Exercise 3

Zhiqian Chen, Yi Zeng, Qihang Liang

4/8/2021

##1. What cause what?

#(a) The reason here is that data on police and crime cannot tell the difference between more police causing crime or more crime leading to the need for more police. In fact, we would like to see a potential positive correlation between crime and police in different cities, and the government may respond to the increase in crime by hiring more police. But we are unable to randomly place the police on the streets of the city on different days to see what happens.

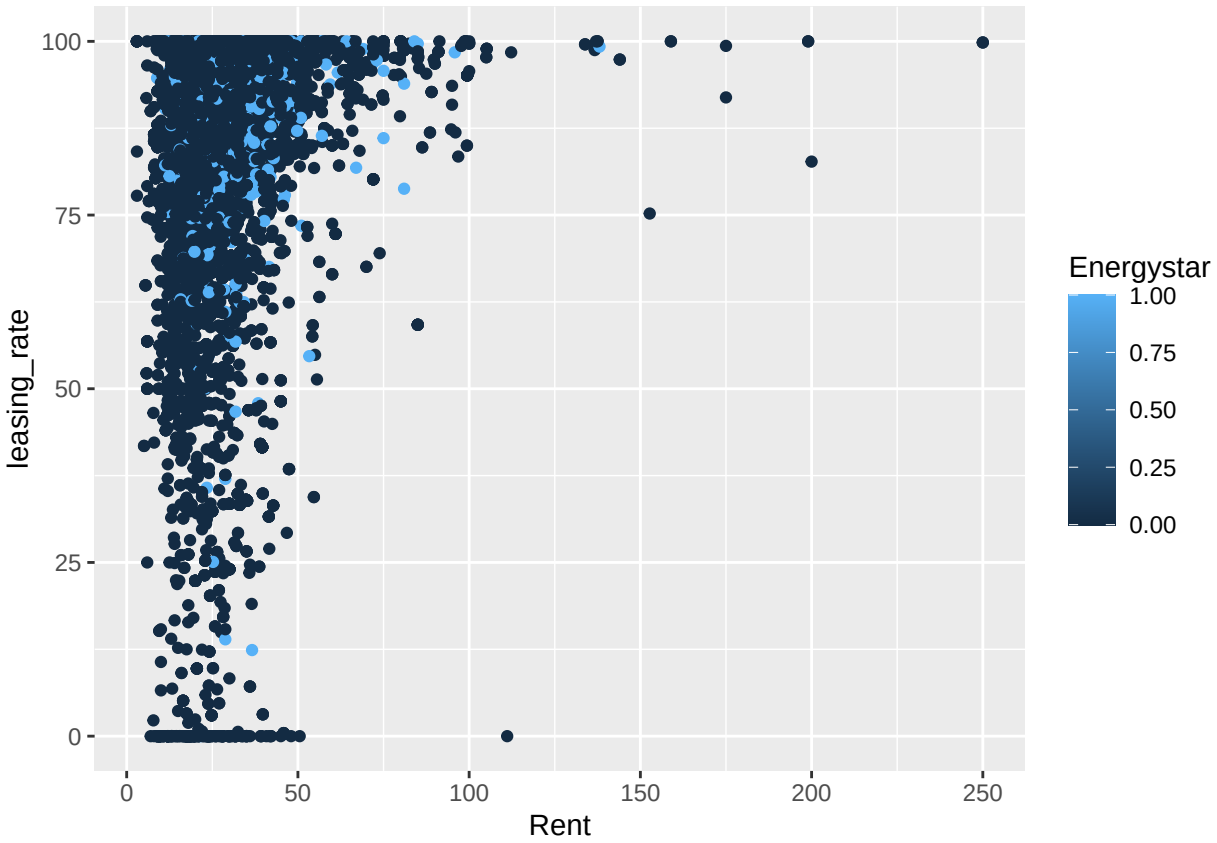
#(b) Researchers from UPenn want to use an estimation method called natural experiment to show that more police were recruited for reasons unrelated to crime. On days of high alert, the mayor of DC must send more police officers on the street. The decision has nothing to do with crime. They collect data on crimes in Washington, D.C., and link it to the days of increased vigilance for possible terrorist attacks. From Table 2, we can see that the coefficients of high alert are -7.316 and -6.046 respectively, which indicates that there are some confounding effects, which may induce omitted variable bias. For example, the midday ridership may be related to crime, because on the day of high alert, people travel less, so the crime rate drops. This impact is not caused by the increase in police. The results from the Table 2 tells us that holding midday ridership fix more police has a negative impact on crime.

