Lab 9: Predicting house prices

Minsu Kang and Bomin Lyoo

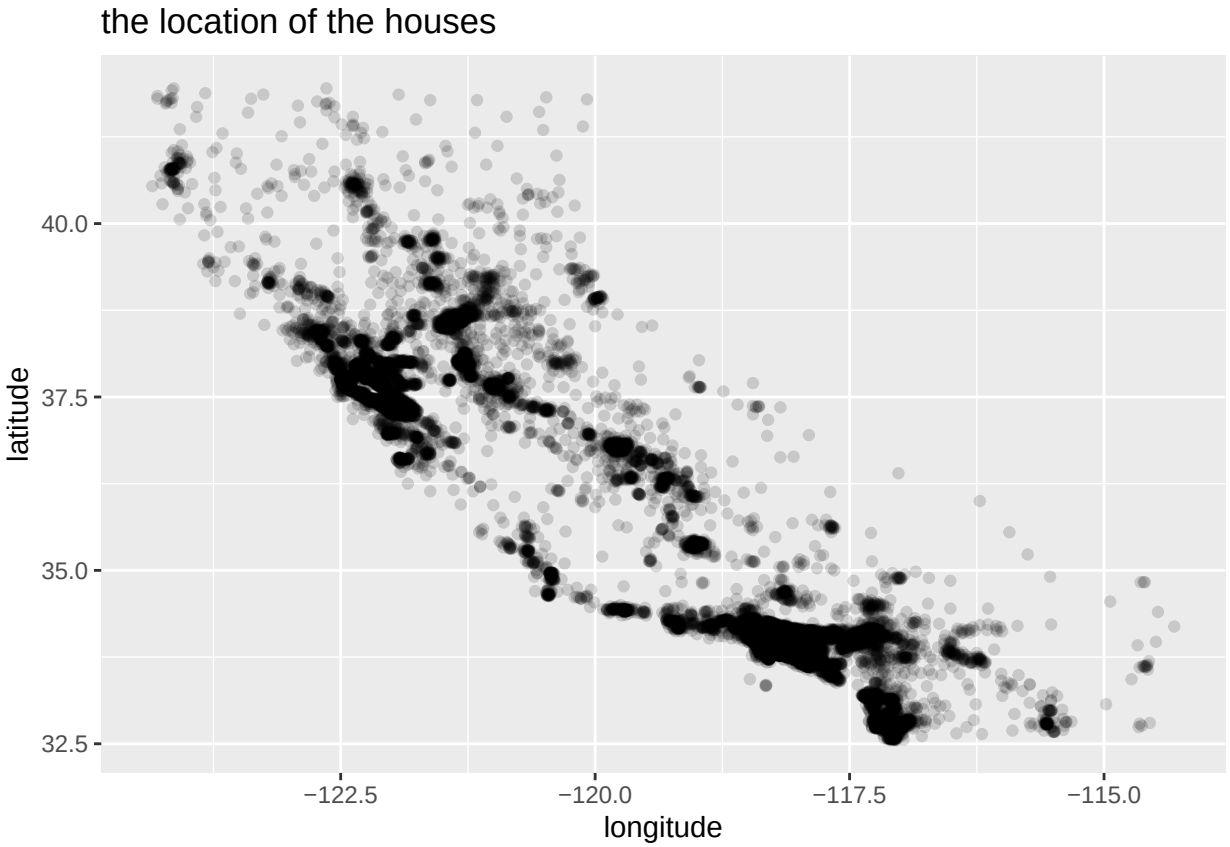
2022-07-17

Lab report

Exercises

Exercise 1

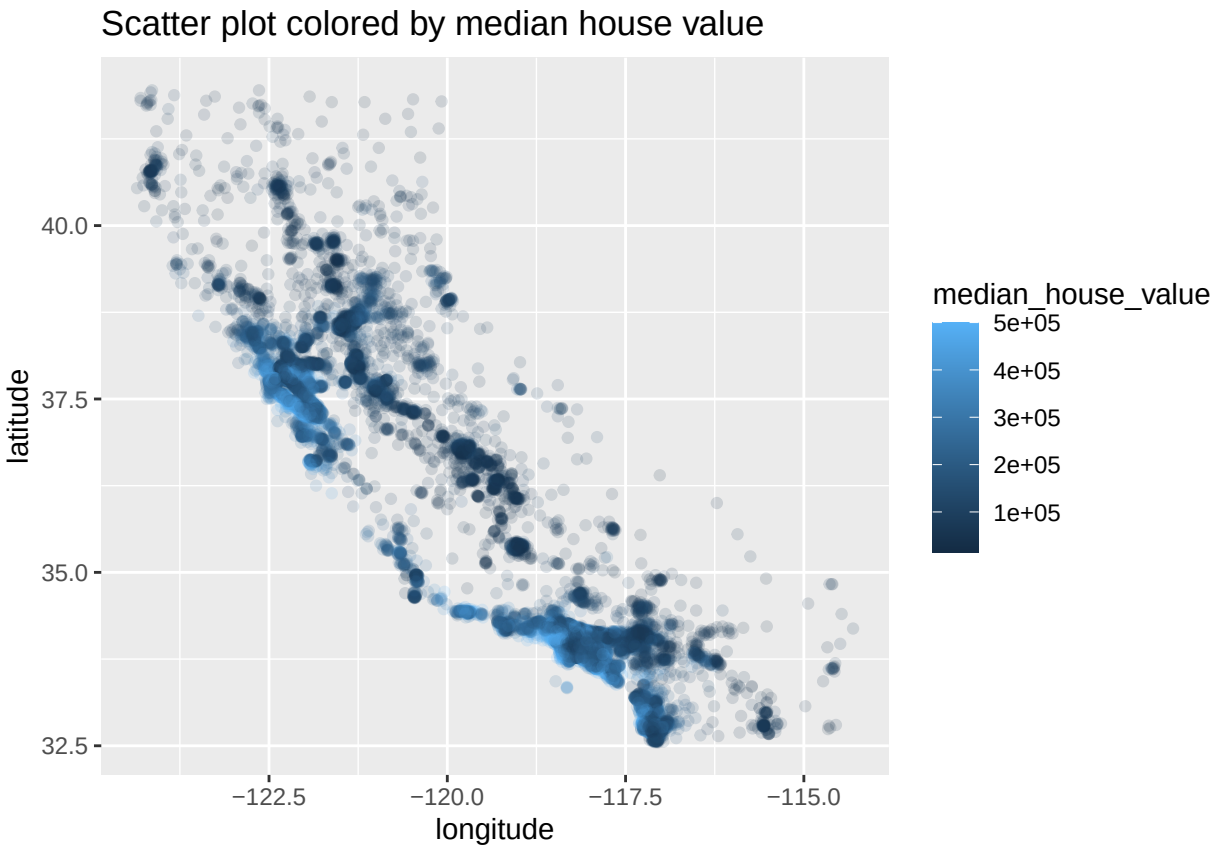
```
 train %>%
ggplot() +
  geom_point(mapping= aes(x= longitude, y= latitude),
             alpha = 0.15) +
  labs(title="the location of the houses", x="longitude", y="latitude")
```



- There are dots, which indicate house merged in northern coastal region and southern coastal region. More people live along the coast. The representative cities are San Francisco and Los Angles.

Exercise 2

```
 train %>%
ggplot() +
  geom_point(mapping= aes(x= longitude, y= latitude, color = median_house_value),
             alpha = 0.15) +
  labs(title="Scatter plot colored by median house value", x="longitude", y="latitude")
```

