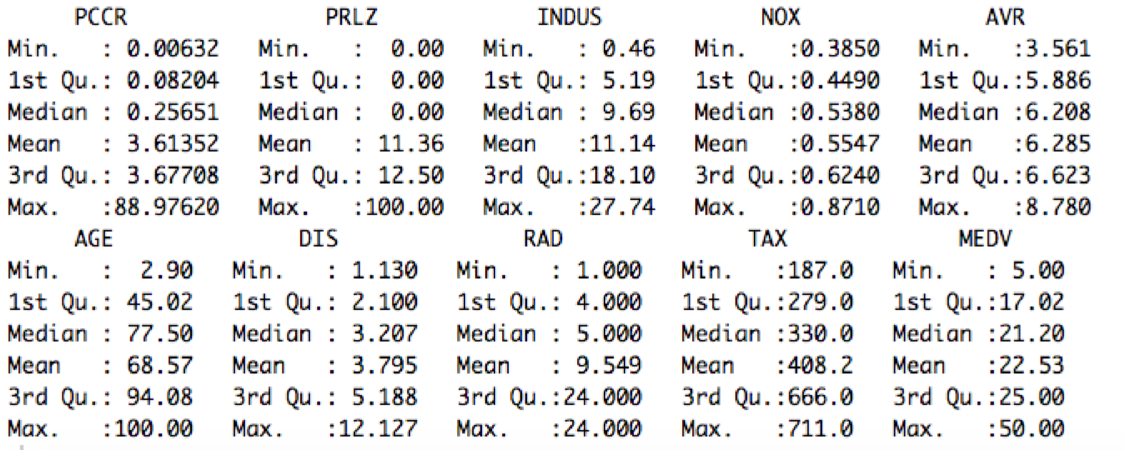
## Regression

1. *Load and study the dataset and comment on the dataset's characteristics / structure. Do not scale your data for this part.*
    - The dataset has 506 observations of 10 variables in integer and numeric form.
    - There is no NA/null/duplicate value in the dataset.
2. *Conduct exploratory data analysis for all variables to get a better understanding about each of them (numerically and visually). Comment on the results and findings.*

The summary of all 10 variables in the housing data was conducted in order to understand about each of them.

