Applying regression method in Machine Learning to predict house prices in Boston, US.

Nguyen Tran

August 2020

This project is coded in ***R language*** dealing with large dataset of houses in Boston, U.S. We want to make sense of the dataset from different perspectives and to build a regression model that attempts to explain the prices of the house in that area.

We obtain the "Housing_data" dataset which contains information about different houses in a specific town. Information about the dataset is as following:

- PCCR: Per capita crime rate by town
- PRLZ: Proportion of residential land zoned for lots over 25,000 sq. ft
- INDUS: Proportion of non-retail business acres per town
- NOX: Nitric Oxide Concentration (parts per 10 million)
- AVG: Average number of rooms per dwelling
- AGE: Proportion of owner-occupied units built prior to 1940
- DIS: Weighted distance to five Boston employment centre
- RAD: Index of accessibility to radial highways
- TAX: Full-value property tax rate per $10,000
- MEDV: Median value of owner-occupied homes in $1000s

The aim is to predict the house price in thousands of dollars using the given features/ variables. The prices of the house that are represented by the variable MEDV is our target variable and the rest of variables provide information to predict the value of a house.

Descriptive Statistics

The data set "Housing_data.csv" is loaded in the R studio. Using the **str( )** function. we first check the structure of all variables of the dataset.

```
> str(data0)
'data.frame':   506 obs. of  10 variables:
 $ PCCR : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ PRLZ : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
 $ INDUS: num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
 $ NOX  : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ AVR  : num  6.58 6.42 7.18 7 7.15 ...
 $ AGE  : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
 $ DIS  : num  4.09 4.97 4.97 6.06 6.06 ...
 $ RAD  : int  1 2 2 3 3 3 5 5 5 5 ...
 $ TAX  : int  296 242 242 222 222 222 311 311 311 311 ...
 $ MEDV : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

