Boston Housing Price Prediction Project

1. Research Scenario Description

Investing in real estate has always been a popular topic in the market.

Nowadays, more and more people are considering living in Boston, and even many overseas investors are also attracted to invest in real estate in Boston.

In this project, I would download the relevant dataset (Boston Housing Price) and use R programming language to establish the house price prediction model through multiple linear regression. By observing the outputs, we would accurately predict what attributes and how they affect the housing prices in Boston.

Based on the Boston Housing Price dataset, 13 factors affect the housing prices: CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT, MEDV. Housing prices are positively correlated with some influencing factors, some are negatively correlated, and the relevant procedures are also different. The housing price is Y (response variable), and the 13 influencing factors are X (explanatory variables). Therefore, we can construct a definite regression equation by the meaning of multiple linear regression.

2. Describe the data set

1. Dataset:

The dataset contains information about housing prices collected by the U.S. Census Bureau in Boston, Massachusetts. It consists of 506 observations and 13 independent variables, and one dependent variable in each class. Thus, the dataset has the following 14 attributes.

2. Each columns of the dataset:

	Column_name	**Column Description**
1	CRIM	per capita crime rate by town
2	ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
3	INDUS	proportion of non-retail business acres per town
4	CHAS	Charles River dummy variable(=1 if tract bounds river；0 otherwise)
5	NOX	nitric oxides concentration (parts per 10 million)
6	RM	average number of rooms per dwelling

7	AGE	proportion of owner-occupied units built prior to 1940
8	DIS	weighted distances to five Boston employment centres
9	RAD	index of accessibility to radial highways
10	TAX	full-value property-tax rate per $10,000
11	PTRATIO	pupil-teacher ratio by town
12	B	1000(Bk-0.63)^2where Bk is the proportion of blacks by town
13	LSTAT	lower status of the population
14	MEDV	Median value of owner-occupied homes in $1000's

3.Link to the main data set source:

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

3. Research Question

The main research question:

- Does the proportion of low-income landlords have an impact on housing prices?
- What are the main attributes affecting the housing prices?

4. The solution R code

```
## Step1: Load libraries
library(MASS);library(car);library(corrplot)
library(funr);library(openxlsx);library(dplyr)
library(caret);library(psych);library(plyr)
library(ggplot2);library(zoo);library(lmtest)
library(graphics))

## Step 2: Import data
```