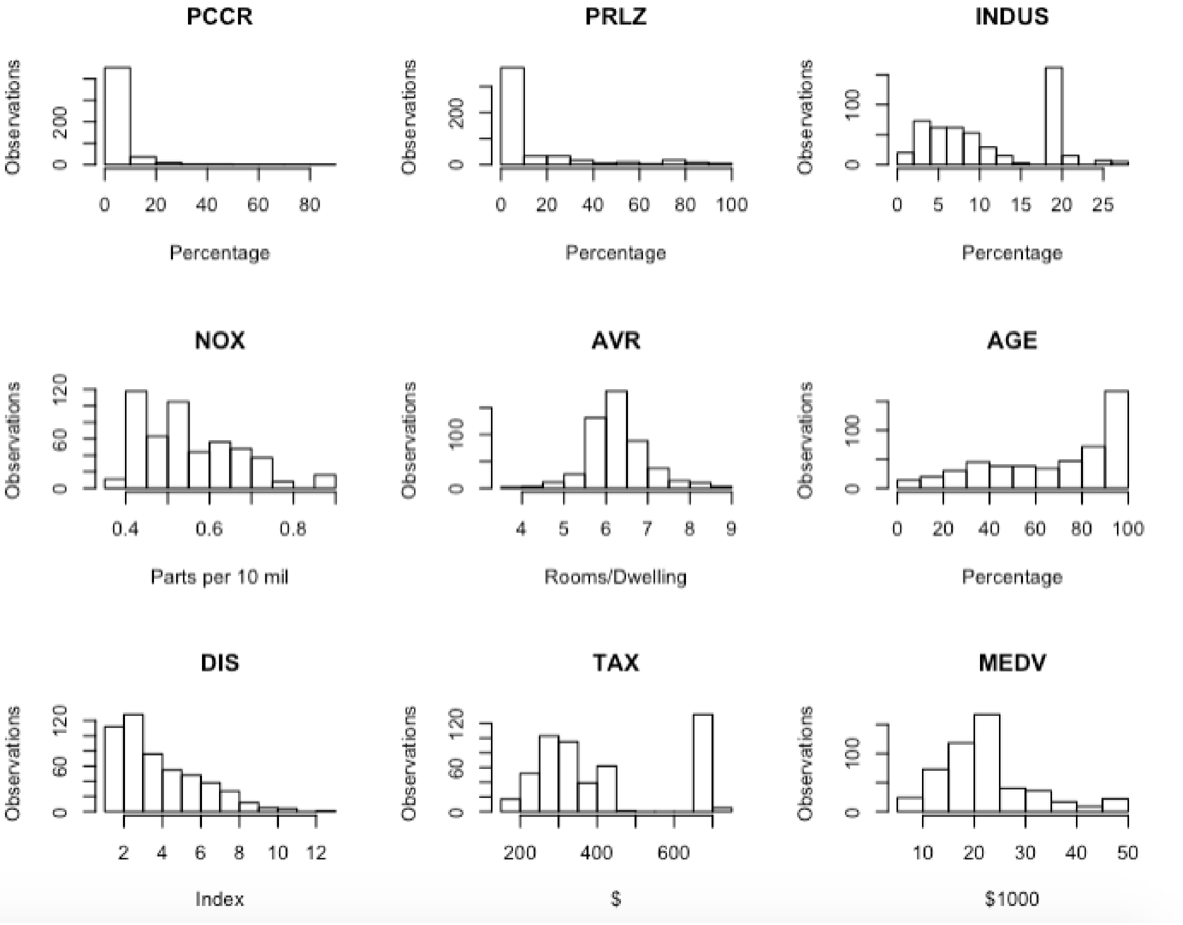


```
      PCCR              PRLZ            INDUS             NOX              AVR
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.3850   Min.   :3.561
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.4490   1st Qu.:5.886
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.5380   Median :6.208
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.5547   Mean   :6.285
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.6240   3rd Qu.:6.623
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :0.8710   Max.   :8.780
      AGE              DIS              RAD              TAX             MEDV
 Min.   :  2.90   Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   : 5.00
 1st Qu.: 45.02   1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.02
 Median : 77.50   Median : 3.207   Median : 5.000   Median :330.0   Median :21.20
 Mean   : 68.57   Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :22.53
 3rd Qu.: 94.08   3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:25.00
 Max.   :100.00   Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :50.00
```

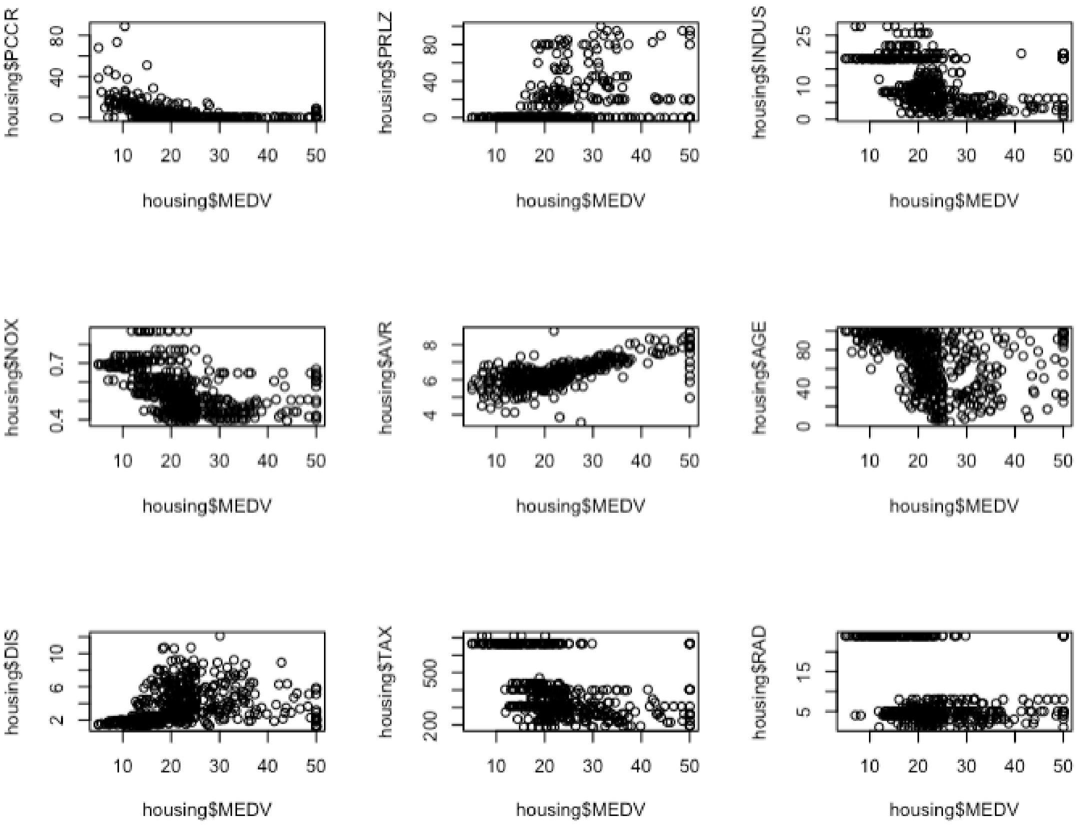
- PCCR:  Average per capita crime rate by town, minimum rate is 0%, maximum rate is 89%,  mean is quite low with only 3.6%. Majority of the data is close to 0%.
- PRLZ: Proportion of residential land zoned for lots over 25,000 sq. ft , minimum rate is 0%, maximum is 100%,  mean is 11.36%. Majority of the data is close to 0%.
- INDUS: Proportion of non-retail business acres per town, minimum rate is 0.5%, maximum is 27.7%, average is 11.14%.
- NOX: Nitric oxide concentration (parts per 10 million) , all value is from 0.4 to 0.9, mean is 0.55
- AVR: Average number of rooms per dwelling, mean is 6.3 rooms, minimum value is 3.6 rooms and maximum value is 8.8 rooms.
- AGE: Proportion of owner-occupied units built prior to 1940, minimum proportion is 2.9% and maximum proportion is 100%, mean is 68.6%,
- DIS: Weighted distances to five Boston employment centers – Range from 1.1 to 12.1, mean distance is 3.78.
- RAD: Index of accessibility to radial highways - Range from 1 to 24, average index is 9.55.

