1. Description of the Boston Data Set

The Boston data set contains information collected in 1970 by the U.S. Census Service concerning housing on the Boston, MA area. It includes 506 observations of 14 variables.

```
Console   Terminal
E:/Dropbox/RU DataScience/MSDS660/Week2/Assignment/
> install.packages("MASS") #Install the MASS package for the first time
Installing package into ‘C:/Users/RodneyWeakly/Documents/R/win-library/3.5’
(as ‘lib’ is unspecified)
trying URL 'https://cran.rstudio.com/bin/windows/contrib/3.5/MASS_7.3-50.zip'
Content type 'application/zip' length 1172610 bytes (1.1 MB)
downloaded 1.1 MB

package ‘MASS’ successfully unpacked and MD5 sums checked

The downloaded binary packages are in
        C:\Users\RodneyWeakly\AppData\Local\Temp\RtmpOKmv0A\downloaded_packages
> library(MASS) #load the MASS package
> data("Boston") #Load the Boston dataset
> ?Boston #INfo about the dataset
> View(Boston) #Preview the dataset
> names(Boston) #List the names of the variable in the Boston dataset
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
> 
```

The variables and their descriptions are as follow:

- `crim` – per capita crime rate by town;
- `zn` – proportion of residential land zoned for lots over 25,000 sq.ft.;
- `indus` – proportion of non-retail business acres per town;
- `chas` – Charles River dummy variable (= 1 if tract bounds river; 0 otherwise);
- `nox` – nitrogen oxides concentration (parts per 10 million);
- `rm` – average number of rooms per dwelling;
- `age` – proportion of owner-occupied units built prior to 1940;
- `dis` – weighted mean of distances to five Boston employment centers;
- `rad` – index of accessibility to radial highways;
- `tax` – full-value property-tax rate per \$10,000;
- `ptratio` – pupil-teacher ratio by town;
- `black` – *1000(Bk - 0.63)^2* where *Bk* is the proportion of blacks by town;
- `lstat` – lower status of the population (percent);
- `medv` – median value of owner-occupied homes in \$1000s.

```
Console   Terminal
E:/Dropbox/RU DataScience/MSDS660/Week2/Assignment/
        C:\Users\RodneyWeakly\AppData\Local\Temp\RtmpOKmv0A\downloaded_packages
> library(MASS) #load the MASS package
> data("Boston") #Load the Boston dataset
> ?Boston #INfo about the dataset
> View(Boston) #Preview the dataset
> names(Boston) #List the names of the variable in the Boston dataset
 [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"     "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
> str(Boston) #Display the internal structure of the dataset
'data.frame':   506 obs. of  14 variables:
 $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
 $ zn     : 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 ...
 $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
 $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
 $ rm     : 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    : num  296 242 242 222 222 222 311 311 311 311 ...
 $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
> 
```

I looked at the internal structure of the data set using the ***str(Boston)*** command, which returns the data type and a few first observations for each variable. Two variables ("chas" and "rad") use int data type, the rest of the variables use num data type.

2. Data Exploration

In order to look at the sample data, I used ***head(Boston)*** and ***tail(Boston)*** commands.

```
Console   Terminal
E:/Dropbox/RU DataScience/MSDS660/Week2/Assignment/
 $ black  : num  397 397 393 395 397 ...
 $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
 $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
> head(Boston) #Display forst few rows of data
     crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90  4.98 24.0
2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90  9.14 21.6
3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83  4.03 34.7
4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63  2.94 33.4
5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90  5.33 36.2
6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12  5.21 28.7
> tail (Boston) #Display last few rows of data
       crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
501 0.22438  0  9.69    0 0.585 6.027 79.7 2.4982   6 391    19.2 396.90 14.33 16.8
502 0.06263  0 11.93    0 0.573 6.593 69.1 2.4786   1 273    21.0 391.99  9.67 22.4
503 0.04527  0 11.93    0 0.573 6.120 76.7 2.2875   1 273    21.0 396.90  9.08 20.6
504 0.06076  0 11.93    0 0.573 6.976 91.0 2.1675   1 273    21.0 396.90  5.64 23.9
505 0.10959  0 11.93    0 0.573 6.794 89.3 2.3889   1 273    21.0 393.45  6.48 22.0
506 0.04741  0 11.93    0 0.573 6.030 80.8 2.5050   1 273    21.0 396.90  7.88 11.9
>
```