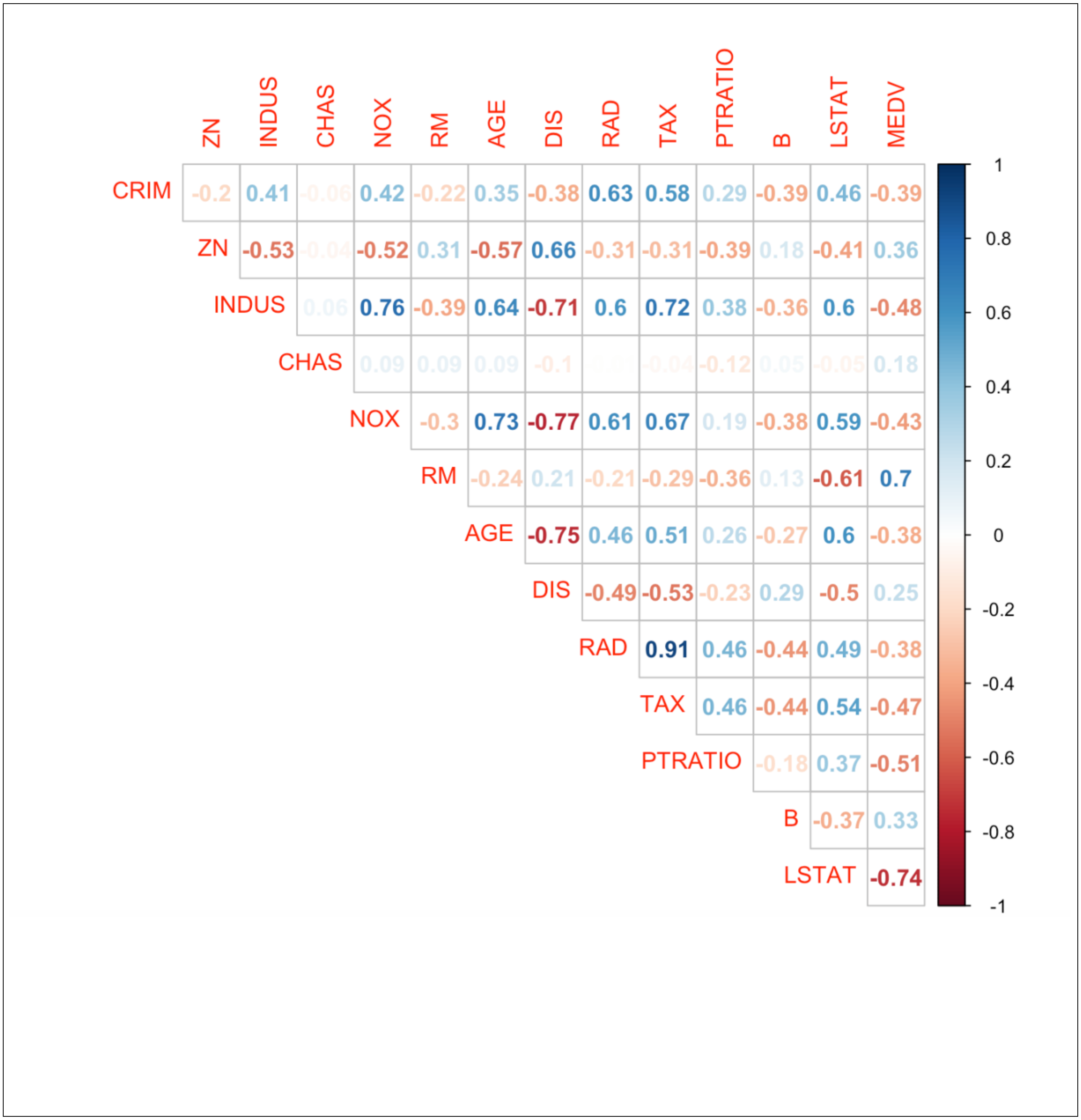
```
df <- read.csv("housingdata.csv",header = T)
train <- df[1:400,]
test <- df[-(1:400),]
View(df)
sum(is.na(df))
write.csv(df,"/Users/zhangluyu/Desktop/CS555_Term Project/housingdata.csv", row.names = FALSE)

## Step 3: Checking correlation between variables
cor<-
cor(df$MEDV,df[c("CRIM","ZN","INDUS","CHAS","NOX","RM","AGE","DIS","RAD","TAX","PTRATIO","B","L
STAT","MEDV")])
corrplot(cor(df), method="number", type = "upper", diag = FALSE)

## Step 4: Building Linear Regression Model
lm1<-lm(MEDV~CRIM+ZN+INDUS+CHAS+NOX+RM+AGE+DIS+RAD+TAX+PTRATIO+B+LSTAT,data=df)
summary(lm1)

## Constructing a new linear model by analyzing the correlation matrix
lm2<-lm(MEDV~INDUS+RM+NOX+TAX+PTRATIO+LSTAT,data=df)
summary(lm2)

##Model Diagnostics for lm2- to check the operation of the model
layout(matrix(c(1,2,3,4),2,2))
plot(lm2)

## Step 5: Model testing - to analyze the variance of two models(verification)
anova(lm1,lm2)

## Step 6: Model fitting
train.pred <- predict(lm2,se.fit=TRUE)
par(mfrow=c(1,1))
plot(train$MEDV,col="lightblue",pch=15,xlab=expression("num"),ylab="MEDV",
     main="Fitting results of MEDV")
lines(train.pred$fit,col="blue")

##Model1
par(mfrow=c(1,1))
test.pred <- predict(lm1,newdata= test,se.fit=TRUE)
plot(test$MEDV,col="lightblue",pch=15,xlab=expression("num"),ylab="MEDV",
     main="Model1:Predicting results of MEDV")
lines(test.pred$fit,col="blue")
estimateError1 <- (test$MEDV-test.pred$fit)
plot(estimateError1,main="Model1:Predicting errors of MEDV")

##Model2
test.pred <- predict(lm2,newdata= test,se.fit=TRUE)
par(mfrow=c(1,1))
plot(test$MEDV,col="lightblue",pch=15,xlab=expression("num"),ylab="MEDV",
     main="Model2:Predicting results of MEDV")
lines(test.pred$fit,col="blue")
estimateError2 <- (test$MEDV-test.pred$fit)
```

```
plot(estimateError2,main="Model2:Predicting errors of MEDV")
```

