

3-2) partial effect of 'green_rating'

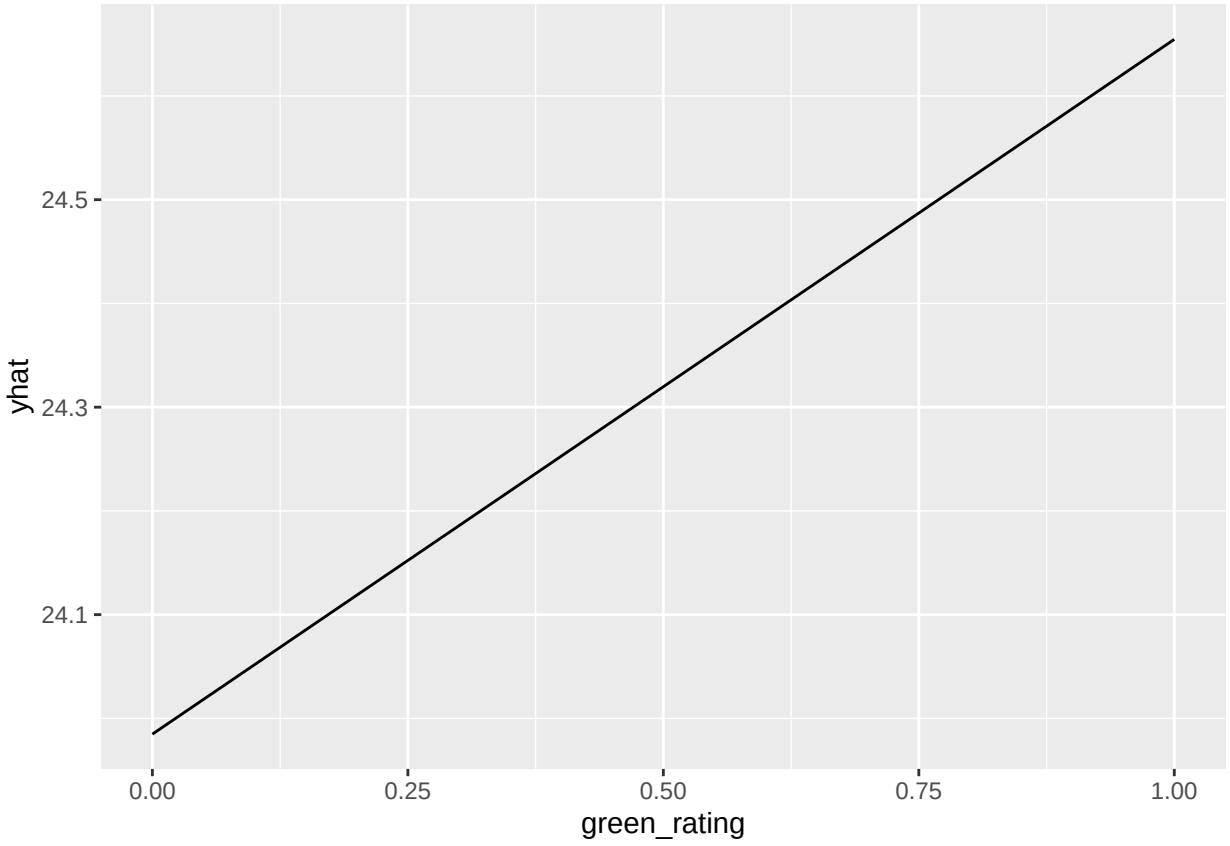
Partial effect of 'green_rating' is small (around 0.7 unit)

```
##   green_rating     yhat
## 1            0 23.99050
## 2            1 24.75868
```