There are total 10 variables and 506 observations in this data set. All of them are quantitative variables. While RAD and TAX are classified as integer, the other variables are number category.

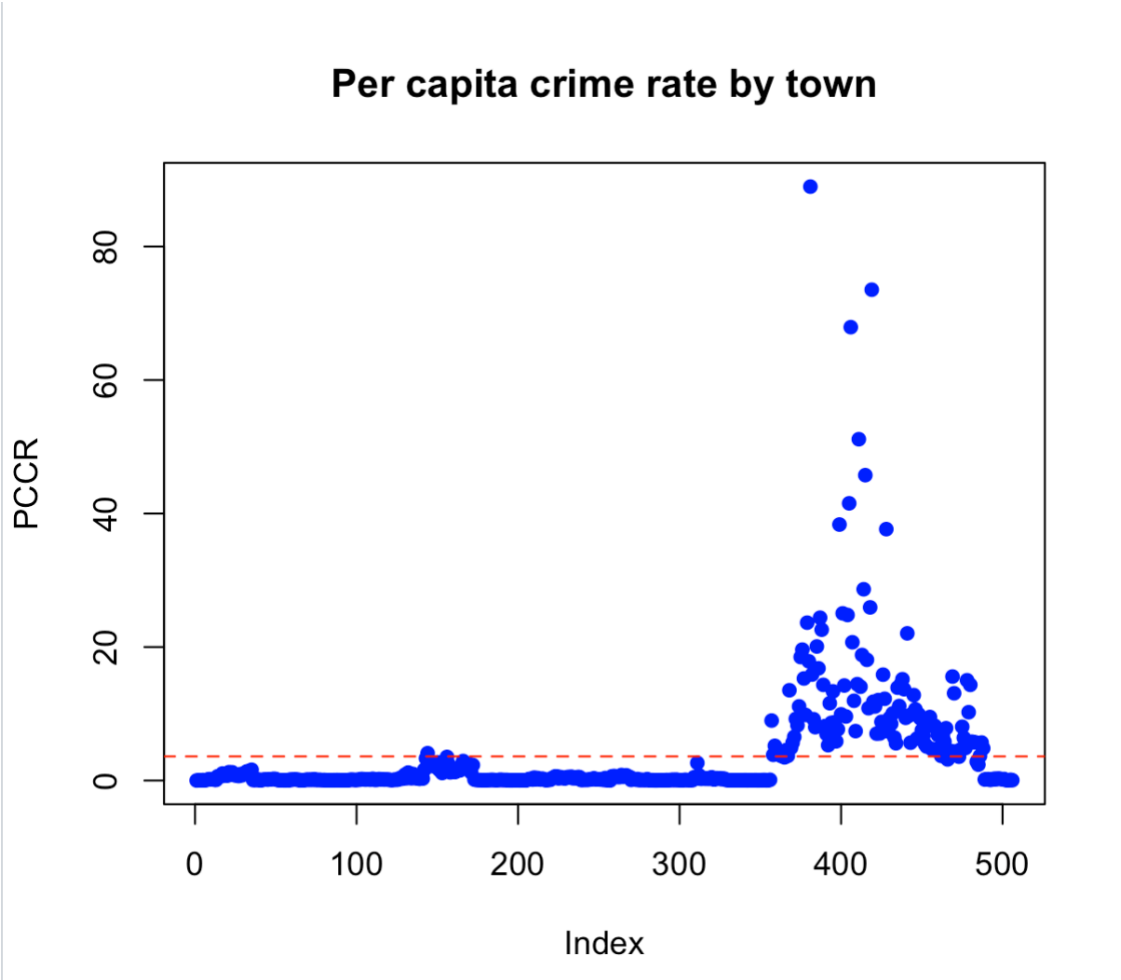
1. PCCR variable

PCCR variable indicates the number of crime per capita by town. Many functions are used to understand this variable. With the **class( )** function, the PCCR variable is categorised as quantitative type and values are continuous numbers. The minimum value of this variable is 0.00632 while the maximum value is 88.9762. This indicates that there is a wide range in between of those two values. However, the median number is 0.25651 implying that half of all the values is quite small and varies from 0.00632 to 0.25651, whilst the rest ranges from the mean to the maximum value. From this information, an anticipation can be drawn that the first half from the minimum value to the median value is denser than the other part.

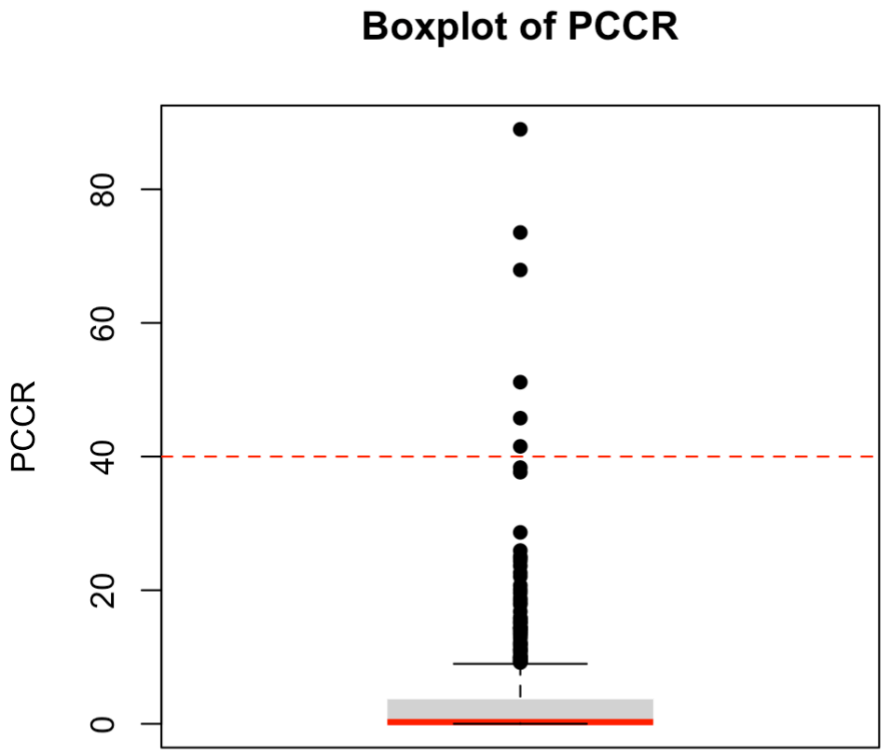
```
> mean(data0$PCCR)
[1] 3.613524
> min(data0$PCCR)
[1] 0.00632
> max(data0$PCCR)
[1] 88.9762
> class(data0$PCCR)
[1] "numeric"
> median(data0$PCCR)
[1] 0.25651
```

Plots of this data will prove that our anticipation is true or not. In this report, the author plotted data in 2 both the base package and the ggplot2 package.

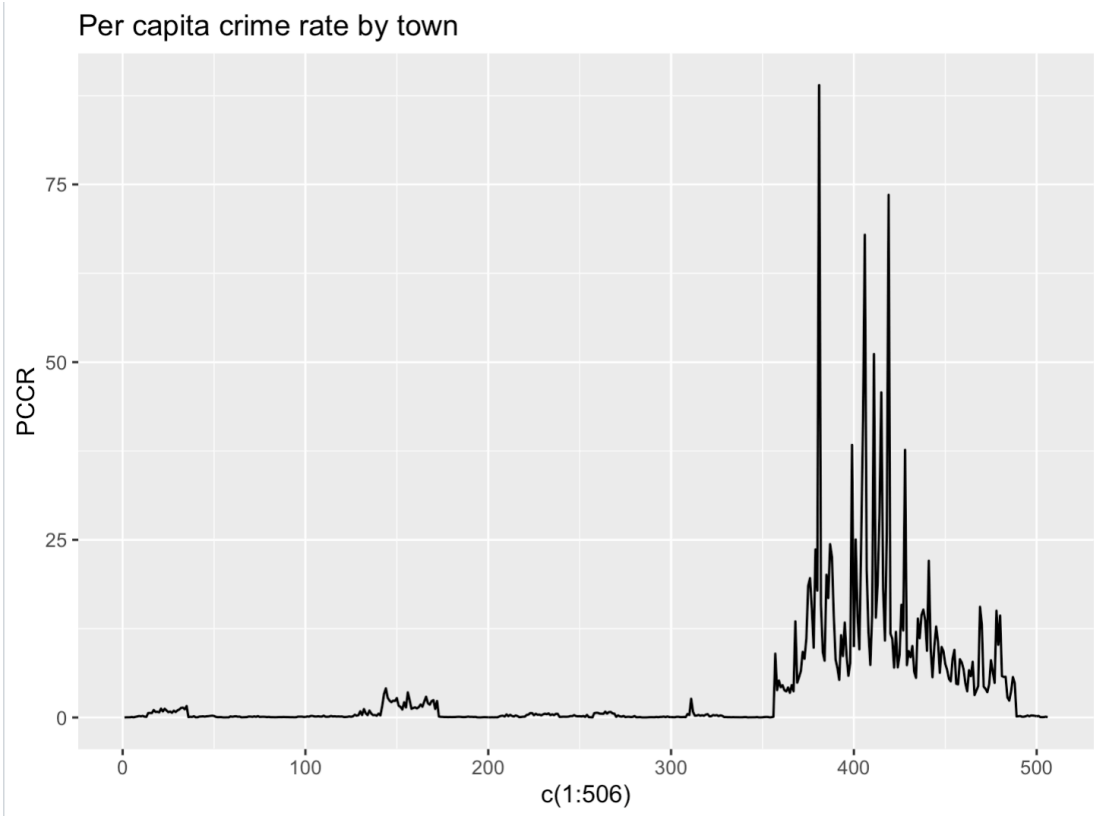
First of all, a scatter plot from base package is displayed as in figure 1 using function below.



The graph proves the anticipation above. According to that, about 25-30% of the observations is higher than the mean value.



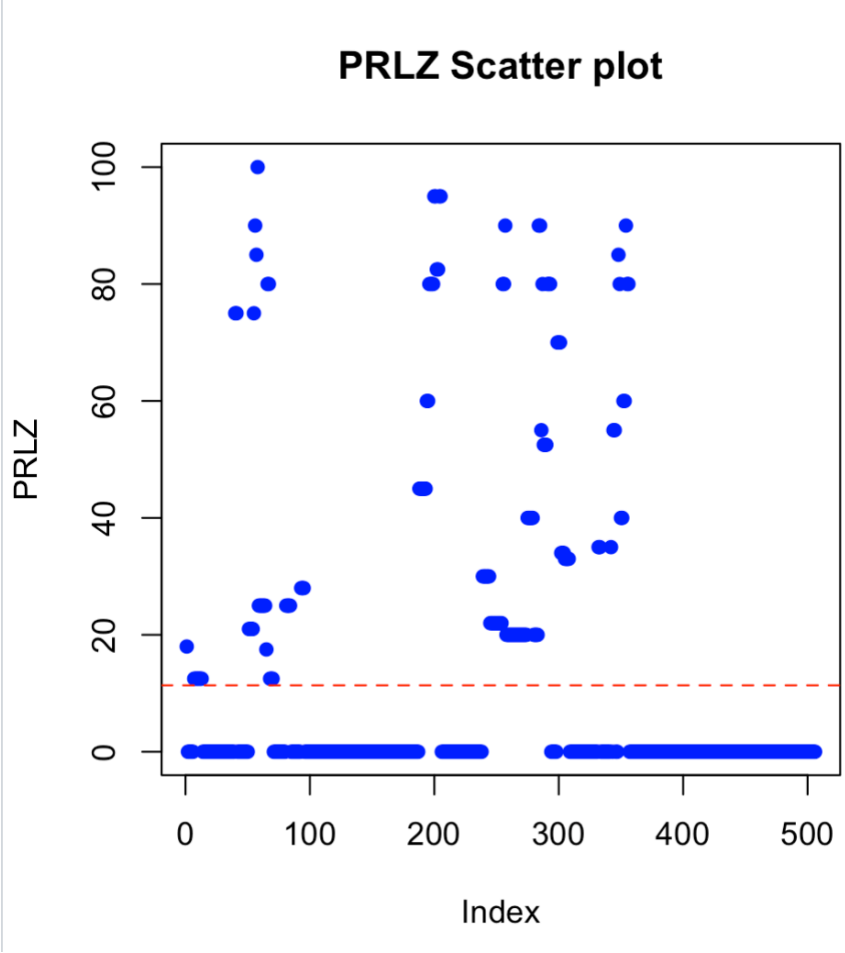
As in this graph, there are 6 points, which are outliers that have value more than 40. The range of values is clearer in the line graph below.



2. PRLZ variable

PRLZ variable indicates the proportion of residential land zoned for lots over 25,000 sq. ft. Many functions are used to understand this variable. With the **class( )** function, the PRLZ variable is categorised as quantitative type and values are continuous numbers. The minimum value is 0 and the maximum value is 100. The mean value is 11 and median is 0. According to the Scatter plot PRLZ below, most of the observations has the value of 0.