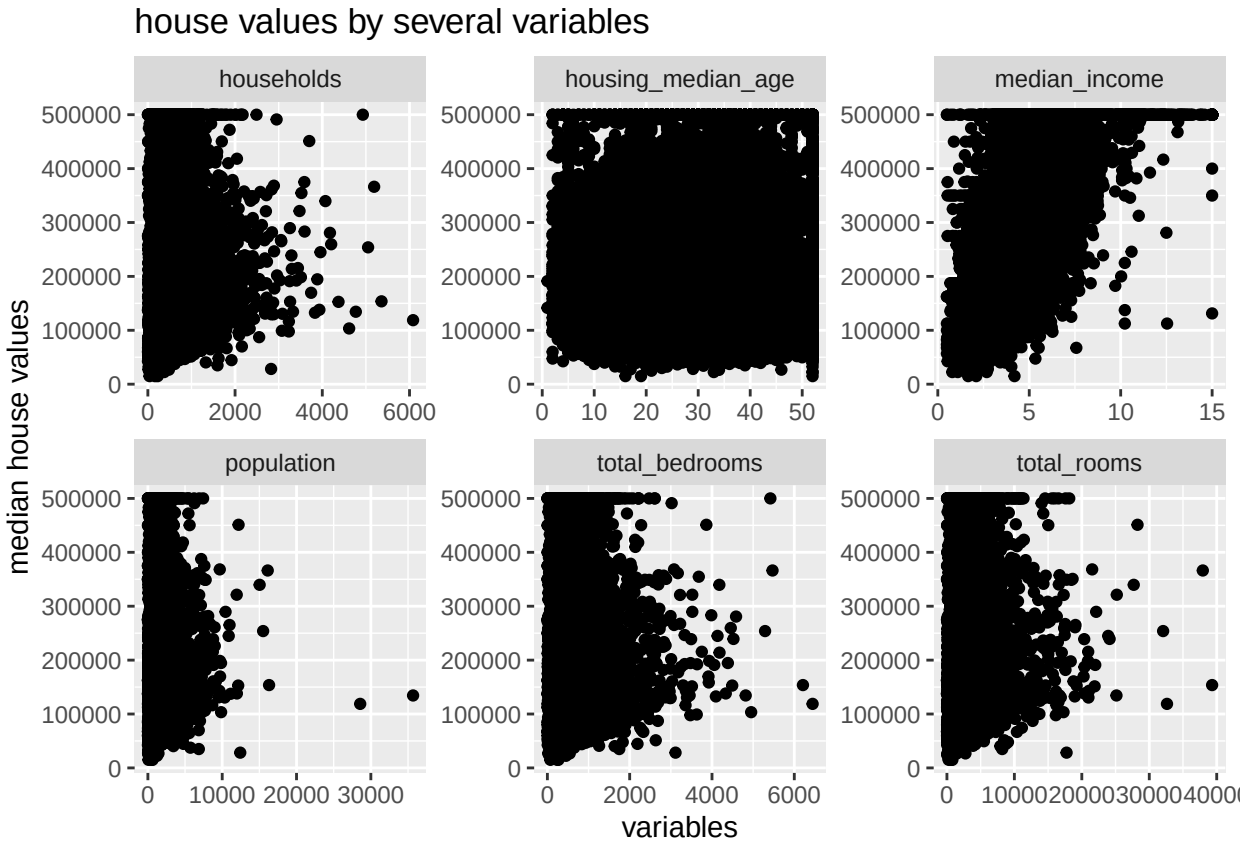


- The milder color means more expensive median house value. northern coastal region, which silicon valley located, shows the most mild color and merged dots.

Exercise 3

```
options("scipen"=100)
train %>% pivot_longer(cols=housing_median_age:median_income, names_to="name",
                       values_to = "value"
 ) %>%
  ggplot() +
  geom_point(mapping = aes(x= value, y= median_house_value)) +
  facet_wrap(~name, scales="free") +
  labs(title="house values by several variables", x="variables",
       y="median house values")
```

