Introduction

Firstly, this data set is introduced. This data set comes from the house price data set of UCI machine learning knowledge base. Boston house began to count these data in 1978, with 506 observations of 13 input variables and 1 output variable. Each data contains detailed information about the house and its surroundings, including urban crime rate, nitric oxide concentration, average number of rooms in the house, weighted distance to the central area, and average house price of the house, etc.

The following is the meaning of each feature.

1.Crim: per capita crime rate in cities and towns. 2. Zn: proportion of residential land over 25000 sq.ft. 3. Industries: the proportion of Urban Non retail commercial land. 4. Chas: Charles River Air variable (1 if the boundary is a river; otherwise 0). 5. NOx: concentration of nitric oxide. 6. RM: average number of rooms in the house. Age: the proportion of self use houses built before 1940. 8. Dis: weighted distance to Boston's five central areas. 9. Rad: approach index of radial road. 10. Tax: full value property tax rate per 10000 US dollars. 11. Ptratio: proportion of teachers and students in cities and towns. 12. B: 1000 (bk-0.63) ^ 2, where BK refers to the proportion of black people in cities and towns. 13. LSTAT: the proportion of people with lower status in the population. 14. MEDV: average house price of self housing, in thousands of US dollars.

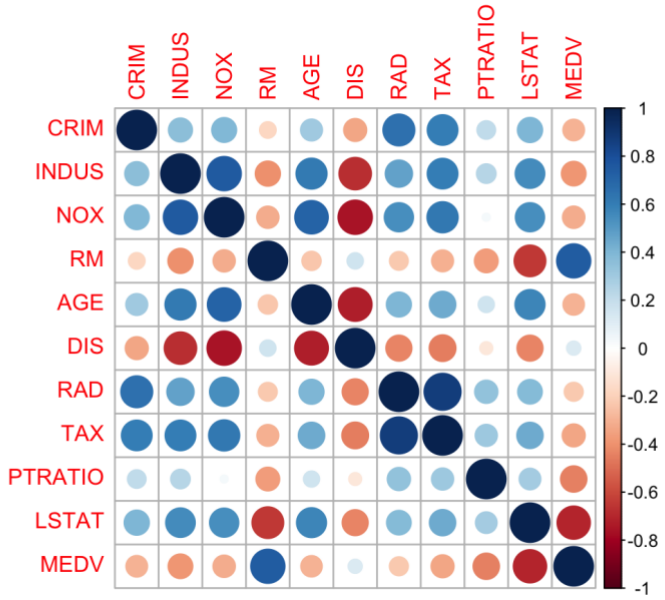
Our purpose is to explore the relationship between variables and housing prices, carry out descriptive statistics and related hypothesis tests, and finally establish a regression model and predict housing prices.

Source: https://archive.ics.uci.edu/ml/machine-learning-databases/housing/housing.data

Data Description

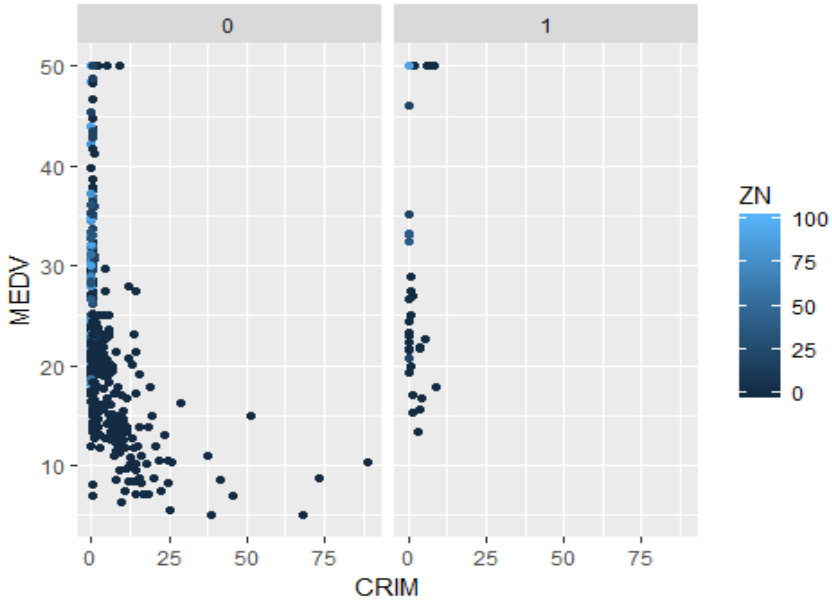
First observe the correlation coefficient between the variables, as shown in the figure below. The correlation coefficient can well reflect the linear relationship between each variable and MEDV, and our purpose is to mine the variables with the highest correlation. As shown in the figure, the most relevant are LSTAT and RM. That is, the more people in low status, the less the average room number, the higher the average house price. This is a surprising finding, because generally speaking, the higher status people live in the more expensive places, while the data shows the opposite. So it's not hard to think that people with low status may have higher requirements for living environment and choose more expensive houses.

```
data = read.csv('house_data.csv')
corrplot::corrplot(cor(data))
```