```
> class(data0$PRLZ)
[1] "numeric"
> mean(data0$PRLZ)
[1] 11.36364
> median(data0$PRLZ)
[1] 0
> min(data0$PRLZ)
[1] 0
> max(data0$PRLZ)
[1] 100
```



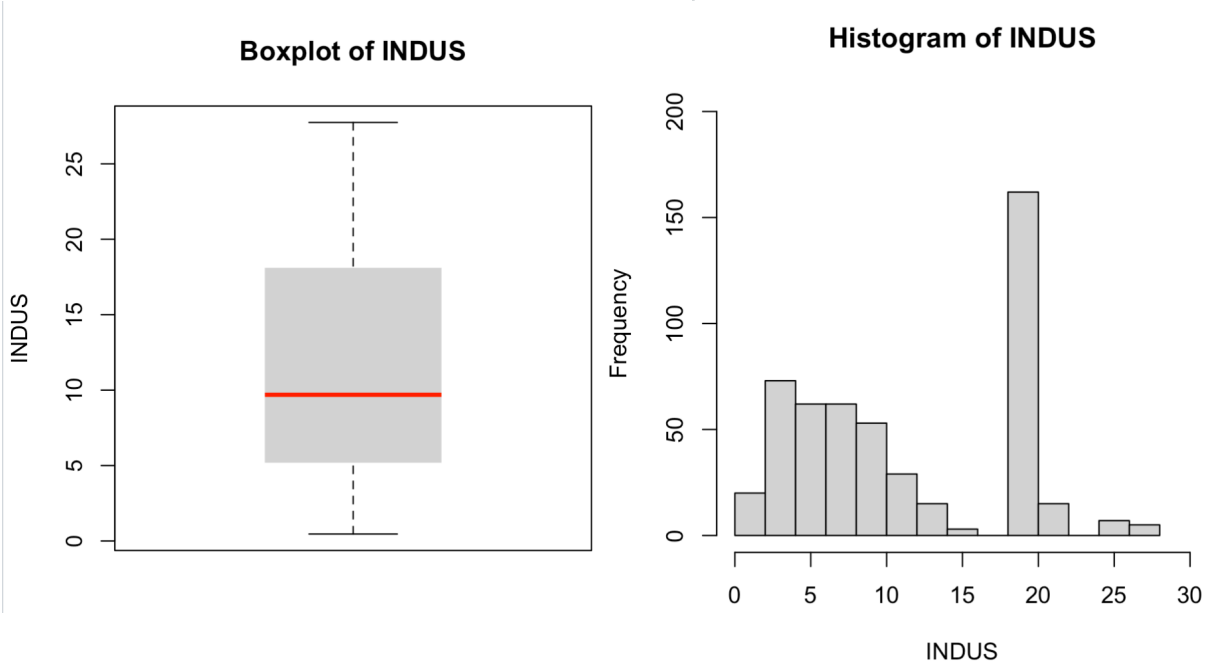
3. INDUS variable

INDUS variable indicates the proportion of non-retail business acres per town. This is a quantitative variable and values are continuous numbers. The minimum value is 0.46 while

the maximum value is 27.74. The mean value is 11.13578 and median is 9.69.

As in Boxplot of INDUS, there is no outlier. The amount from first quartil to the second is less than the one from the second quantil to the third. This can also be seen in the Histogram of INDUS (number of bins = 15), where the highest bin is the "18-20" bin.