- TAX: Full-value property tax rate per $10,000 - Range from 187 to 711, average value is 330.
- MEDV: Median value of owner-occupied homes in $1000s – Range from 5,000 to 50,000, average value is 21,200.

  - Conduct Histogram of all variables

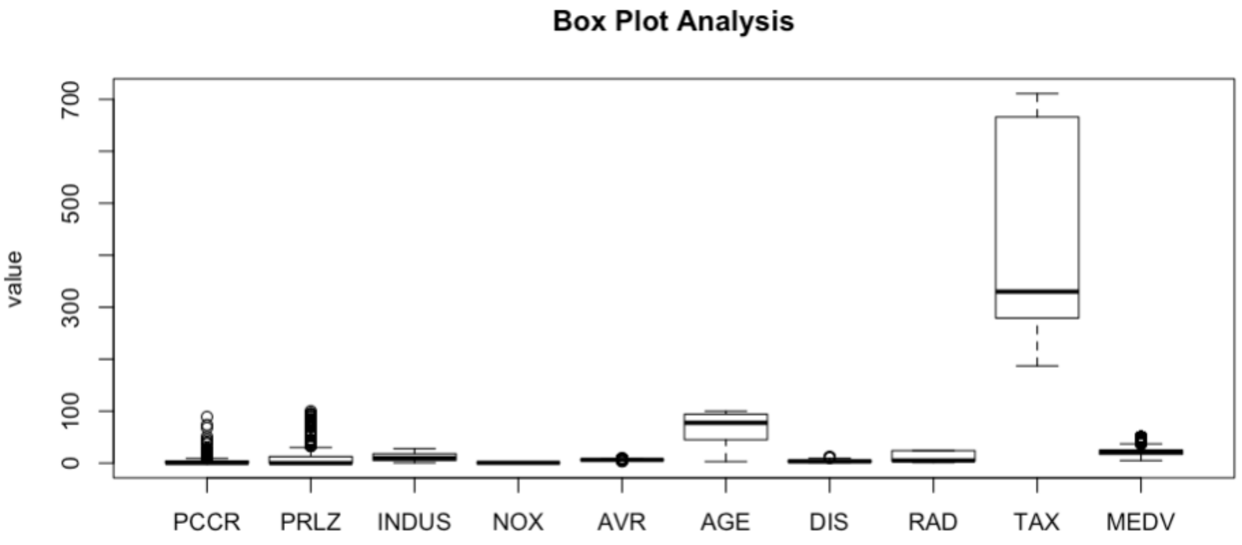


- A closer look to the relationships between housing price (MEDV) and other variables. It can also be seen from the graph that there is some relationship between price and number of rooms per dwelling (AVR) and Nitric Oxide Concentration