#(c) If people go out on a high alert day, there would be fewer opportunities for crimes and hence less crime, which is unrelated to more police. But even though we control midday ridership, we still cannot prove that more police can reduce crime. This is because if that day is a high-vigilance day, criminals may therefore not plan to go out to commit a crime, which will lead to a situation where there are more police and fewer crimes.

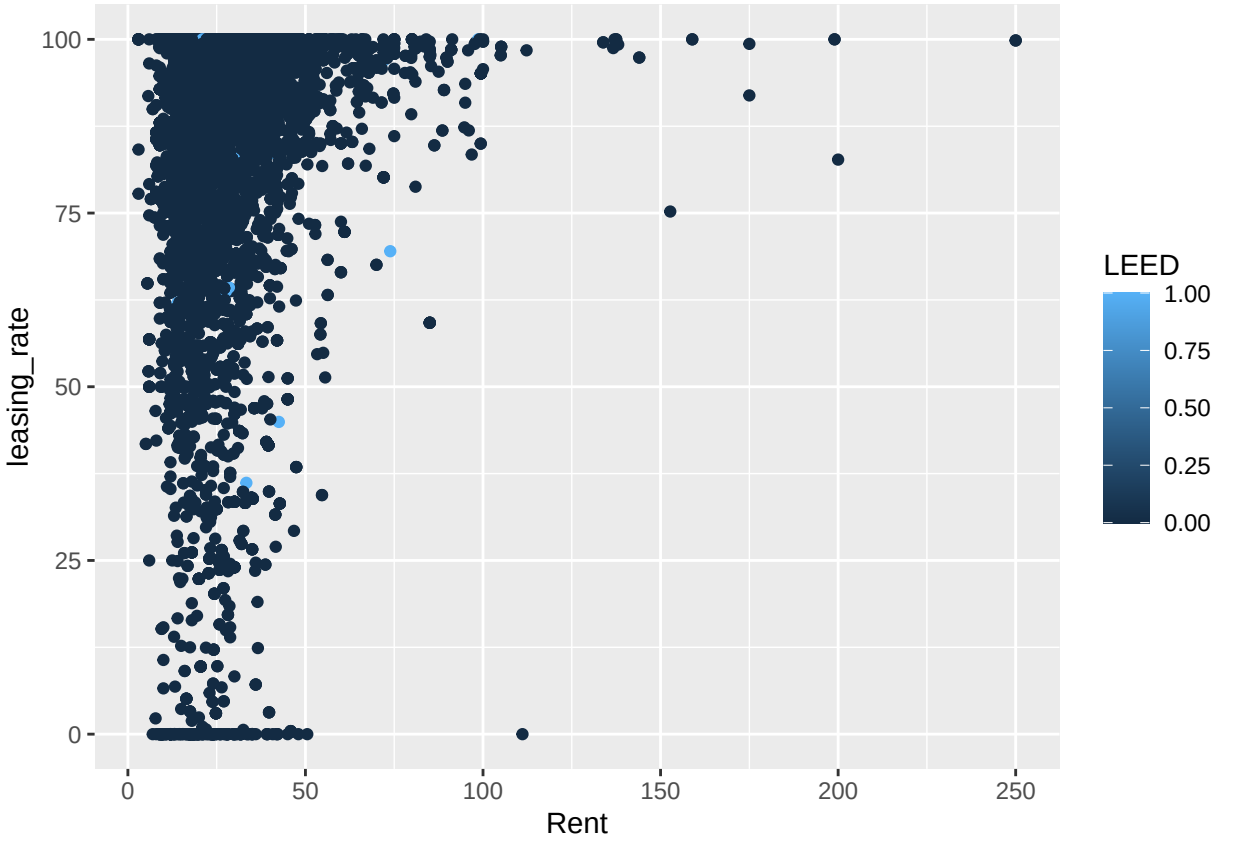
#(d) In Table 4, we can see that the researchers added new variables to determine whether the impact of high alert days on crime is the same in all areas of the town. We can see that the coefficients of Hight Alert*District 1 and Hight Alert*Other Districts are -2.621and -0.571. By using the interaction between the location and the high alert day, we can find that the impact is only obvious in zone 1, which makes sense because most potential terrorist targets are in zone 1, so it is more likely more police were deployed. Although the coefficient of Hight Alert*Other Districts is still negative (-0.571), the coefficient is very close to 0. Considering the standard error in parenthesis, we can conclude that this effect is 0.