```
> class(data0$INDUS)
[1] "numeric"
> mean(data0$INDUS)
[1] 11.13678
> median(data0$INDUS)
[1] 9.69
> min(data0$INDUS)
[1] 0.46
> max(data0$INDUS)
```



4. NOX variable

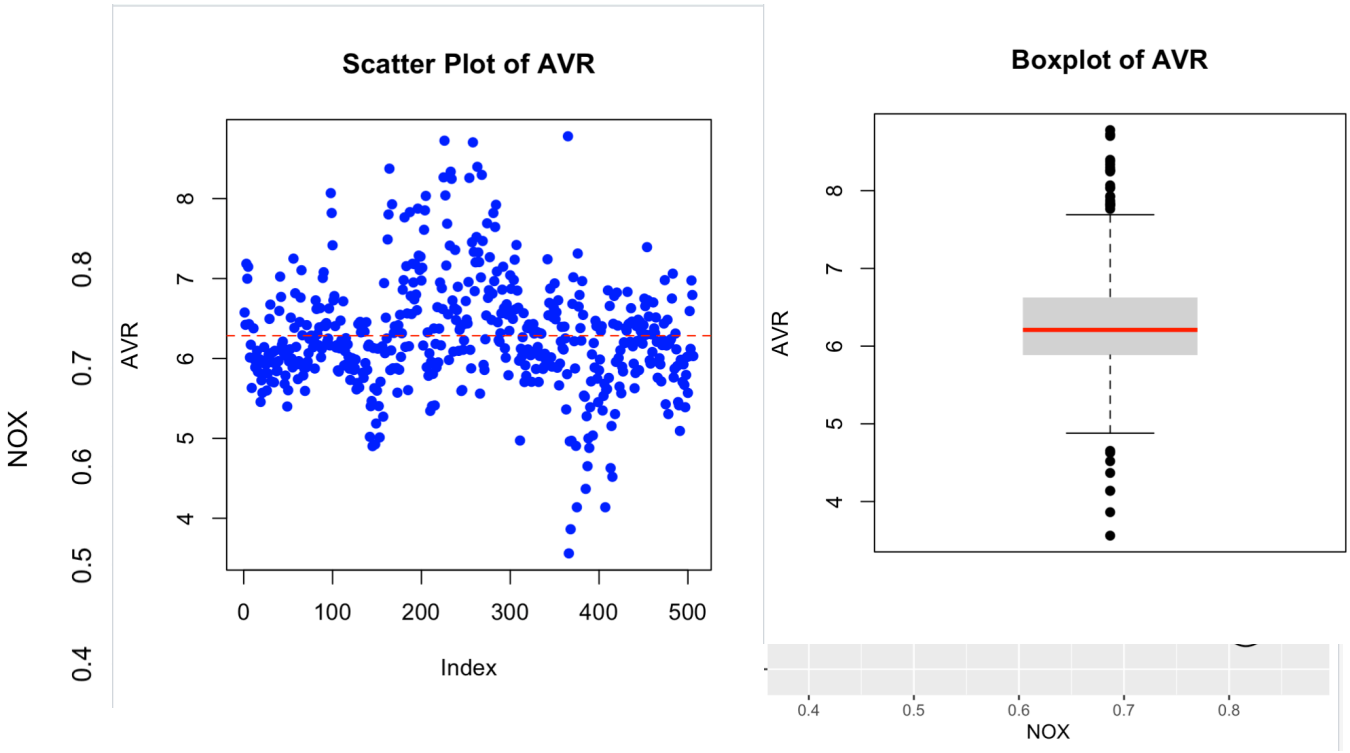
NOX variable indicates the nitric oxide concentration (parts per 10 million). This is also a quantitative variable and values are continuous numbers. The minimum value is 0.385 while the maximum value is 0.871. The mean value is 0.5546951 and median is 0.538.

Same as INDUS variable, NOX has no outlier in the box plot below. The amount from third quartil to the maximum value is large. This is seen in the Density Graph of NOX. The graph seems to be right-skewed (positive skewness).

```
> mean(data0$NOX)
[1] 0.5546951
> median(data0$NOX)
[1] 0.538
> min(data0$NOX)
[1] 0.385
> max(data0$NOX)
[1] 0.871
```

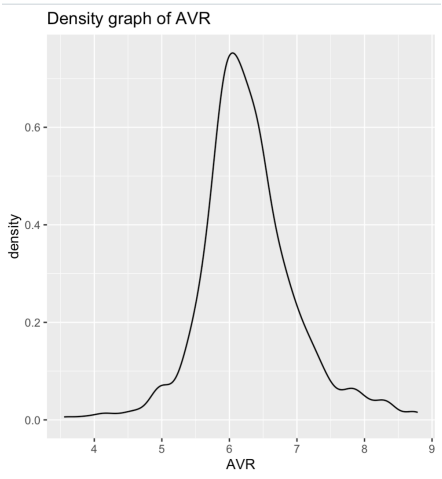
5. AVR variable

AVR is average number of rooms per dwelling. This is also a quantitative variable and values are continuous numbers. The minimum value is 3.561 while the maximum value is 8.78,



which means that the range will be 5.219.
The mean is 6.284634 and the difference between the mean and the median is not much with the value 6.2085 of the median. We can predict the graph may be symmetric, which is proved by the density graph of AVR below.

The scatter plot also proves our anticipation. According to this graph, the line red, the mean of 506 observations, lies in the middle of the graph implying this sample is well distributed.



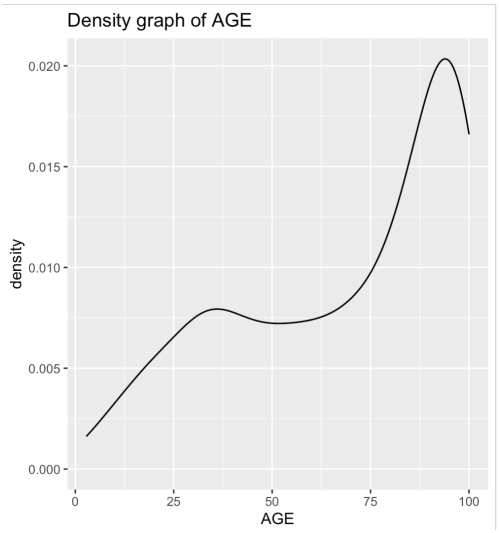
In the Boxplot of AVR variable, the range between the first and second quartil and the range between the second and third quantil are not much different. However, even though the graph seems to be symmetric, there are several outliers. The outliers above the maximum value outnumbers than the outliers below the minimum value.

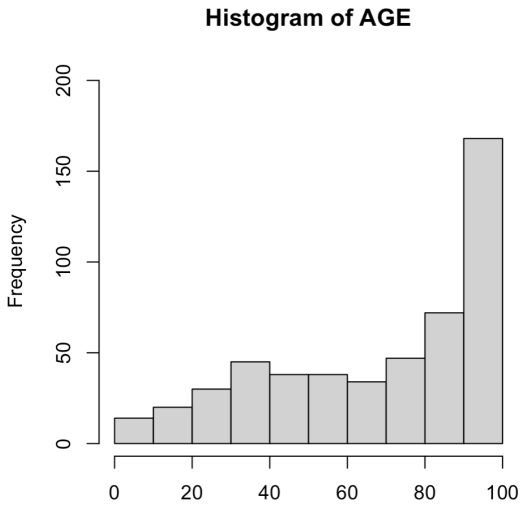
6. AGE variable

AGE is proportion of owner-occupied units built prior to 1940, another quantitative variable and values are continuous numbers. The min value is 2.9, while the max value is 100. Therefore, the range is 97.1. The mean is 68.5749 and the median is 77.5. As the mean is less than the median, the graph of this sample could be asymmetric, specifically left-skewed.

The density graph below proves our predication. The graph is left-skewed. The histogram of AGE variable shows that the value range from 90 to 100 has the highest frequency.