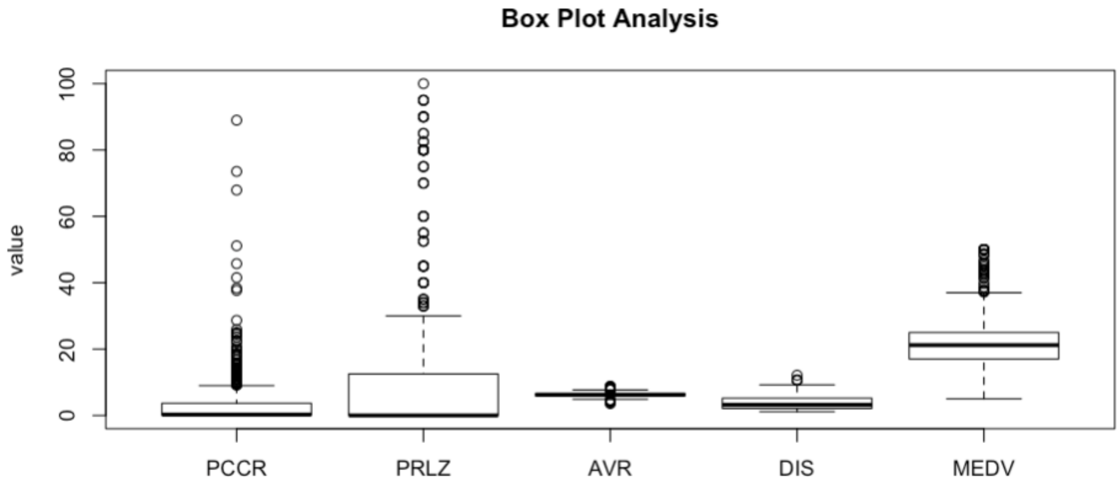


## Box Plot Analysis

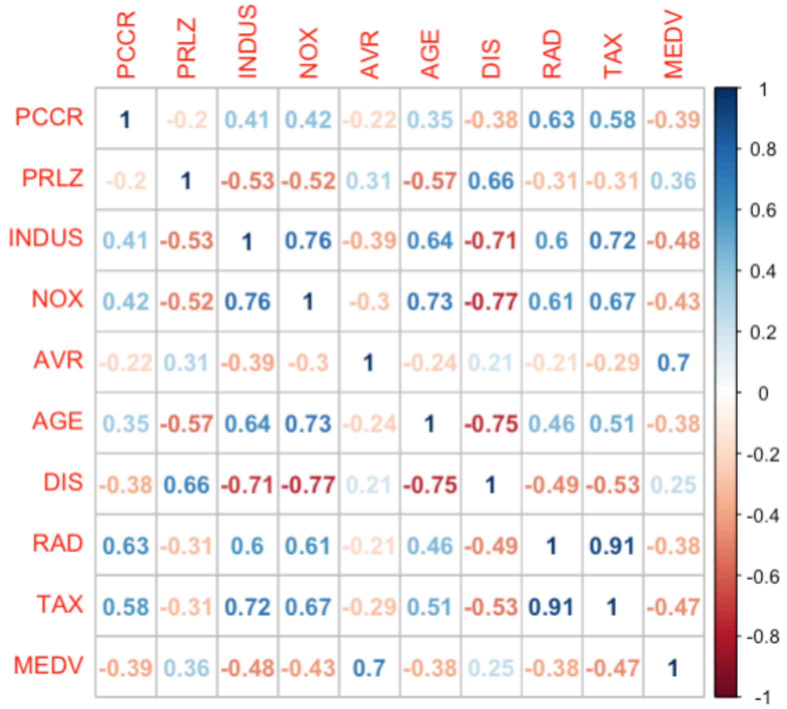
- There are 5 variables seem to have outliers: PCCR, PRLZ, AVR, DIS, MEDV