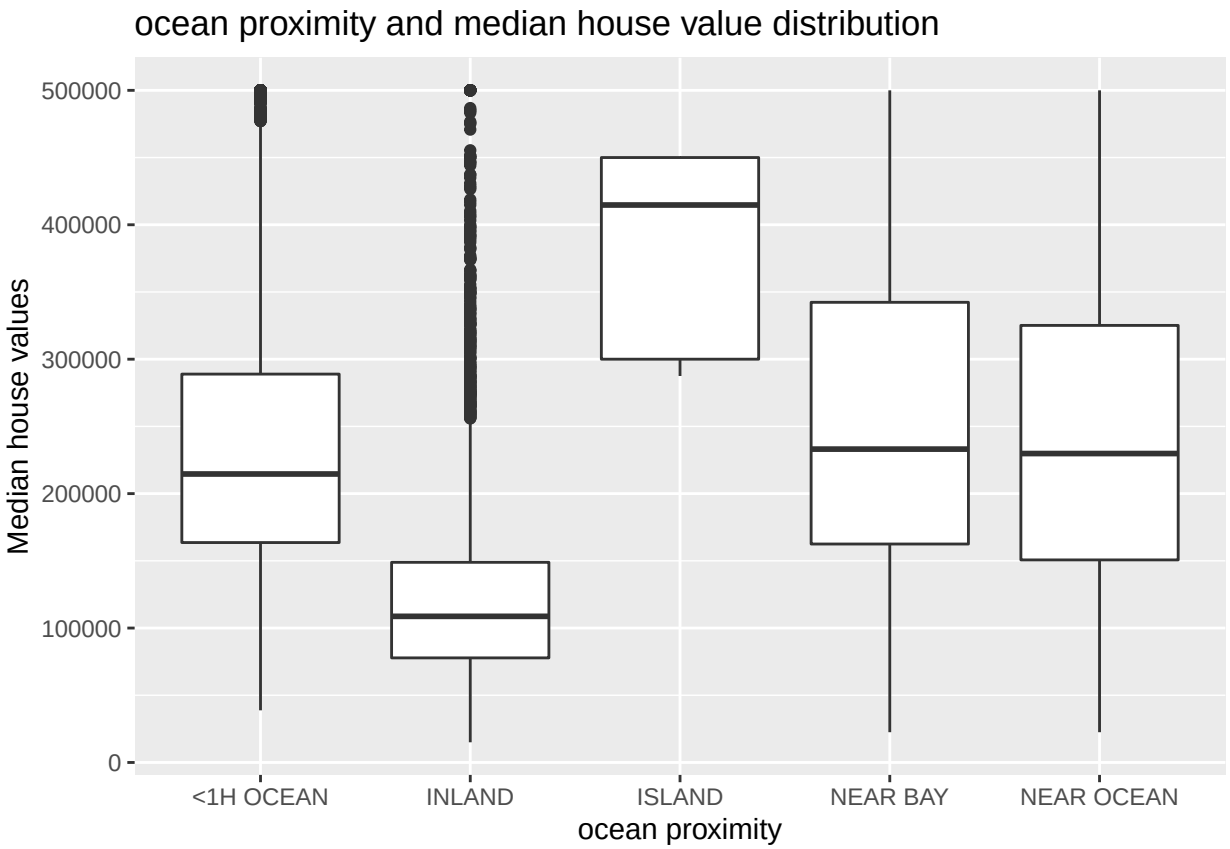
```
## Warning: Removed 164 rows containing missing values (geom_point).
```



- median_income shows positive relationship with median_house_value.
- The response variable goes up to (the celling)500000.
Although the values over the celling can effect on linear model,
however, in this graph we have to underestimate the values over the celling,
even it is still valid data that over 500,000.

Exericse 4

```
train %>%
ggplot() +
  geom_boxplot(mapping = aes(x= ocean_proximity, y = median_house_value)) +
  labs(title= "ocean proximity and median house value distribution",
       x= "ocean proximity",
       y= "Median house values")
```



- The category names "INLAND" has low distributed at median house price. This tells us weather it located near by coast or not, it effect on the house price.

Exericse 5

```
model_1 <- lm(median_house_value ~ median_income, data= train, y= TRUE, x= TRUE)
 cv.lm(model_1, k =5)
```

```
## Mean absolute error        :  62569.56
## Sample standard deviation  :  669.1511
##
## Mean squared error         :  6990000512
## Sample standard deviation  :  206744858
##
## Root mean squared error    :  83598.8
## Sample standard deviation  :  1245.984
```

```
 - RMSE for this model is 84592.31.(first try)
- RMSE is the standard deviation of the residuals.
 It indicates the degree of differences between real data and model's prediction.
 The low number of RMSE means there is less difference between models' expectations
 and real observations in data, which means less errors(residuals).
```

Exericse 6

```
model_2 <- lm(median_house_value ~ ., data= train, y= TRUE, x= TRUE)
 cv.lm(model_2, k =5)
```

```
## Mean absolute error        :  49713.19
## Sample standard deviation  :  962.9679
##
## Mean squared error         :  4719530499
## Sample standard deviation  :  330789928
##
## Root mean squared error    :  68665.29
## Sample standard deviation  :  2400.124
```

```
- Compare RMSEs in model 1 and 2, model_2 have lower RMSE, so it fit more accurately.
This means model_2 performs better.
We cans consider model_2 dealt with various explanatory variables,
so it could approached nearer to the actual observations.
```

Exericse 7

```
 rmse(model_2, test)
```

```
## [1] 69274.46
```

- RMSE is 69274.46(first try), it shows model_2 for the test data set did worse than train data set.
- Compare to the cross-validations for two data set, when we apply test data, the out put was worse than trained case. This caused by our model is trained by train data sets. In terms of model, as test data is whole-new data, we can expect our model will be more fit to the train data that we feed to the model_2.