```
> class(data0$AGE)
[1] "numeric"
> mean(data0$AGE)
[1] 68.5749
> median(data0$AGE)
[1] 77.5
> min(data0$AGE)
```





)

7. DIS variable

DIS is weighted distances to five Boston employment centres. This variable is also a quantitative type and values are continuous numbers. The min value is 1.1296 while the max value is 12.1265. Therefore, the range is 10.9969. The median is 3.20745 and the mean is 3.795043.

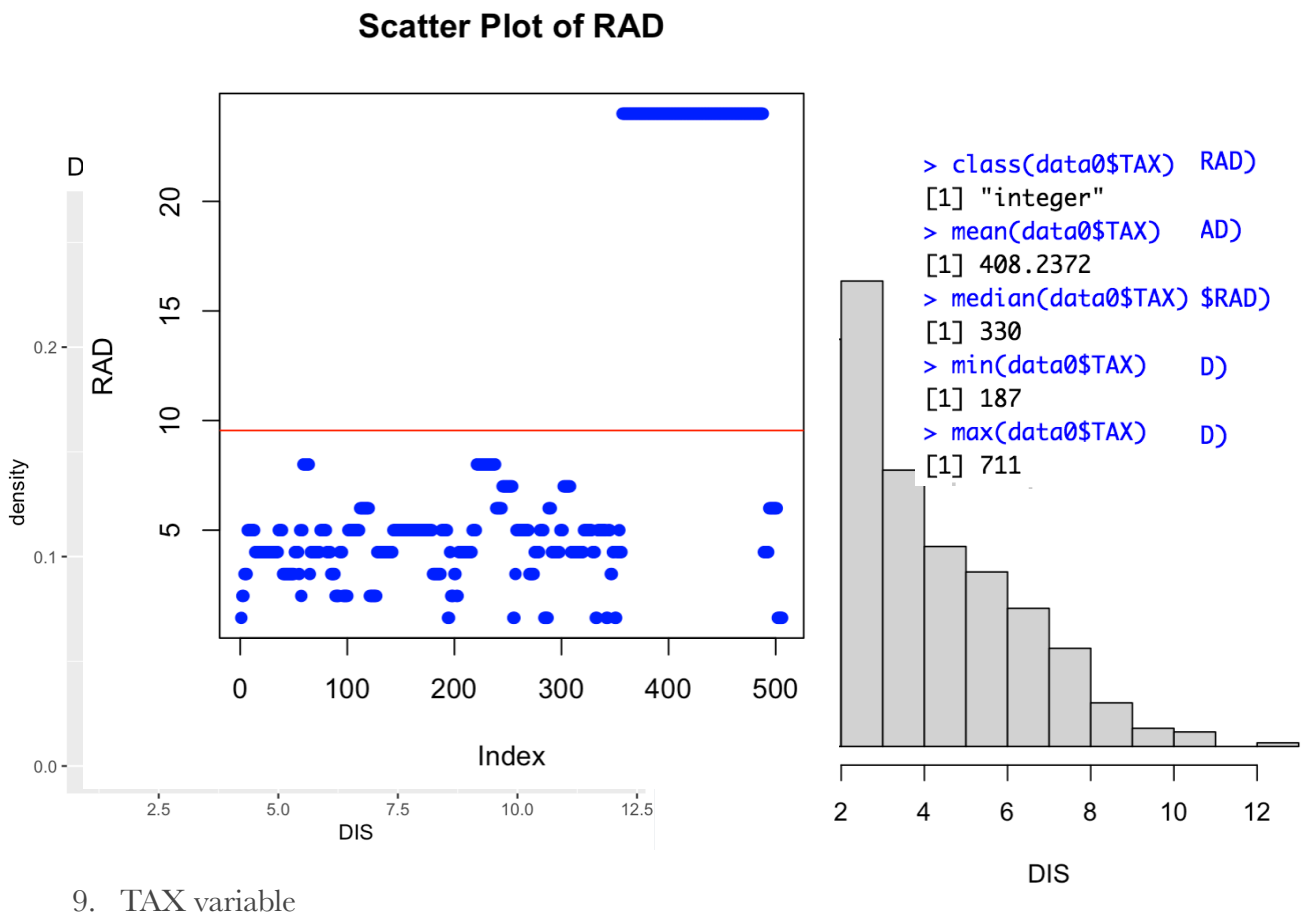
```
> class(data0$DIS)
[1] "numeric"
> mean(data0$DIS)
[1] 3.795043
> median(data0$DIS)
[1] 3.20745
> min(data0$DIS)
[1] 1.1296
> max(data0$DIS)
[1] 12.1265
```

The density graph of this data set below is definitely asymmetric, right-skewed. In the histogram, the value of DIS between 3 and 4 has the highest frequency. Overall, the highest value of DIS, the less the frequency is.

8. RAD variable

RAD is index of accessibility to radial highways. This variable is also a quantitative type but unlike others, the value is a discrete number. With the **class( )** function, this variable is classified as 'integer', not 'number' as other variables above. The min value is 1, max value is 24 and range is 23. The median value is 5 and the mean is 9.549407.

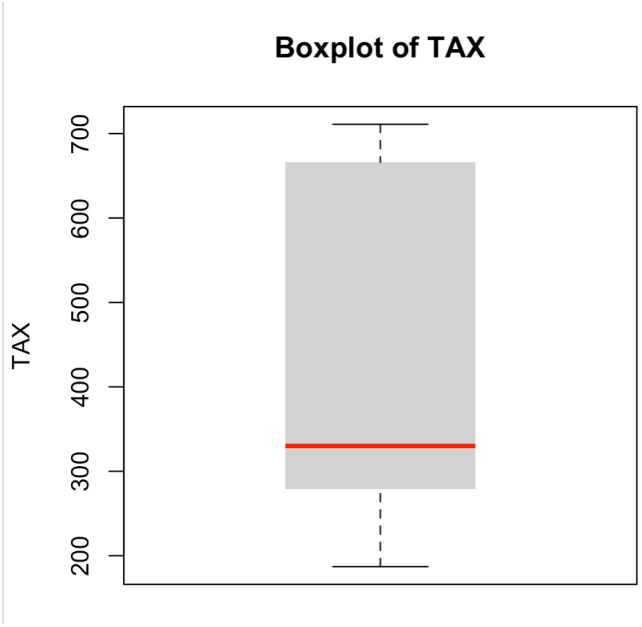
In the Scatter plot, the maximum value, 24, accounts for a large proportion in the total sample. The other values are less than the mean.



9. TAX variable

TAX is full-value property tax rate per $10,000. Like RAD variable, this is the quantitative type and the value is a discrete number. This variable varies a lot as the range is 524 with the min of 187 and the max of 711. The mean is 408.2372 and the median is 330.

According to box plot of TAX, there is no outliers in this graph. The most values of this data set is between 300 and 700.

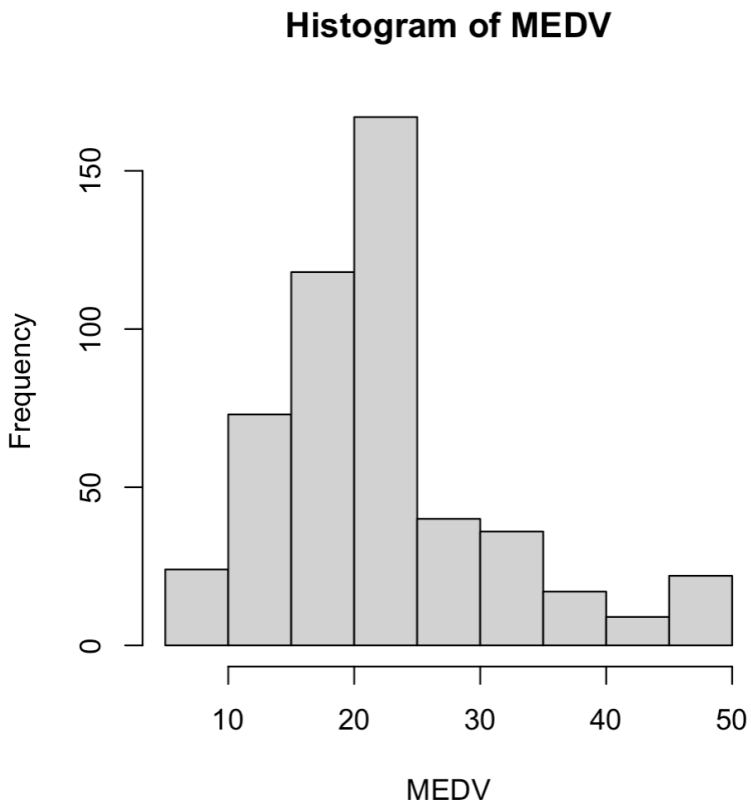


10. MEDV variable

MEDV is median value of owner-occupied homes in $1000s. The mean value is 22.53281 and the median value us 21.2. The min and max value are 5 and 50 respectively. The range is 45.

In the histogram of MEDV, the most value is between 20 and 25. Its frequency is even higher than 150 accounts for almost a third of the total number of observations.