- Closer look to the variables which have outliers

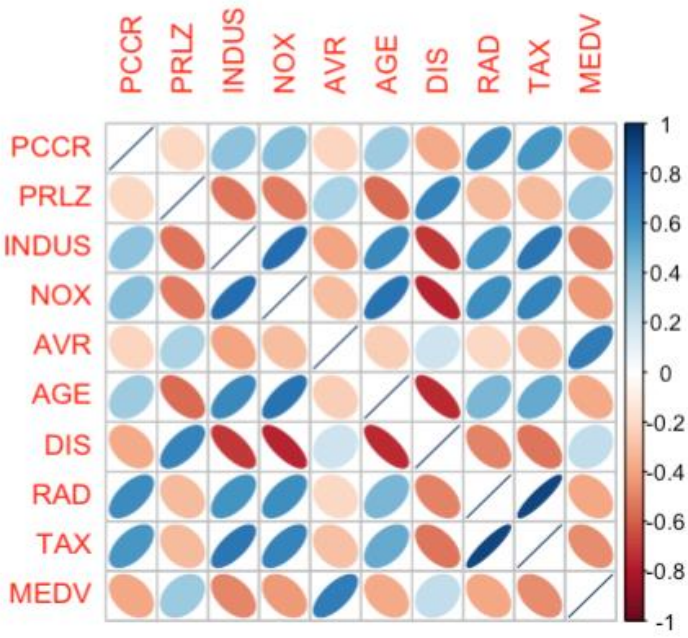


3. *Perform a correlation analysis between all variables. In addition, comment on the variables that have the highest (absolute) linear associations with the prices of the house. Also, comment on your findings.*
    - AVR have the highest linear associations with the prices of the house. The correlation between the two variables is strong and positive which is +0.7.
    - PRLZ and DIS have moderate positive relationship with the prices of the house, 0.36 and 0.25 respectively.
    - Other variables such as PCCR, INDUS, NOX, AGE, RAD, TAX have moderate negative relationship with the price of house.



| | PCCR | PRLZ | INDUS | NOX | AVR | AGE | DIS | RAD | TAX | MEDV |
|---|---|---|---|---|---|---|---|---|---|---|
| PCCR | 1 | -0.2 | 0.41 | 0.42 | -0.22 | 0.35 | -0.38 | 0.63 | 0.58 | -0.39 |
| PRLZ | -0.2 | 1 | -0.53 | -0.52 | 0.31 | -0.57 | 0.66 | -0.31 | -0.31 | 0.36 |
| INDUS | 0.41 | -0.53 | 1 | 0.76 | -0.39 | 0.64 | -0.71 | 0.6 | 0.72 | -0.48 |
| NOX | 0.42 | -0.52 | 0.76 | 1 | -0.3 | 0.73 | -0.77 | 0.61 | 0.67 | -0.43 |
| AVR | -0.22 | 0.31 | -0.39 | -0.3 | 1 | -0.24 | 0.21 | -0.21 | -0.29 | 0.7 |
| AGE | 0.35 | -0.57 | 0.64 | 0.73 | -0.24 | 1 | -0.75 | 0.46 | 0.51 | -0.38 |
| DIS | -0.38 | 0.66 | -0.71 | -0.77 | 0.21 | -0.75 | 1 | -0.49 | -0.53 | 0.25 |
| RAD | 0.63 | -0.31 | 0.6 | 0.61 | -0.21 | 0.46 | -0.49 | 1 | 0.91 | -0.38 |
| TAX | 0.58 | -0.31 | 0.72 | 0.67 | -0.29 | 0.51 | -0.53 | 0.91 | 1 | -0.47 |
| MEDV | -0.39 | 0.36 | -0.48 | -0.43 | 0.7 | -0.38 | 0.25 | -0.38 | -0.47 | 1 |

4. *Visualize the correlation for all variables to identify which pairs of variables are correlated strongly with each other in absolute terms. Comment on the results and findings.*
   - TAX and RAD have the highest absolute relation with each other (+0.91)
   - There are strong negative relationship between variables INDUS, NOX, AGE and DIS (-0.71, -0.77 and -0.75 respectively)



5. *Write down the regression equation for the prices of the house with respect to the explanatory variables (note: use the notation for the coefficient as in the class). Comment on the meaning of each component in the equation.*

MEDV= $\theta_0 + \theta_1 * PCCR + \theta_2 * PRLZ + \theta_3 * INDUS + \theta_4 * NOX + \theta_5 * AVR + \theta_6 * AGE + \theta_7 * DIS + \theta_8 * RAD + \theta_9 * TAX$