2.Predictive model building: green certification

Check the data: At first, we check the basic data, and drop the missing value. Then we create the rent revenue variable which is leasing rate times rent. After that we simply check the relationship between the rent and green ratting, and the relationship between the leasing rate and green rate. Also, we check the relationship between the rent revenue and LEED, and the relationship between the leasing rate and LEED.



From this graph, we can find there are some relationship between leasing rate and energystar, but we find there is almaost no relationship between rent and energystar.



From this graph, we can find there are also some relationship between leasing rate and LEED, but we find there is almaost no relationship between rent and LEED.

Model build: we want to built a best predictive model to estimate the revenue per square foot per calendar year, and to use this model to quantify the average change in rental income per square foot (whether in absolute or percentage terms) associated with green certification. Therefore, we built three models which is the liner regression model, single tree model, and random forest model. We use the revenue as the dependent variable, green rating and LEED as independent variable, and we add some control variable which we think they may affect the rent revenue.

```
##               Model      RMSE  RSquared
## 1 linear regression 1460.6225 0.3067093
## 2       single tree 1095.7192 0.6100147
## 3     random forest  983.3427 0.6858494
```

Model selection: we compute the RMSE for these three models. We find that the RMSE for the random forest model is the lowest in these three model with the highest rsquare. Therefore, the random forest model is the best model we can choose.

```
##                 Length Class  Mode
## call               4   -none- call
## type               1   -none- character
## predicted       6257   -none- numeric
## mse              500   -none- numeric
## rsq              500   -none- numeric
## oob.times       6257   -none- numeric
## importance        34   -none- numeric
```

```
## importanceSD      17   -none- numeric
## localImportance    0   -none- NULL
## proximity          0   -none- NULL
## ntree              1   -none- numeric
## mtry               1   -none- numeric
## forest            11   -none- list
## coefs              0   -none- NULL
## y               6257   -none- numeric
## test               0   -none- NULL
## inbag              0   -none- NULL
## terms              3   terms  call
```

we use this model as our final model

```
##
##  Paired t-test
##
## data:  predict_LEED1 and predict_LEED0
## t = 23.673, df = 1562, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  104.4270 123.2956
## sample estimates:
## mean of the differences
##                113.8613
```

```
##
##  Paired t-test
##
## data:  predict_Energystar1 and predict_Energystar0
## t = 17.769, df = 1562, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  38.75249 48.37001
## sample estimates:
## mean of the differences
##                43.56125
```

We do the t-test for LEED variable and found that the p-value is less than 0.05 which means it is statistically significant at 5% level. we can reject the null hypothesis that the there is no relationship between LEED and rent revenue. Also, we do the t_test for Energystar, we also can not reject the nuall hypothesis that there is no relationship between energystar and rent revenue.Therefore, LEED and Energystar both have significant affect for rent revenue.

Predictive model building: California housing

1.Abstract

In this exercise, our task is to build the best predictive model to predict median house value in the state of California, using the other available feature, like longitude. And also, we need to include three figures in our model. For the model's choice, we consider three option: 1. linear model 2. boosting model and 3. random forest model. Recall that our goal is to predict the median house value by some features of the house, and the random forest model can search for the best feature among a random set of features and results in a wide diversity. So, we think the random forest model is best fit for our goal.

2.Model

Before we start to use the model to analyze the data, we need to clean the data, in other words, we need to processing the data first. For each census tract, variables "totalRooms" and "totalBedrooms" are the total amount for house hold in the tract, we nneed to figure the average amount. And also, for the variables "household" and "population", population is the total number of household in the tract, we need to find average population in each household of the tract. Those avergae results can help us to predict the value of each house.

After processing the data, what we do next is to decide which model is the best model. The first model we build is the linear model. As we said before, we think random forest model is the best model to predict the value of median house. So, the linear model we build here is to compare with the random model, we can think of the linear model as a control model.

```
## [1] 72329.73
```