```
> class(data0$MEDV)
[1] "numeric"
> mean(data0$MEDV)
[1] 22.53281
> median(data0$MEDV)
[1] 21.2
> min(data0$MEDV)
[1] 5
> max(data0$MEDV)
[1] 50
```



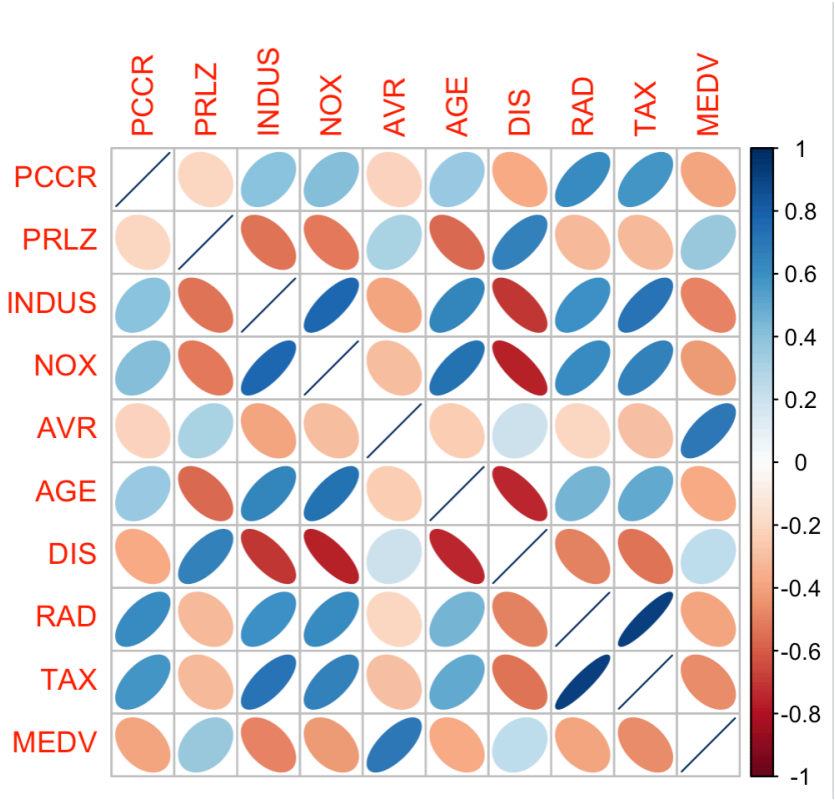
Correlation analysis between all variables

```
> cor(data0)
             PCCR       PRLZ      INDUS        NOX        AVR        AGE        DIS        RAD        TAX       MEDV
PCCR    1.0000000 -0.2004692  0.4065834  0.4209717 -0.2192467  0.3527343 -0.3796701  0.6255051  0.5827643 -0.3883046
PRLZ   -0.2004692  1.0000000 -0.5338282 -0.5166037  0.3119906 -0.5695373  0.6644082 -0.3119478 -0.3145633  0.3604453
INDUS   0.4065834 -0.5338282  1.0000000  0.7636514 -0.3916759  0.6447785 -0.7080270  0.5951293  0.7207602 -0.4837252
NOX     0.4209717 -0.5166037  0.7636514  1.0000000 -0.3021882  0.7314701 -0.7692301  0.6114406  0.6680232 -0.4273208
AVR    -0.2192467  0.3119906 -0.3916759 -0.3021882  1.0000000 -0.2402649  0.2052462 -0.2098467 -0.2920478  0.6953599
AGE     0.3527343 -0.5695373  0.6447785  0.7314701 -0.2402649  1.0000000 -0.7478805  0.4560225  0.5064556 -0.3769546
DIS    -0.3796701  0.6644082 -0.7080270 -0.7692301  0.2052462 -0.7478805  1.0000000 -0.4945879 -0.5344316  0.2499287
RAD     0.6255051 -0.3119478  0.5951293  0.6114406 -0.2098467  0.4560225 -0.4945879  1.0000000  0.9102282 -0.3816262
TAX     0.5827643 -0.3145633  0.7207602  0.6680232 -0.2920478  0.5064556 -0.5344316  0.9102282  1.0000000 -0.4685359
MEDV   -0.3883046  0.3604453 -0.4837252 -0.4273208  0.6953599 -0.3769546  0.2499287 -0.3816262 -0.4685359  1.0000000
```

According to the figure above, regarding to MEDV variable, AVR variable has the strongest positive correlation with 0.69. PRLZ and DIS also have positive correlation with MEDV but as not strong as AVR. The numbers are 0.36 and 0.25 respectively. In contrast, INDUS,

TAX, NOX, PCCR, RAD, AGE have negative correlation with MEDV. The strongest negative correlation with MEDV is INDUS (-0.48), followed by TAX (-0.47), NOX (-0.43), PCCR (-0.39), RAD (-0.38) and AGE (-0.38).

To sum up, the more AVR, PRLZ and DIS increase, the more MEDV value rises. Likewise, the more INDUS, TAX, NOX, PCCR, RAD and AGE value decrease , the more MEDV value gains. Among AVR, PRLZ and DIS, AVR has the most significant impact on MEDV numbers. On the other hand, among all, INDUS has the most negative effect on MEDV values.



The figure is the illustration of correlation between all pairs of variables. In this figure, the pair that has the strongest positive correlation is TAX and RAD (0.91), followed closely INDUS & NOX (0.76) and INDUS & TAX (0.72).

On the other hand, the strongest negative correlation is -0.76, the pair of NOX & DIS. Other strong negative correlated pairs are DIS & INDUS and DIS & AGE.

Building regression model to predict house prices:

The regression equation for the prices of the house (MEDV) with respect to all explanatory variables is:

as T is short for Theta.

$$MEDV = T0 + T1*PCCR + T2*PRLZ + T3*INDUS + T4*NOX + T5*AVR + T6*AGE + T7*DIS + T8*RAD + T9*TAX$$

T0 is intercept. Given that all variables are 0, the MEDV value will be T0. It's a constant value. In this case, the medium value of oner-occupied homes will be T0 even if all other variables reach 0.

T1 is coefficient of PCCR. This means that if PCCR increases by 1 unit, MEDV will increase by T1 units given that all other variables are 0 or have the same amount as the previous.

T2, T3, T4, …, T9 is same as T1. They are coefficient of variables.

Implement the regression model

According to the summary of model0 as in the figure below, the intercept, which is T0, is 2.14. Consequently, T1 to T9 are -0.19, 0.07, -0.079, -11.56, 6.82, -0.05, -1.75, 0.18, -0.02 respectively. If the number is larger than 0 then that variable has positive correlation with variable MEDV. Otherwise, it has negative correlation with MEDV.

The next value we should focus on is the p-value, which is "Pr(>|t|)" in the figure below. The p-value should be smaller than 0.1. The smaller the p-value is, the more significant that variable can affect the MEDV. As in figure below, the intercept (T0) has the p-value of 0.6. It is quite high. We should remove it to make the model fit better.

PCCR , PRLZ, AVR and DIS has a relatively small p-value. This is good. It means that those variables have significant impacts on the MEDV values.

INDUS variable has high p-value of 0.27. It's more than 10%. This variable should be removed to make the model fit better. The p-value of other variables are not high so they may not need to be removed from the model.

The multiple R-squared is 0.6275. This is not a good number. The model needs to be adjusted to fit well. However, the F-statistic and p-value of overall are small, which is good.