5. Execute results.

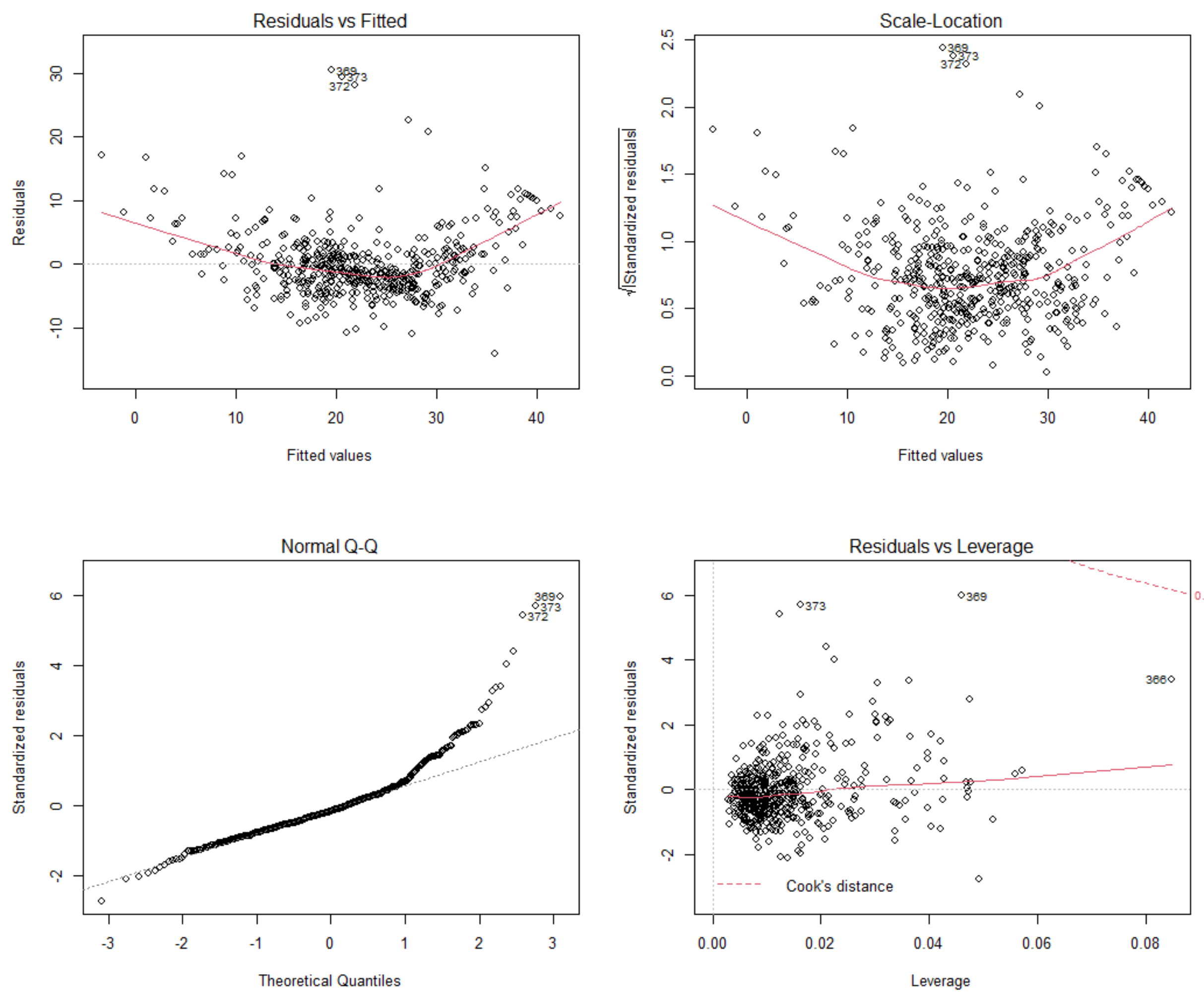
	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
1	0.00632	18.0	2.31	0	0.5380	6.575	65.2	4.0900	1	296	15.3	396.90	4.98	24.0
2	0.02731	0.0	7.07	0	0.4690	6.421	78.9	4.9671	2	242	17.8	396.90	9.14	21.6
3	0.02729	0.0	7.07	0	0.4690	7.185	61.1	4.9671	2	242	17.8	392.83	4.03	34.7
4	0.03237	0.0	2.18	0	0.4580	6.998	45.8	6.0622	3	222	18.7	394.63	2.94	33.4
5	0.06905	0.0	2.18	0	0.4580	7.147	54.2	6.0622	3	222	18.7	396.90	5.33	36.2
6	0.02985	0.0	2.18	0	0.4580	6.430	58.7	6.0622	3	222	18.7	394.12	5.21	28.7
7	0.08829	12.5	7.87	0	0.5240	6.012	66.6	5.5605	5	311	15.2	395.60	12.43	22.9
8	0.14455	12.5	7.87	0	0.5240	6.172	96.1	5.9505	5	311	15.2	396.90	19.15	27.1
9	0.21124	12.5	7.87	0	0.5240	5.631	100.0	6.0821	5	311	15.2	386.63	29.93	16.5
10	0.17004	12.5	7.87	0	0.5240	6.004	85.9	6.5921	5	311	15.2	386.71	17.10	18.9
11	0.22489	12.5	7.87	0	0.5240	6.377	94.3	6.3467	5	311	15.2	392.52	20.45	15.0
12	0.11747	12.5	7.87	0	0.5240	6.009	82.9	6.2267	5	311	15.2	396.90	13.27	18.9
13	0.09378	12.5	7.87	0	0.5240	5.889	39.0	5.4509	5	311	15.2	390.50	15.71	21.7