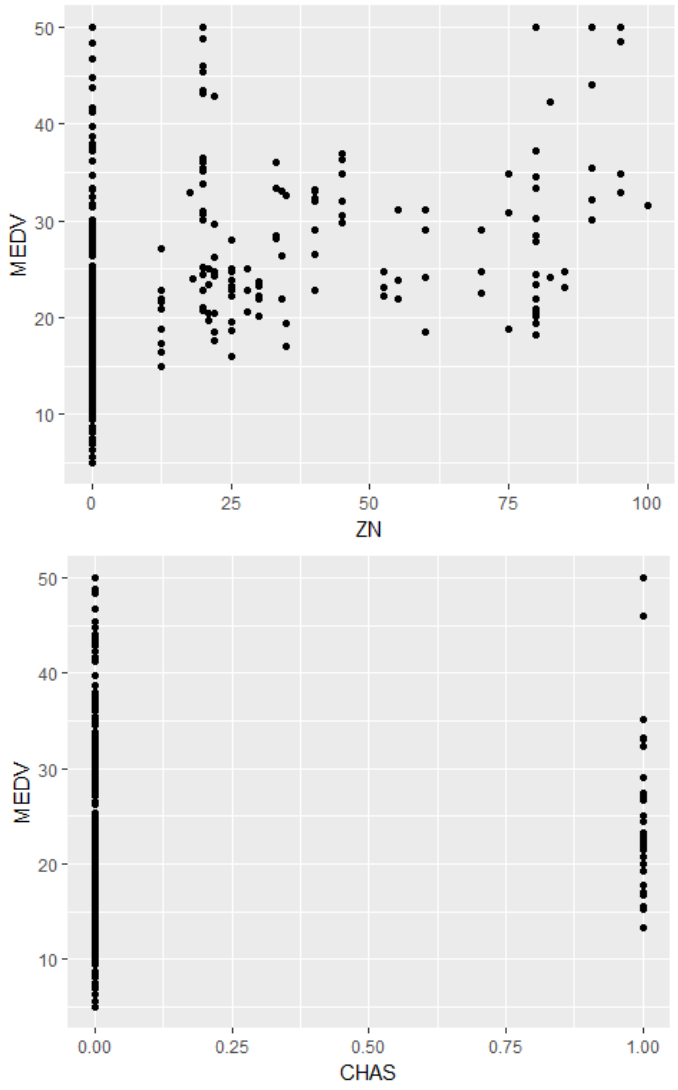
Then we find that for Crim and Zn, they are mainly concentrated in smaller values. Because crim is the urban crime rate, we know that the crime rate is generally low, fluctuating in a small value, and the larger value should affect the house price. We select crim and Zn as two variables to observe the change of house price on the basis of Chas. First of all, we can be sure that the distribution of data will not be caused by abnormal values, because for larger data, such as larger crime rate, the lower house price is reflected in the lower house price, and for the house with Chas 1, the house price distribution is also regular. However, for these biased distributions, the total number of feature values is less than 10% of the sample number, we should consider whether the feature has discrimination.

ggplot(data, aes(x=CRIM,y=MEDV))+geom_point(aes(colour=ZN))+facet_wrap(~CHAS)



We can see from the figure below that for different values of Zn and Chas, the distribution of house prices is not significantly different. Therefore, we can think that these two characteristics are lack of differentiation and should be deleted.

```
ggplot(data, aes(x=ZN, y=MEDV))+geom_point()
ggplot(data, aes(x=CHAS, y=MEDV))+geom_point()
```



Linear Regression Analysis

Next, establish a linear regression model. Based on the previous discussion, the correlation variables are eliminated and the regression model is established for the remaining variables to explore the more accurate relationship between each variable and MEDV. The result is as follow.

```
data.train = data %>% slice(1:(0.8*n()))
data.test = data %>% slice((0.8*n()):n())
x.train = data.train[,c('RM','LSTAT','PTRATIO','DIS','CRIM','RAD','NOX','TAX', 'MEDV')]
x.test = data.test[,c('RM','LSTAT','PTRATIO','DIS','CRIM','RAD','NOX','TAX')]
reg = lm(MEDV~.-1, x.train)
stargazer::stargazer(reg)
```

	Dependent variable:
	MEDV
RM	6.871***
	(0.305)
LSTAT	-0.445***
	(0.055)
PTRATIO	-0.510***
	(0.103)
DIS	-0.579***
	(0.154)
CRIM	-0.225***
	(0.055)
RAD	0.354***
	(0.085)
NOX	-2.034
	(3.483)
TAX	-0.010**
	(0.004)
Observations	404
R_2	0.962
Adjusted R_2	0.962
Residual Std. Error	5.076 (df = 396)
F Statistic	1,263.694*** (df = 8; 396)
Note:	*p<0.1; **p<0.05; ***p<0.01

The coefficients of each variable are shown above, and the size of the coefficients reflects the degree of influence on MEDV. It is not difficult to find that RM has the largest coefficient and the biggest influence, which shows that the house price is mainly affected by the number of rooms. Next is NOx, the concentration of nitric oxide also affects the house price to some extent, and the smaller the concentration, the higher the house price, which also conforms to common sense. Tax is the least influential variable, and the property tax rate has little influence on the average house price.

Next is the t test of the coefficient. The original assumption is that the coefficient is 0 and the optional assumption is that the coefficient is not 0. of all the above variables, only NOX is not significant, even if the coefficient is large, which indicates that its impact on MEDV is due to the large absolute value, rather than the real significant impact. On the contrary, although the tax coefficient is small, it is significant. Finally, the F-test of coefficients is also very significant, and the model fits well.

Acknowledgments

I would like to thank several authors in reference for their books that have given me a better understanding of statistical theory. After reading these books, I have mastered the basic data analysis ability, so as to successfully analyze the data set completely.

Reference

1. Robert Kabacoff, R in Action, Manning Publications
2. Trevor Hastie / Robert Tibshirani / Jerome Friedman, The Elements of Statistical Learning, Springer