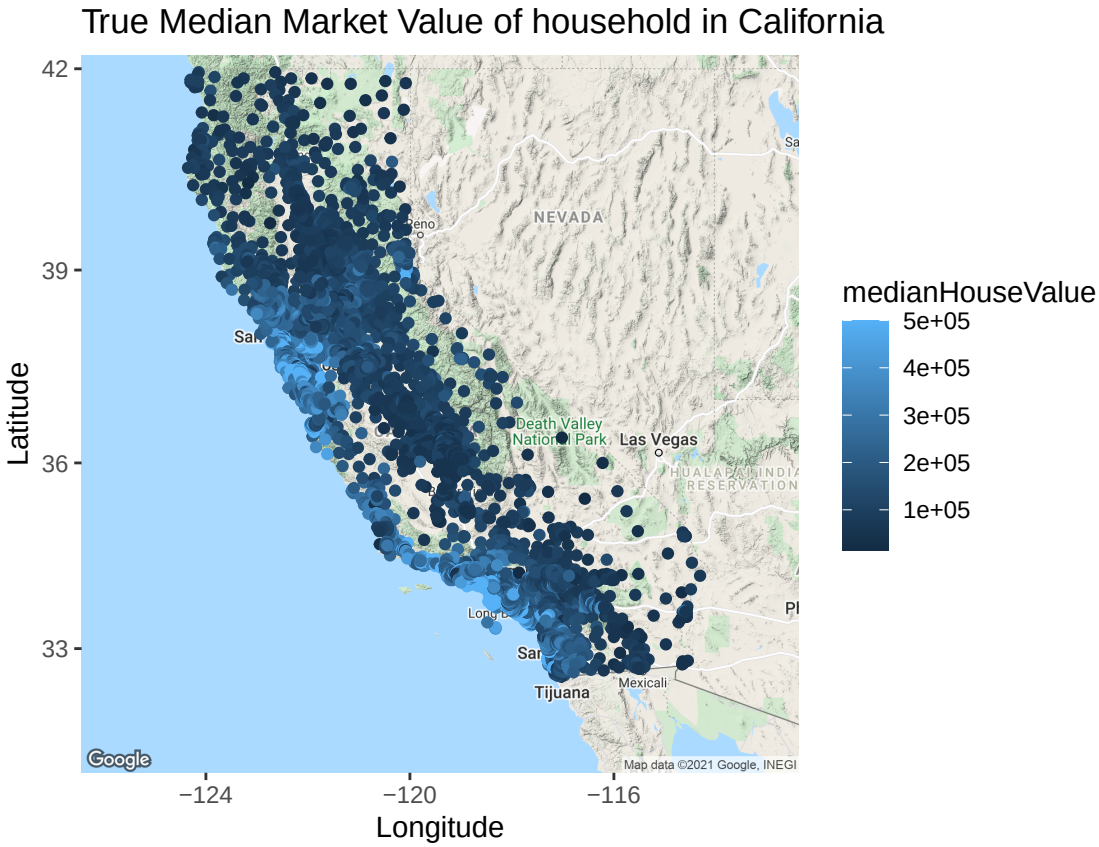
The RMSE of linear model is around 70000. Then, we build a random forest model to compare with the linear model.