14	0.62976	0.0	8.14	0	0.5380	5.949	61.8	4.7075	4	307	21.0	396.90	8.26	20.4
15	0.63796	0.0	8.14	0	0.5380	6.096	84.5	4.4619	4	307	21.0	380.02	10.26	18.2
16	0.62739	0.0	8.14	0	0.5380	5.834	56.5	4.4986	4	307	21.0	395.62	8.47	19.9
17	1.05393	0.0	8.14	0	0.5380	5.935	29.3	4.4986	4	307	21.0	386.85	6.58	23.1
18	0.78420	0.0	8.14	0	0.5380	5.990	81.7	4.2579	4	307	21.0	386.75	14.67	17.5
19	0.80271	0.0	8.14	0	0.5380	5.456	36.6	3.7965	4	307	21.0	288.99	11.69	20.2
20	0.72580	0.0	8.14	0	0.5380	5.727	69.5	3.7965	4	307	21.0	390.95	11.28	18.2
21	1.25179	0.0	8.14	0	0.5380	5.570	98.1	3.7979	4	307	21.0	376.57	21.02	13.6
22	0.85204	0.0	8.14	0	0.5380	5.965	89.2	4.0123	4	307	21.0	392.53	13.83	19.6
23	1.23247	0.0	8.14	0	0.5380	6.142	91.7	3.9769	4	307	21.0	396.90	18.72	15.2
24	0.98843	0.0	8.14	0	0.5380	5.813	100.0	4.0952	4	307	21.0	394.54	19.88	14.5
25	0.75026	0.0	8.14	0	0.5380	5.924	94.1	4.3996	4	307	21.0	394.33	16.30	15.6
26	0.84054	0.0	8.14	0	0.5380	5.599	85.7	4.4546	4	307	21.0	303.42	16.51	13.9
27	0.67191	0.0	8.14	0	0.5380	5.813	90.3	4.6820	4	307	21.0	376.88	14.81	16.6