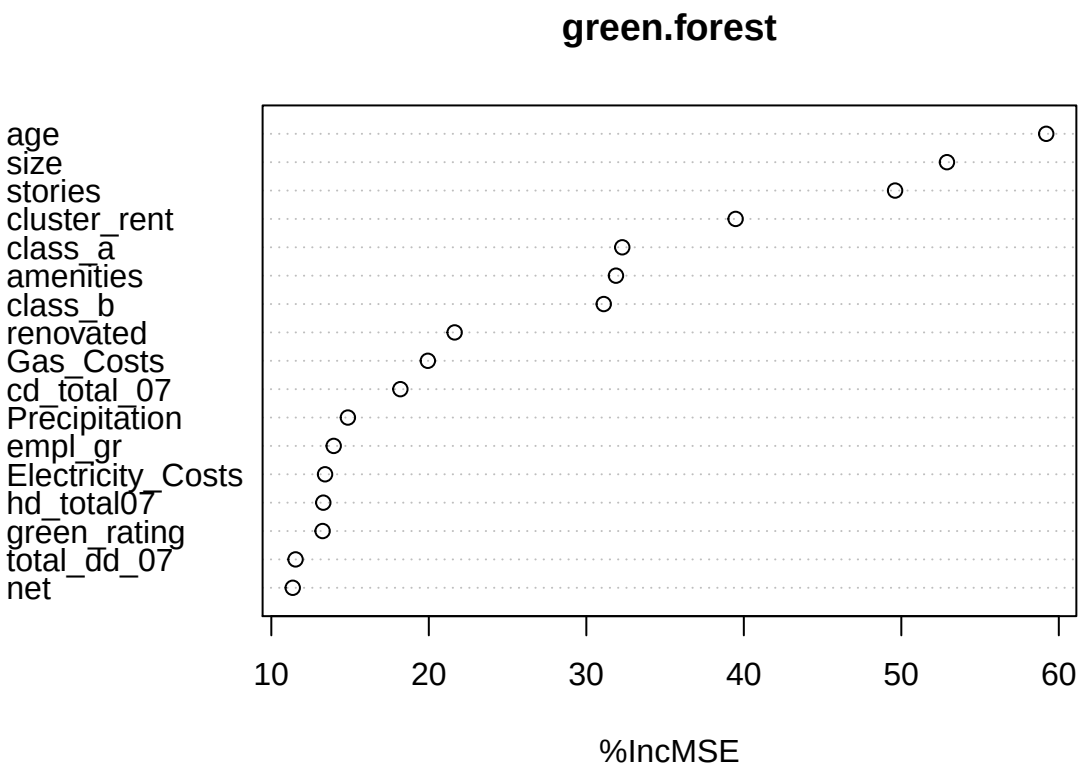


In 'Boosting' model, the partial value of 'green_rating' is simillar to that of 'Random Forests' model (around 0.8 unit)

```
##   green_rating     yhat
## 1            0 23.98479
## 2            1 24.65455
```



In variable importance measures, 'green_rating' variable is less important than other variables.



4) Conclusion

In order to get best model, we used 'Step wise', 'CART', 'Random Forests', Boosting' models. The RMSE of 'Random Forests' model is lowest, so we selected 'Random Forests' model. And by using this model, we quantified the average change in rental income('revenue') associated with 'green_rating'. Lastly, we calculated partial effect of 'green_rating', the value is small. And 'green_rating' is less important than other variables.

3. Predictive model building: California housing

1) Overview

The purpose of this analysis is to fit the prediction model for median housing price. The data contains 20,640 observations about housing tract in California with nine variables. We can guess that house price would be highly related with median income among variables, and especially, the interaction variable between latitude and longitude would be dominating effect on house price from the scatter plot. So, we started fitting the prediction model for house price with these three variables.