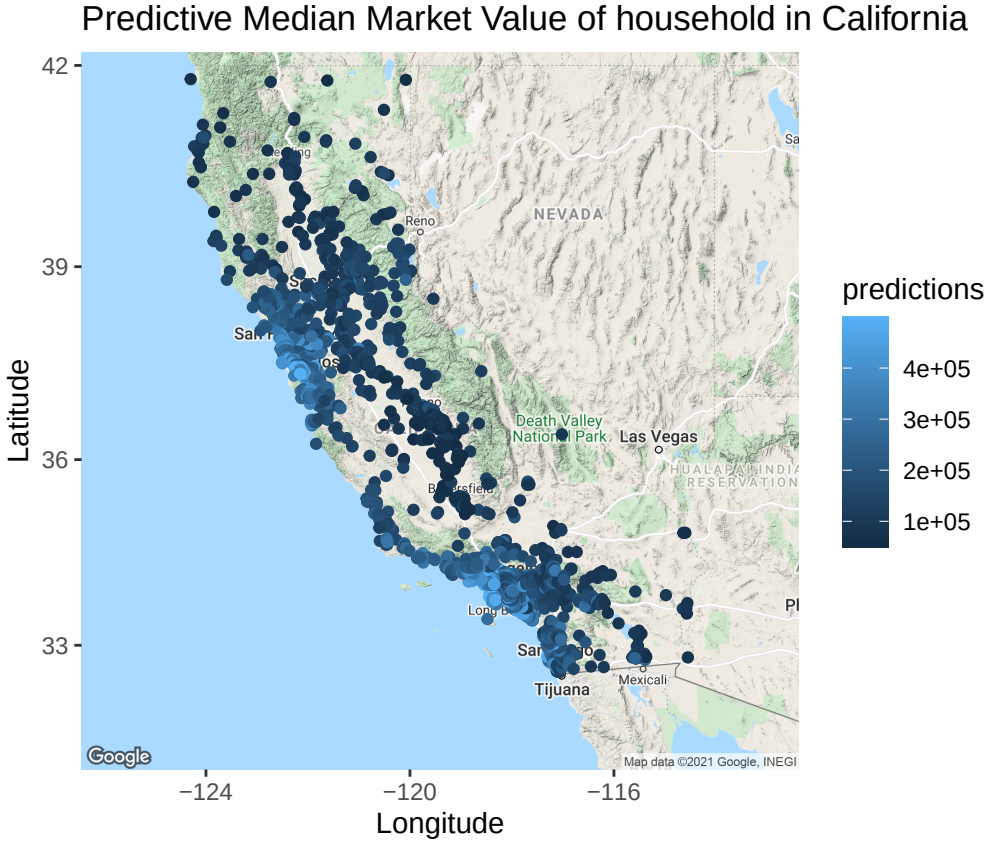
```
## [1] 49333.23
```

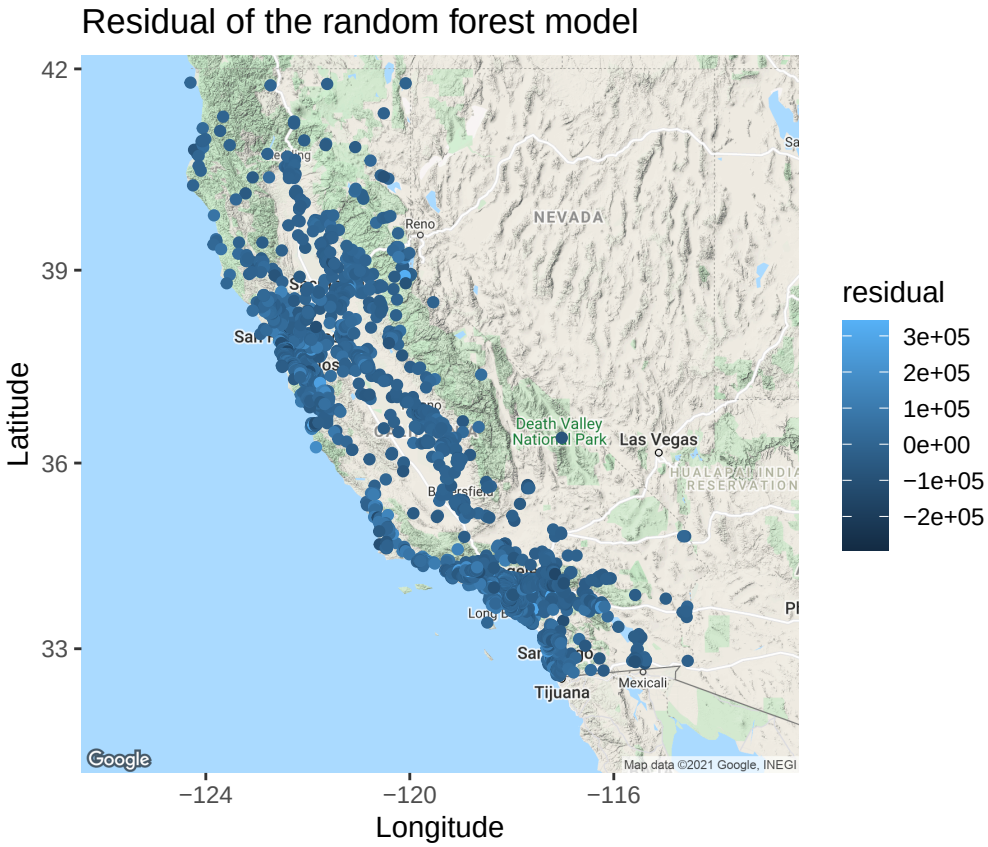
```
## [1] 0.1880716
```

The RMSE of random forest is around 52000. Compared to the RMSE of linear model, the RMSE of Random Forest is nearly 25 percent lower than linear model one. The results verify our previous inference, random forest model is better in predictive than the linear model. And the overall out-of-sample accuracy of the random forest model is showed above. Then, we have generate the predictive value of median house, we use the true value and predictive value to get residuals, then we have the true value, predictive value and residuals, we use those variables to generate those three figures. For the mapping package, since we do not learn about this yet, so we google about the maaping package. So, we google about "google API" and install the ggmap library to create those three figures, also we google about the key, there is tutorials online and we follow the tutorials. I use the google to get the map of California and set as map_CA, then I use the command "ggmap" to get those three figures.

3.Three figures:






Residual of the random forest model
Latitude
42
39
36
33
−124
−120
−116
Longitude
NEVADA
Reno
Death Valley
National Park
Las Vegas
Tijuana
Mexicali
Google
Map data ©2021 Google, INEGI
residual
3e+05
2e+05
1e+05
0e+00
−1e+05
−2e+05