```
> summary(model0)

Call:
lm(formula = MEDV ~ PCCR + PRLZ + INDUS + NOX + AVR + AGE + DIS +
    RAD + TAX, data = data0)

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
```

Adjust the model

As discussed above, the p-value of the intercept is too high. Therefore, the intercept needs to be removed from the model.

```
#No Intercept
model0.1<-lm(MEDV~PCCR+PRLZ+INDUS+NOX+AVR+AGE+DIS+RAD+TAX+0,data=data0)
summary(model0.1)
```

We rebuild the linear regression as below with a new model: model0.1.

As we can see in the new model (where there is no more intercept), all variables has the p-value is significantly small except INDUS with the p-value of 0.28. Most of the p-value is less than 10%. Overall p-value is extremely small, which is good along with the F-statistic.

```
> summary(model0.1)

Call:
lm(formula = MEDV ~ PCCR + PRLZ + INDUS + NOX + AVR + AGE + DIS +
    RAD + TAX + 0, data = data0)

Residuals:
    Min      1Q  Median      3Q     Max
-15.811  -2.949  -0.708   1.983  38.204

Coefficients:
        Estimate Std. Error t value Pr(>|t|)
PCCR   -0.188135   0.038049  -4.945 1.05e-06 ***
PRLZ    0.072283   0.015088   4.791 2.20e-06 ***
INDUS  -0.077645   0.072054  -1.078 0.281736
NOX   -10.390755   3.610620  -2.878 0.004177 **
AVR     6.986317   0.263458  26.518  < 2e-16 ***
AGE    -0.052305   0.014620  -3.578 0.000381 ***
DIS    -1.679667   0.185187  -9.070  < 2e-16 ***
RAD     0.175094   0.074255   2.358 0.018759 *
TAX    -0.016181   0.004322  -3.744 0.000203 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.66 on 497 degrees of freedom
Multiple R-squared:  0.9469,    Adjusted R-squared:  0.9459
F-statistic:   984 on 9 and 497 DF,  p-value: < 2.2e-16

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
```

Multiple R-squared increases significantly from 0.6275 to 0.9469 proving this model is better fitted than the last one, model0.

However, the p-value of INDUS variable is still over 0.1. Therefore, we run a new model in which we remove the INDUS variable.

```
> summary(model1)

Call:
lm(formula = MEDV ~ PCCR + PRLZ + NOX + AVR + AGE + DIS + RAD +
    TAX + 0, data = data0)