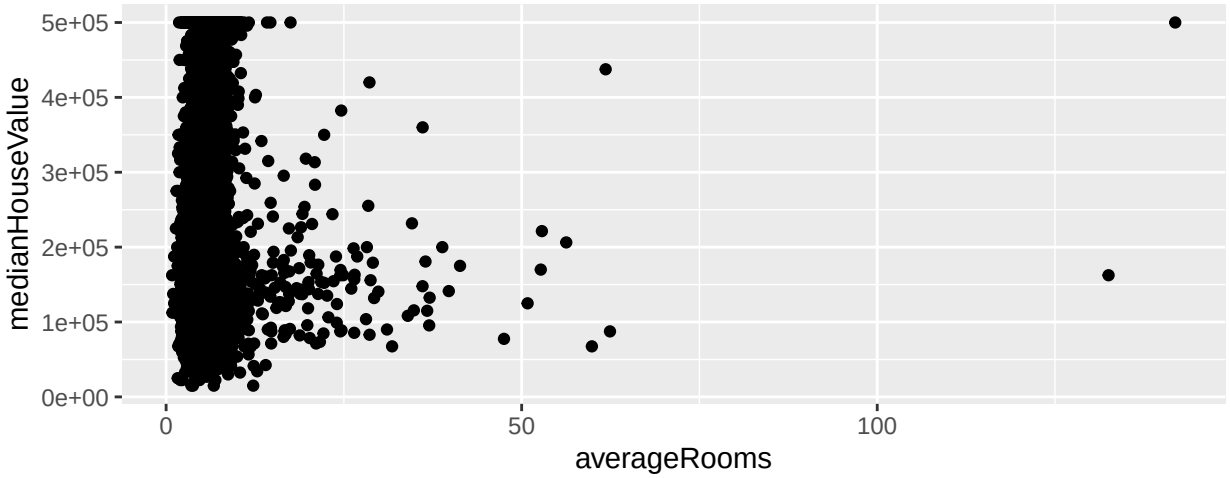
2) Data featuring and data exploring

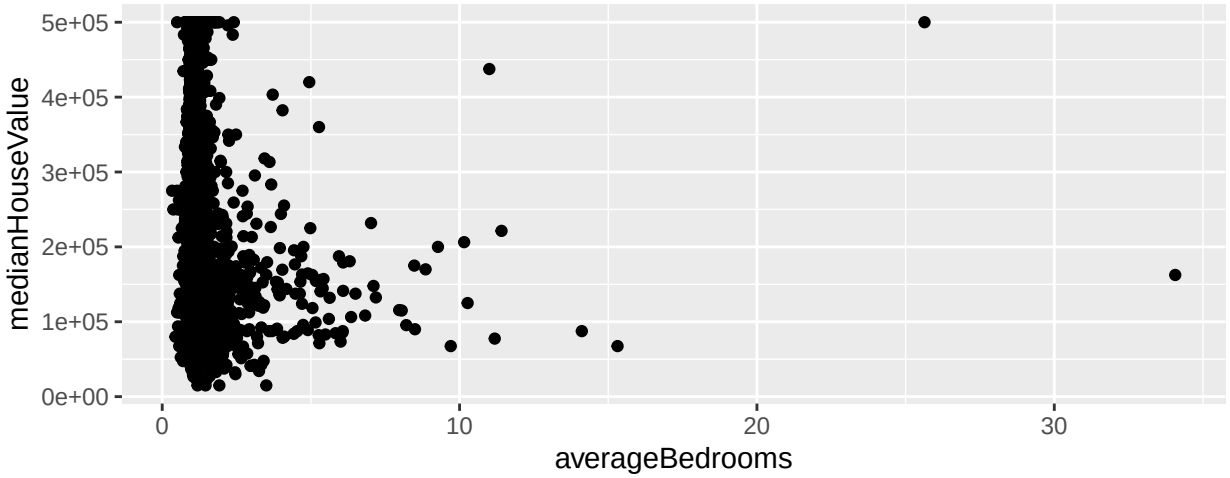
We added new variables: averageRooms, averageBedrooms and unitBedrooms using mutate command, and, scanned the relations betwwen medianHousingValue and other variables. When we plot the data to show medainHouseValue versus longitude(x) and latitude(y), we could see the distriubution of medianHousingValue on geographical grid.