**Comment:**

- Intercept: $\theta_0$ – Constant value
- When these variables increase, housing price also increases: PRLZ, AVR, DIS
- When these variables increase, housing price decreases: PCCR, INDUS, NOX, AGE, RAD, TAX

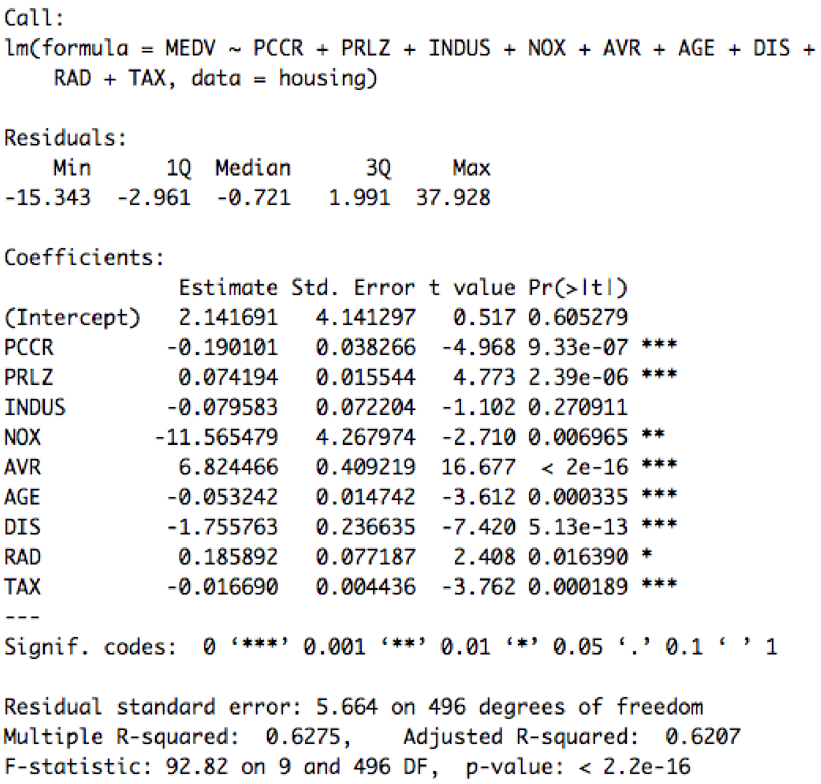
6. *Implement linear regression with all variables and intercept to estimate the coefficients of the model you defined in the previous task to predict the prices of the house. Interpret and comment on the results and findings in detail.*

R formula:

model<-lm(MEDV ~ PCCR + PRLZ + INDUS + NOX + AVR + AGE + DIS + RAD + TAX, housing)

Result interpretation:

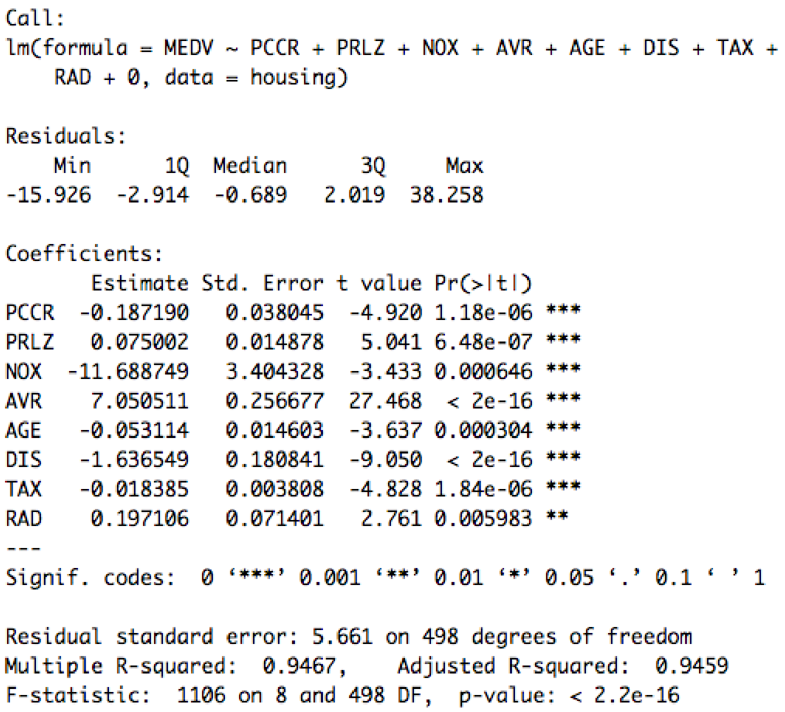
- At first, we set the null hypothesis that the coefficient of all variables to price is equal to 0 (no effect on price). If a variable has low p-value (<0.05), then we can reject the null hypothesis. Based on the model constructed, it can be seen that Intercept and INDUS has p-value larger than 0.05, thus we can consider removing these variables in order to improve the model.
- Multiple R-squared is 0.63 and adjusted R-square is 0.62, thus there are room to improve the model.
- F-statistic is 92.82 with p-value close to 0 indicates that the model with variables is better than the only intercept model.



```
Call:
lm(formula = MEDV ~ PCCR + PRLZ + INDUS + NOX + AVR + AGE + DIS +
    RAD + TAX, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max
-15.343  -2.961  -0.721   1.991  37.928

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.141691   4.141297   0.517 0.605279
PCCR         -0.190101   0.038266  -4.968 9.33e-07 ***
PRLZ          0.074194   0.015544   4.773 2.39e-06 ***
INDUS        -0.079583   0.072204  -1.102 0.270911
NOX         -11.565479   4.267974  -2.710 0.006965 **
AVR           6.824466   0.409219  16.677  < 2e-16 ***
AGE          -0.053242   0.014742  -3.612 0.000335 ***
DIS          -1.755763   0.236635  -7.420 5.13e-13 ***
RAD           0.185892   0.077187   2.408 0.016390 *
TAX          -0.016690   0.004436  -3.762 0.000189 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.664 on 496 degrees of freedom
Multiple R-squared:  0.6275,	Adjusted R-squared:  0.6207
F-statistic: 92.82 on 9 and 496 DF,  p-value: < 2.2e-16
```

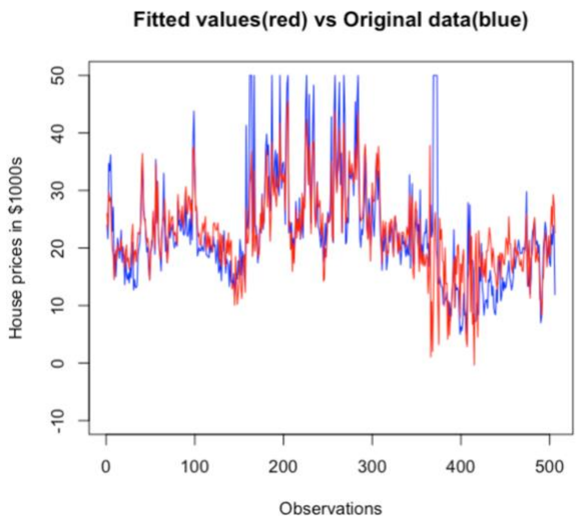
7. *If your analysis of the model shows that this model can be improved, make the necessary adjustments and rerun your model. Discuss in detail how you proceeded and why you proceeded that way as well as your results in each stage. If you consider the model in previous subpart 6 as the optimal model, then explain in detail why it is the optimal model.*

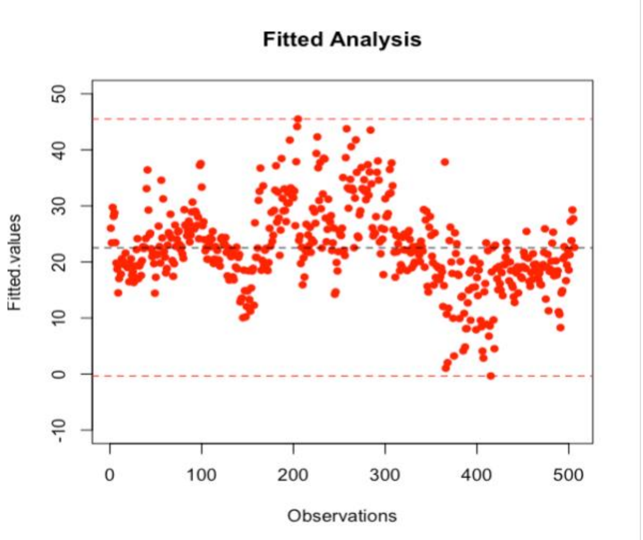
- We can improve the model by removing the variables which have p-value larger than 0.05. Thus, the INDUS and intercept are removed from the model.
- The new model has 8 variables, in which NOX has the largest coefficient. All variables have small p-value (<0.05).
- Multiple R-squared is 0.947 and adjusted R-square is 0.946, significantly improve compare to the original model.
- F-statistic is 1106 with p-value close to 0 indicates that the model with variables is better than the only intercept model.