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 ***
RAD    0.197106   0.071401   2.761 0.005983 **
TAX   -0.018385   0.003808  -4.828 1.84e-06 ***
---
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
```

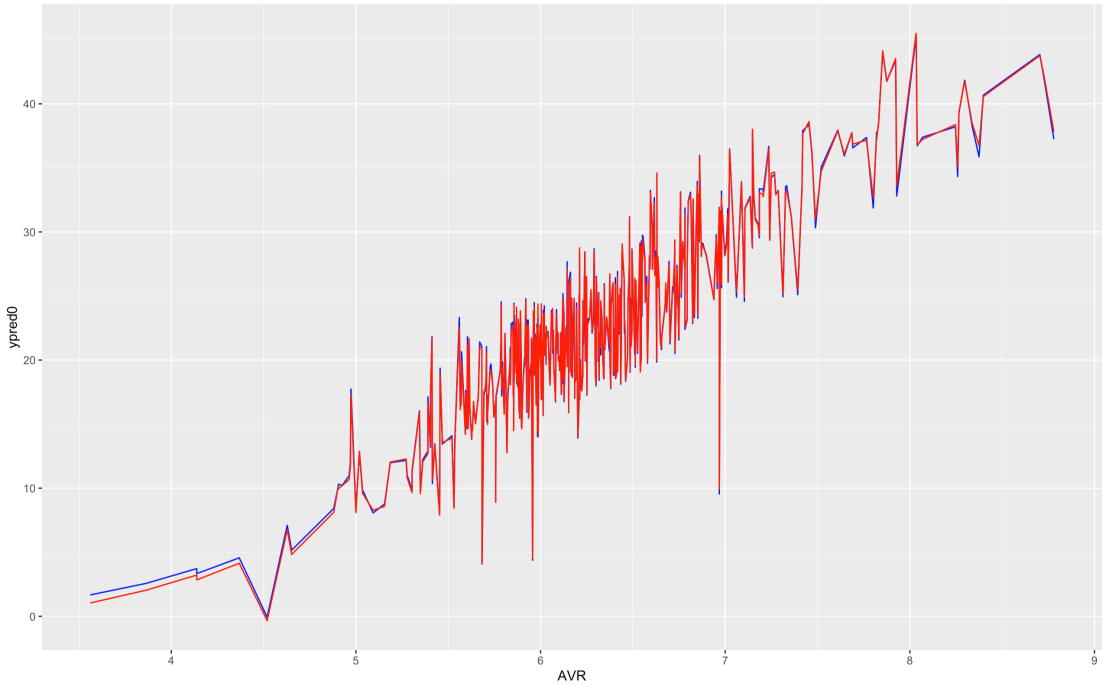
With the latest model, the multiple R-squared doesn't change much, still significant with 0.9467. F-statistic and p-value overall indicated that this model is well fitted. Now we check the p-value of each variable. All variable have significantly small p-value less than 10%. The model is finalised with ***model1***.

Now, we make a plot of original and fitted values to check the goodness of the model.

First, we define the equation of the original model (ypred0) and the adjusted model (ypred1) as below

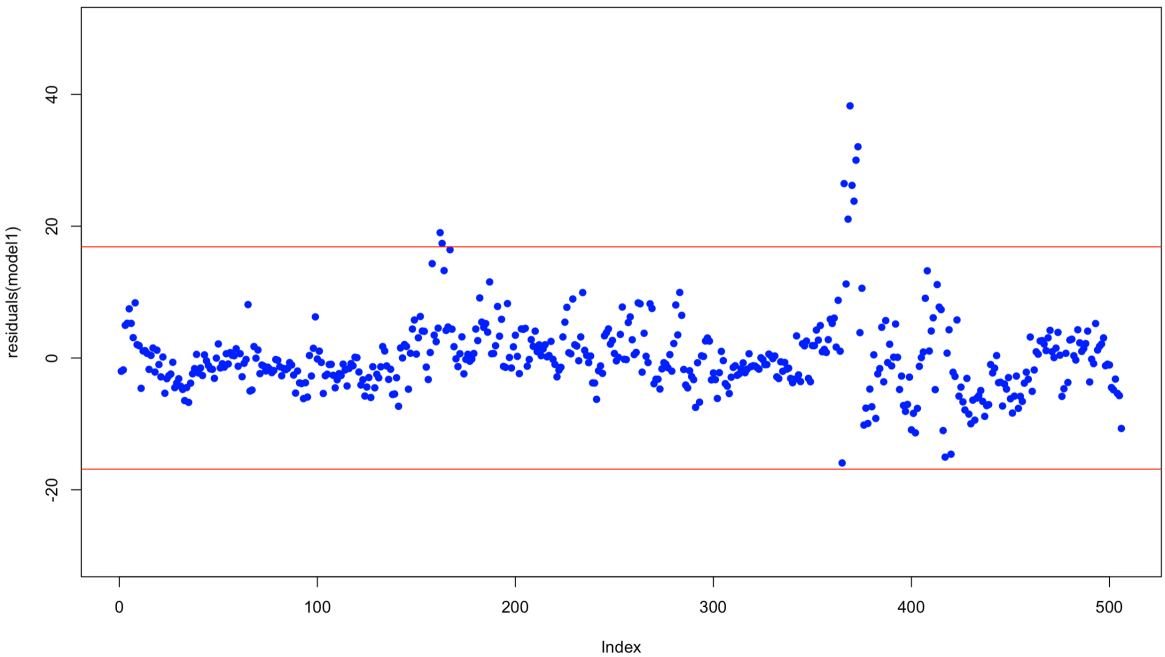
```
ypred0=coef(model0)[1]+coef(model0)[2]*data0$PCCR+coef(model0)[3]*data0$PRLZ+coef(model0)[4]*data0$INDUS+
  coef(model0)[5]*data0$NOX+coef(model0)[6]*data0$AVR+coef(model0)[7]*data0$AGE+coef(model0)[8]*data0$DIS+
  coef(model0)[9]*data0$RAD+coef(model0)[10]*data0$TAX


ypred1=coef(model1)[1]*data0$PCCR+coef(model1)[2]*data0$PRLZ+coef(model1)[3]*data0$NOX+
  coef(model1)[4]*data0$AVR+coef(model1)[5]*data0$AGE+coef(model1)[6]*data0$DIS+
  coef(model1)[7]*data0$RAD+coef(model1)[8]*data0$TAX
```

Next we plot two models.

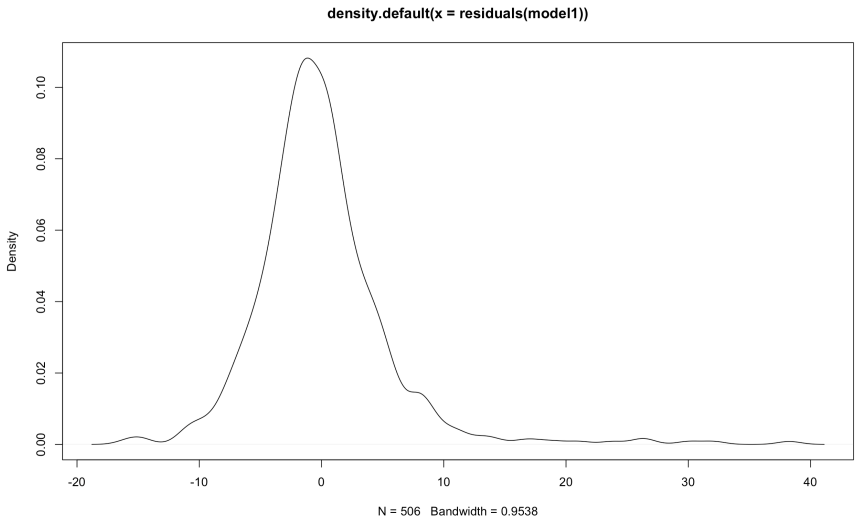


The original is in blue and the fitted one is in red. As we can see, there is not much difference between 2 models.

Next, we plot the residuals of the final model.



As we can see in the figure below most the residuals are in between two lines. However, there are some outliers. There are total 9 points above the higher line, which accounts for 1.77%. So, it is acceptable. the final fitted really well.



To sum up, the findings are:

- Data analysis of all variables given.
- Correlation between variables. The main target is to analyse the factor that can affect the MEDV value. As we have analysed, the AVR value has the most impact on the MEDV value. In contrast, the INDUS and TAX may have negative effect on the price. The housing agency can use this new information to control the price better.
- Correlation of other pairs. Some variables are linked to gather. The housing agency can take note on this, and exploit it to gain advantage for the company.
- The linear equation includes only necessary variables which has huge impact on the price with the smallest p-value. This equation can be used to predict the price of houses for agency.
- The model is adjusted well enough with the smallest p-value and and significant impact to the final price, helping housing agency in decision making.
- The residual plot needs to be rechecked to make sure the model is correctly fitted, not overrated nor underrated.