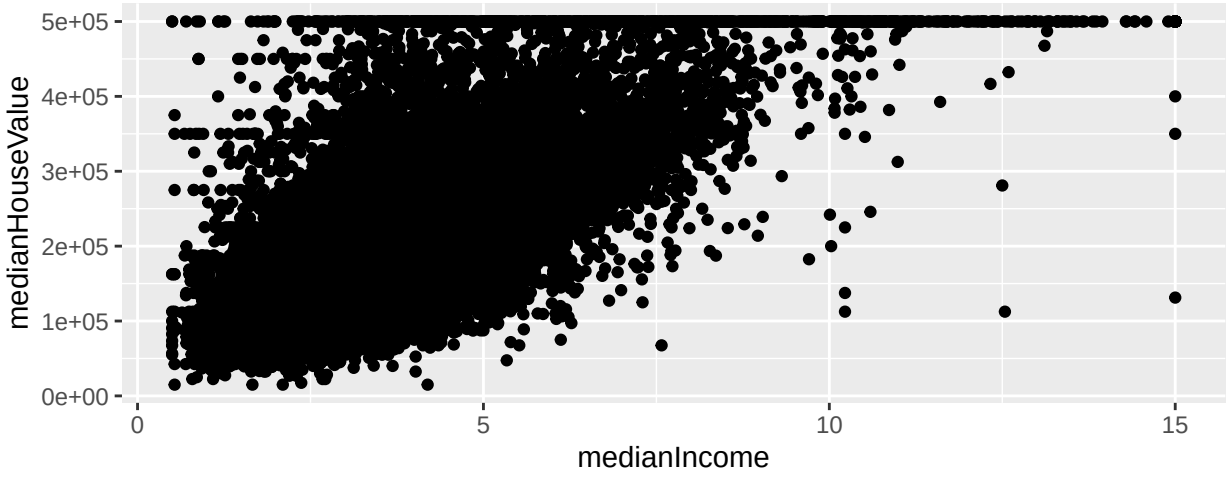
2-1) data featuring

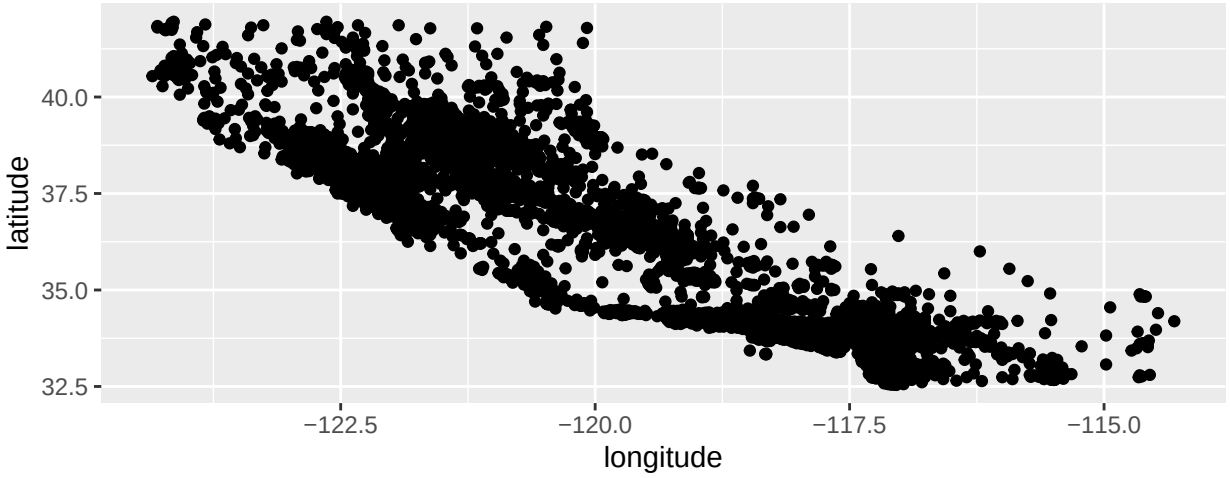
We mutated 'averageRooms = totalRooms/households', 'averageBedrooms = totalBedrooms/households', 'unitBedrooms = totalBedrooms/population'