Showing 1 to 27 of 506 entries, 14 total columns

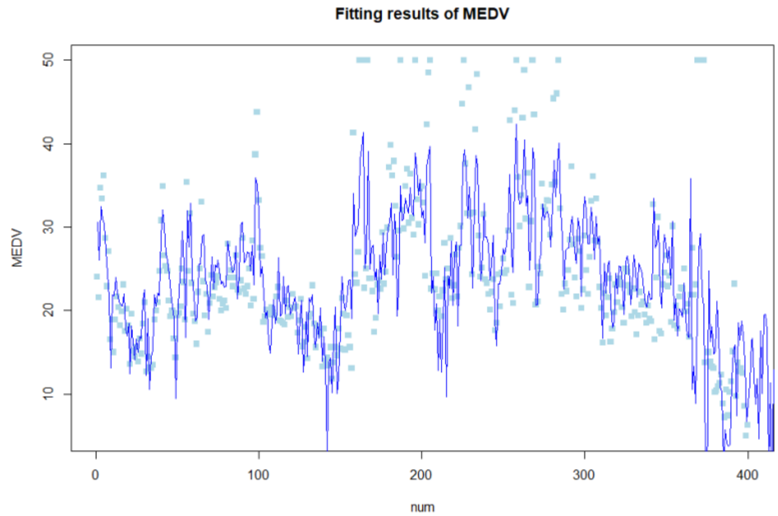

ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO B LSTAT MEDV
CRIM -0.2 0.41 -0.06 0.42 -0.22 0.35 -0.38 0.63 0.58 0.29 -0.39 0.46 -0.39
ZN -0.53 -0.04 -0.52 0.31 -0.57 0.66 -0.31 -0.31 -0.39 0.18 -0.41 0.36
INDUS 0.06 0.76 -0.39 0.64 -0.71 0.6 0.72 0.38 -0.36 0.6 -0.48
CHAS 0.09 0.09 0.09 -0.1 -0.04 -0.12 0.05 -0.05 0.18
NOX -0.3 0.73 -0.77 0.61 0.67 0.19 -0.38 0.59 -0.43
RM -0.24 0.21 -0.21 -0.29 -0.36 0.13 -0.61 0.7
AGE -0.75 0.46 0.51 0.26 -0.27 0.6 -0.38
DIS -0.49 -0.53 -0.23 0.29 -0.5 0.25
RAD 0.91 0.46 -0.44 0.49 -0.38
TAX 0.46 -0.44 0.54 -0.47
PTRATIO -0.18 0.37 -0.51
B -0.37 0.33
LSTAT -0.74
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

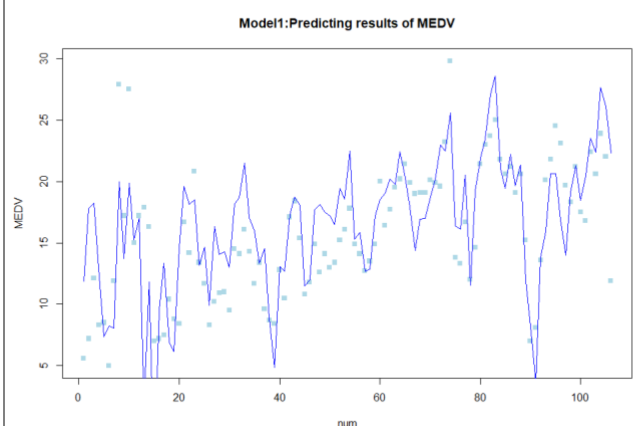
```
Call:
lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
    DIS + RAD + TAX + PTRATIO + B + LSTAT, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-15.595  -2.730  -0.518   1.777  26.199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
CRIM        -1.080e-01  3.286e-02  -3.287 0.001087 **
ZN           4.642e-02  1.373e-02   3.382 0.000778 ***
INDUS        2.056e-02  6.150e-02   0.334 0.738288
CHAS         2.687e+00  8.616e-01   3.118 0.001925 **
NOX         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
RM           3.810e+00  4.179e-01   9.116  < 2e-16 ***
AGE          6.922e-04  1.321e-02   0.052 0.958229
DIS         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
RAD          3.060e-01  6.635e-02   4.613 5.07e-06 ***
TAX         -1.233e-02  3.760e-03  -3.280 0.001112 **
PTRATIO     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B            9.312e-03  2.686e-03   3.467 0.000573 ***
LSTAT       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.745 on 492 degrees of freedom
Multiple R-squared:  0.7406,	Adjusted R-squared:  0.7338
F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

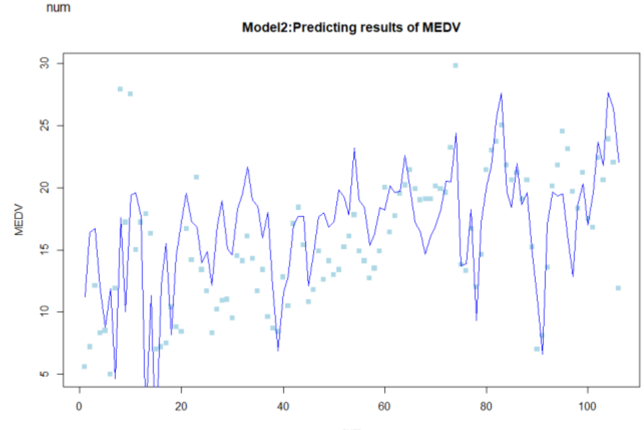
```
Call:
lm(formula = MEDV ~ INDUS + RM + NOX + TAX + PTRATIO + LSTAT,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-13.9802  -3.0470  -0.9347   1.7100  30.4545