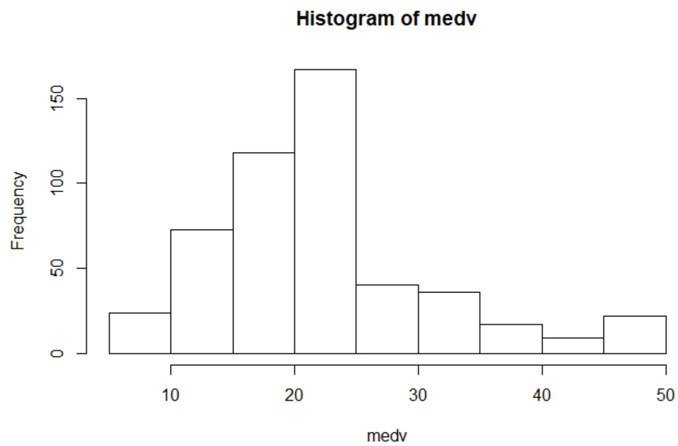
Then, I used ***summary(Boston)*** to display the basic summary statistics for each variable in the dataset.

```
Console   Terminal
E:/Dropbox/RU DataScience/MSDS660/Week2/Assignment/
506 0.04741  0 11.93    0 0.573 6.030 80.8 2.5050   1 273    21.0 396.90  7.88 11.9
> summary(Boston) #Display basic summary statistics for each variable
      crim                zn             indus            chas              nox               rm             age              dis
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
      rad              tax           ptratio          black            lstat            medv
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00
>
```

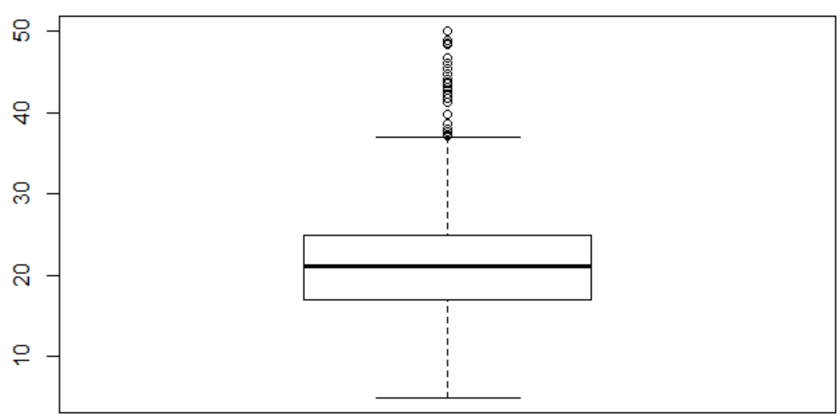
For example, the age variable, showing proportion of the owner-occupied dwellings built before 1940, varies from 2.9 % (min) to 100% (max) with the mean value of 68.57%, and median 77.50%. The interquartile range is 45.02 – 94.08.

I was particularly interested in the ***medv*** variable, as it represents a median value of owner-occupied dwellings expressed in thousands of dollars. The minimum median value in the Boston dataset is 5.00 and the maximum is 50.00. The median and mean are slightly different, with 21.20 thousand for the median and 22.53 thousand for the mean. Since the mean is higher than the median it implies a right skewed distribution. The middle 50% of the data points lay between 17.02 and 25.00 thousand dollars.

The ***hist(medv)*** command I used to create a histogram for median house values confirmed a right skewed distribution, which is not uncommon for real estate prices, where a small number of high-priced houses push the mean values up.