2-2) data explore

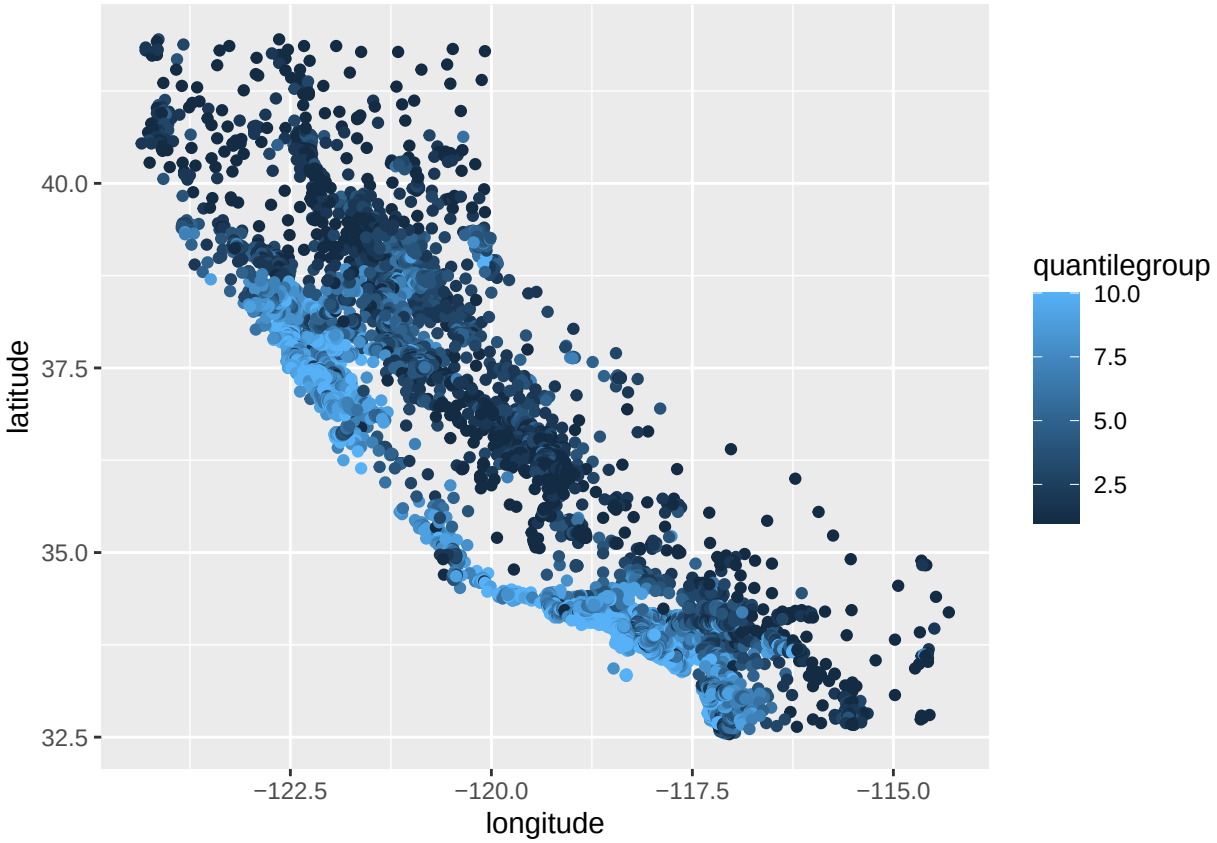



medianHouseValue
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
10
20
30
averageBedrooms


medianHouseValue
5e+05
4e+05
3e+05
2e+05
1e+05
0e+00
0
5
10
15
medianIncome


latitude
40.0
37.5
35.0
32.5
−122.5
−120.0
−117.5
−115.0
longitude

2-3) scatter plot



3) fitting models

In order to fit the prediction model, we did multiple data splits, and fitted several linear model using stepwise selection. Then, we calculated out-of-sample rmses from the models. We could get the average rmses from three models.