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 19.145818   4.309359   4.443 1.09e-05 ***
INDUS        0.087187   0.061080   1.427    0.154
RM           4.655928   0.431815  10.782  < 2e-16 ***
NOX         -3.403117   3.478085  -0.978    0.328
TAX         -0.002901   0.002225  -1.304    0.193
PTRATIO     -0.913819   0.131157  -6.967 1.03e-11 ***
LSTAT       -0.545935   0.050641 -10.780  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.225 on 499 degrees of freedom
Multiple R-squared:  0.681,	Adjusted R-squared:  0.6772
F-statistic: 177.6 on 6 and 499 DF,  p-value: < 2.2e-16
```

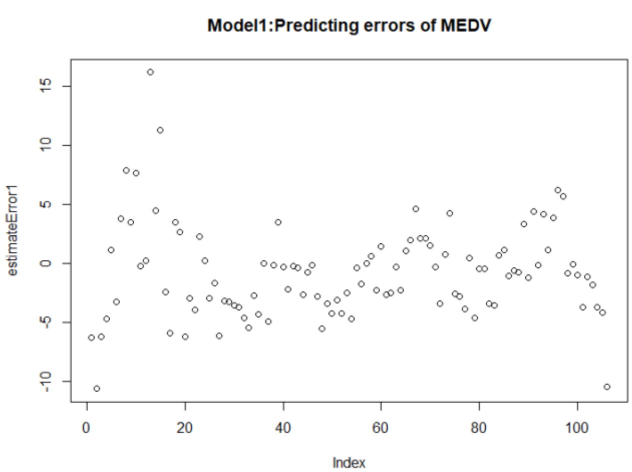


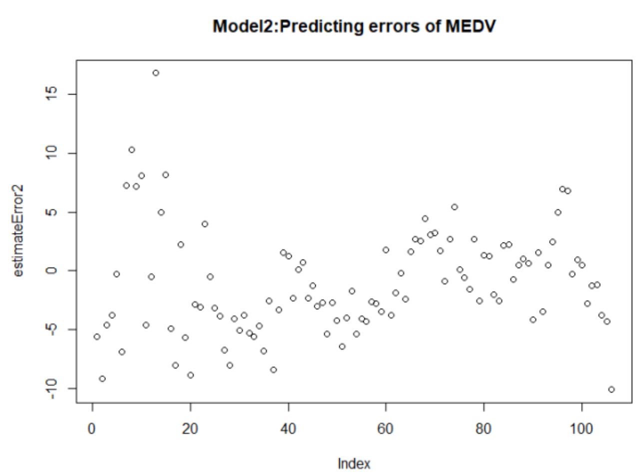
```
Analysis of Variance Table

Model 1: MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
    TAX + PTRATIO + B + LSTAT
Model 2: MEDV ~ INDUS + RM + NOX + TAX + PTRATIO + LSTAT
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    492 11079
2    499 13626 -7   -2546.8 16.157 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```











6. Conclusion

- Boston house prices are mainly affected by these factors, which are "INDUS", "RM", "NOX", "TAX", "PTRATIO","LSTAT". Through correlation analysis, "LSTAT" and "RM" have the highest correlation with housing prices. The results show the proportion of low-income landlords is negatively correlated with housing prices.
- Through the analysis of ANOVA's two models, we conclude model2 is more meaningful, and the fitting effect is more obvious than model1.Thus, we can say " LSTAT " has a significant influence on the housing prices.
- The fitting degree and prediction accuracy of the model are closely related to the selection of attributes.