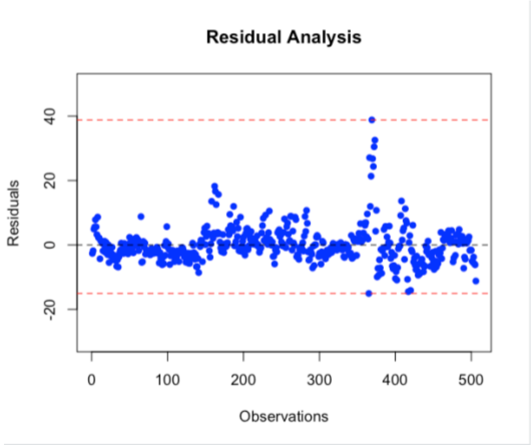
```
Call:
lm(formula = MEDV ~ PCCR + PRLZ + NOX + AVR + AGE + DIS + TAX +
    RAD + 0, data = housing)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.926  -2.914  -0.689   2.019  38.258 

Coefficients:
       Estimate Std. Error t value Pr(>|t|)    
PCCR  -0.187190   0.038045  -4.920 1.18e-06 ***
PRLZ   0.075002   0.014878   5.041 6.48e-07 ***
NOX  -11.688749   3.404328  -3.433 0.000646 ***
AVR    7.050511   0.256677  27.468  < 2e-16 ***
AGE   -0.053114   0.014603  -3.637 0.000304 ***
DIS   -1.636549   0.180841  -9.050  < 2e-16 ***
TAX   -0.018385   0.003808  -4.828 1.84e-06 ***
RAD    0.197106   0.071401   2.761 0.005983 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.661 on 498 degrees of freedom
Multiple R-squared:  0.9467,	Adjusted R-squared:  0.9459 
F-statistic:  1106 on 8 and 498 DF,  p-value: < 2.2e-16
```

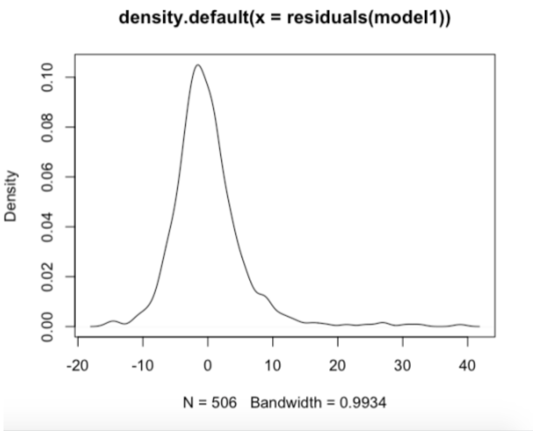
8. *Make a plot of the original vs. the fitted values (based on the final regression model). Comment on the results and findings.*
    - The fitted values (or predicted values) goes in the same direction with original values.
    - Fitted values are distributed equally with the mean, which mean that the model is well constructed.





9. *Do the residuals visually follow the assumption of the normal distribution for residuals in regression analysis? Comment on your findings and reasoning.*
    - The residual is visually followed the assumption of the normal distribution for residuals in regression analysis, except for some outliner cases which could be considered to remove for a better model.
    - Most residual points are 0 and fluctuate from -20 to 40, however very few points above 20 which are outliners.





10. Give a detailed conclusion on your findings and results. Consider and explain how these findings may help the housing agency in the future.

    The regression analysis is constructed to predict the future house prices. It is concluded that house prices are depend on many variables including PCCR, PRLZ, NOX, AVR, AGE, DIS, RAD and TAX. Of those, NOX (Nitric oxide concentration) has the largest coefficient which can negatively influent to the house price. On the other hand, AVR (number of rooms per dwelling) has strongest positive correlation with the house price. The model could be further improved by removing outliners and applying normalization.