3-1) data split

we split data into training and testing

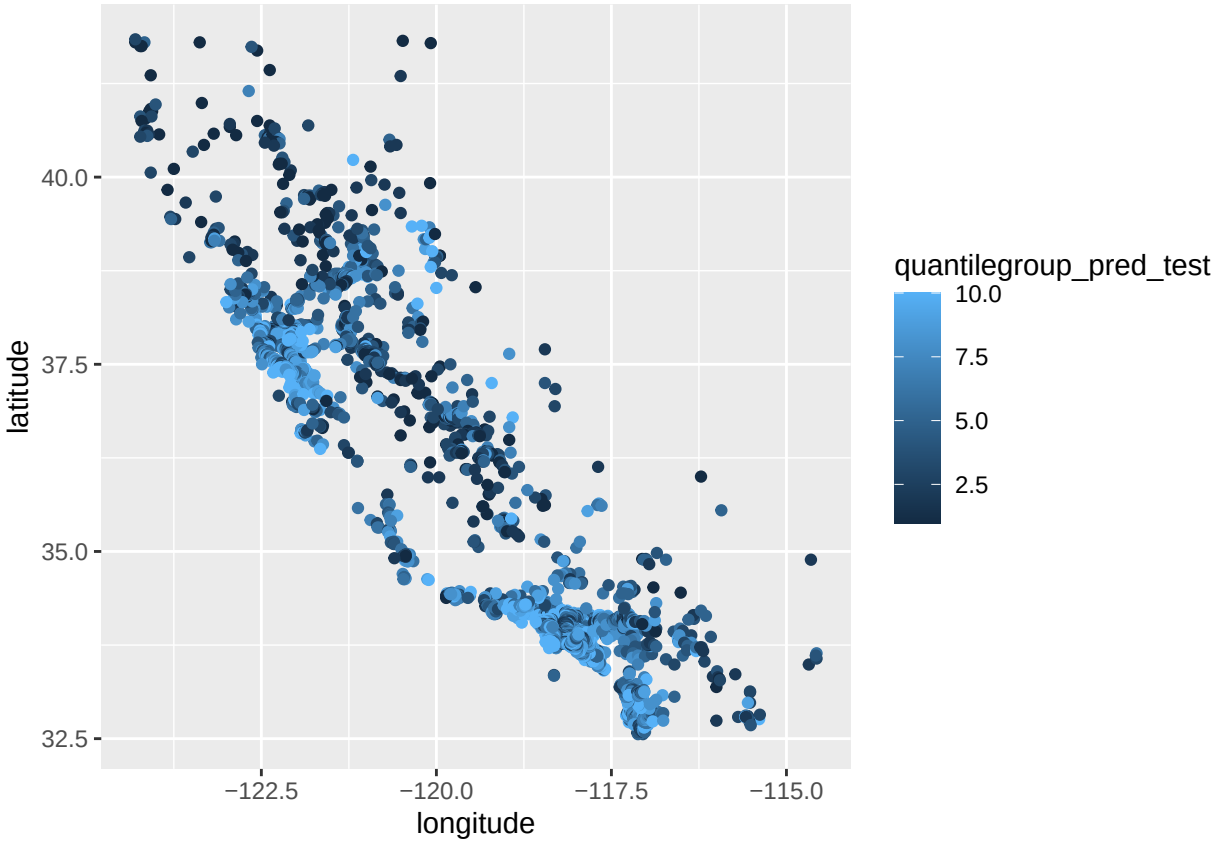
3-2) Fit linear models: base model & step wise selection with AIC

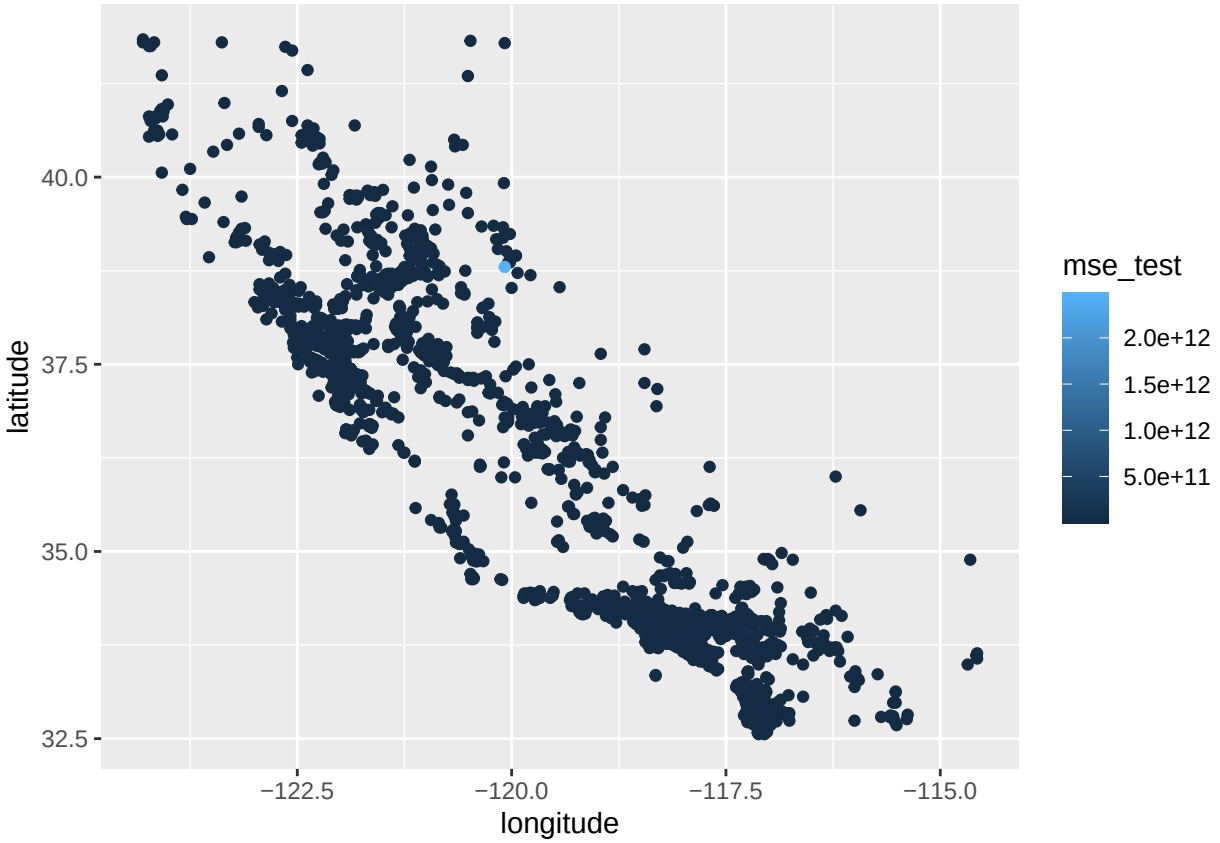
3-3) Averaging the performance over 10 train/test splits

```
##       V1       V2       V3
## 83739.55 82598.65 82807.09
```

4) Prediction with the step wise(lm3) model and plot the predictions and errors

If we predict for medianHousingValue with test set, we can get a plot of predictions as below. Also, we can calculate and plot the error term(mse) from it


40.0
37.5
35.0
32.5
latitude
−122.5
−120.0
−117.5
−115.0
longitude
quantilegroup_pred_test
10.0
7.5
5.0
2.5



5) Conclusion

We can get low level of errors through the overall test data set, which means the model we fit can generally predict the median housing values well.