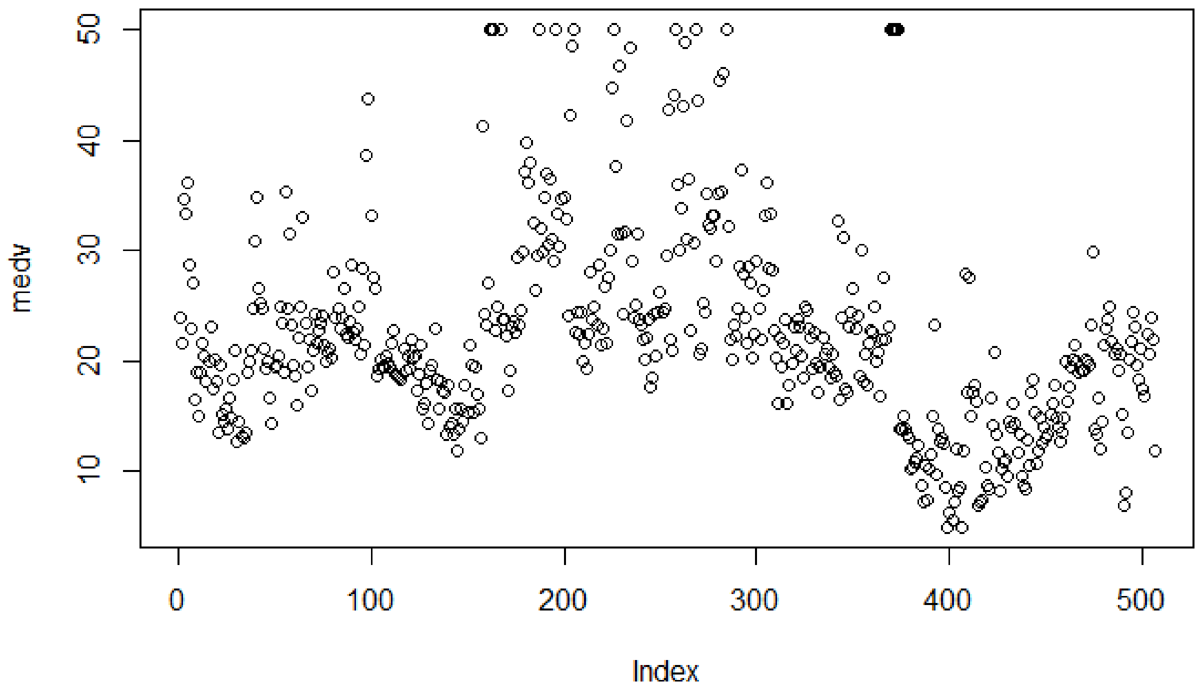


Creating a boxplot for median house values using ***boxplot(medv)*** only reinforced this conclusion:



The boxplot shows a significant number of outliers with median values above approximately 38.00 thousand dollars

I then used ***plot(medv)*** to create a scatter plot for median house values to visually inspect data for possible trends.

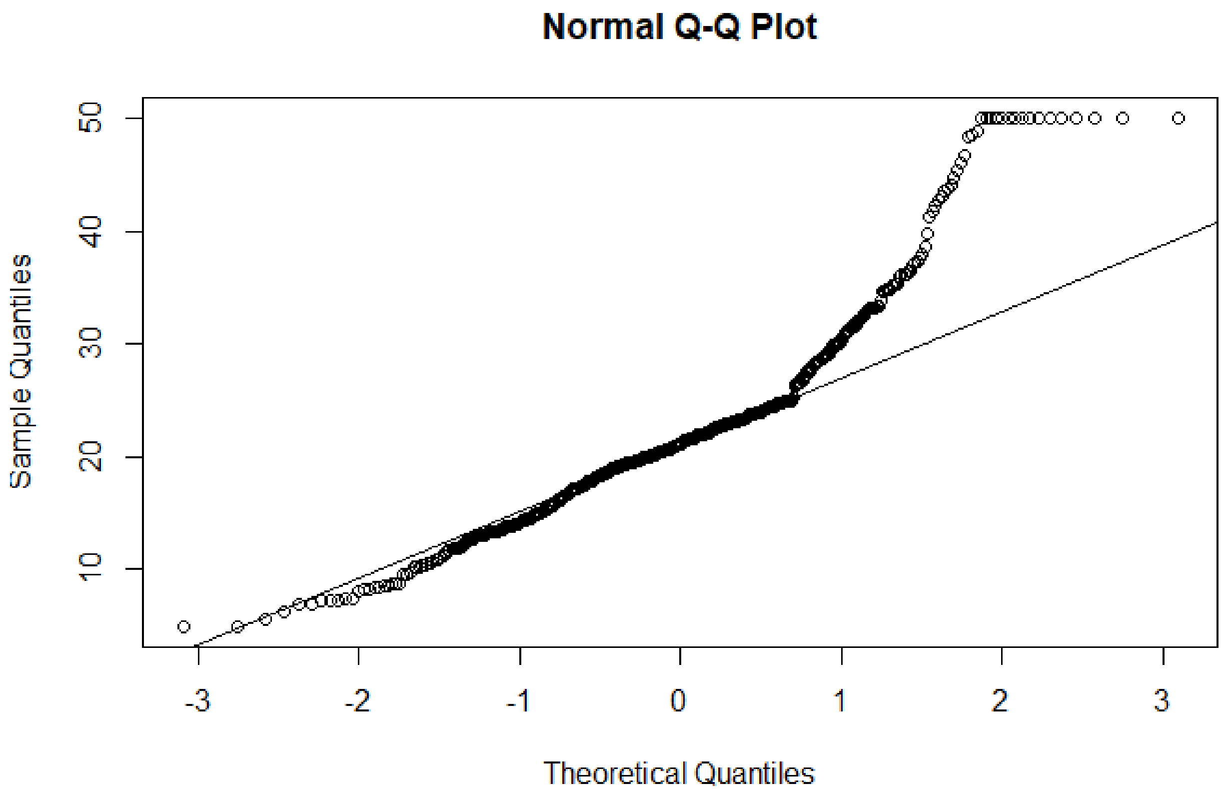


The plot showed a relatively high variability of data, and the standard deviation calculations showed 9.197 thousand dollars.

```
> sd(medv) #Standard Deviation for median house values
[1] 9.197104
```

In order to inspect normality of the mean house values, I looked at Q-Q norm plot:

```
> qqnorm(medv)#Norm Q-Q plot for median house values
> qqline(medv)
```

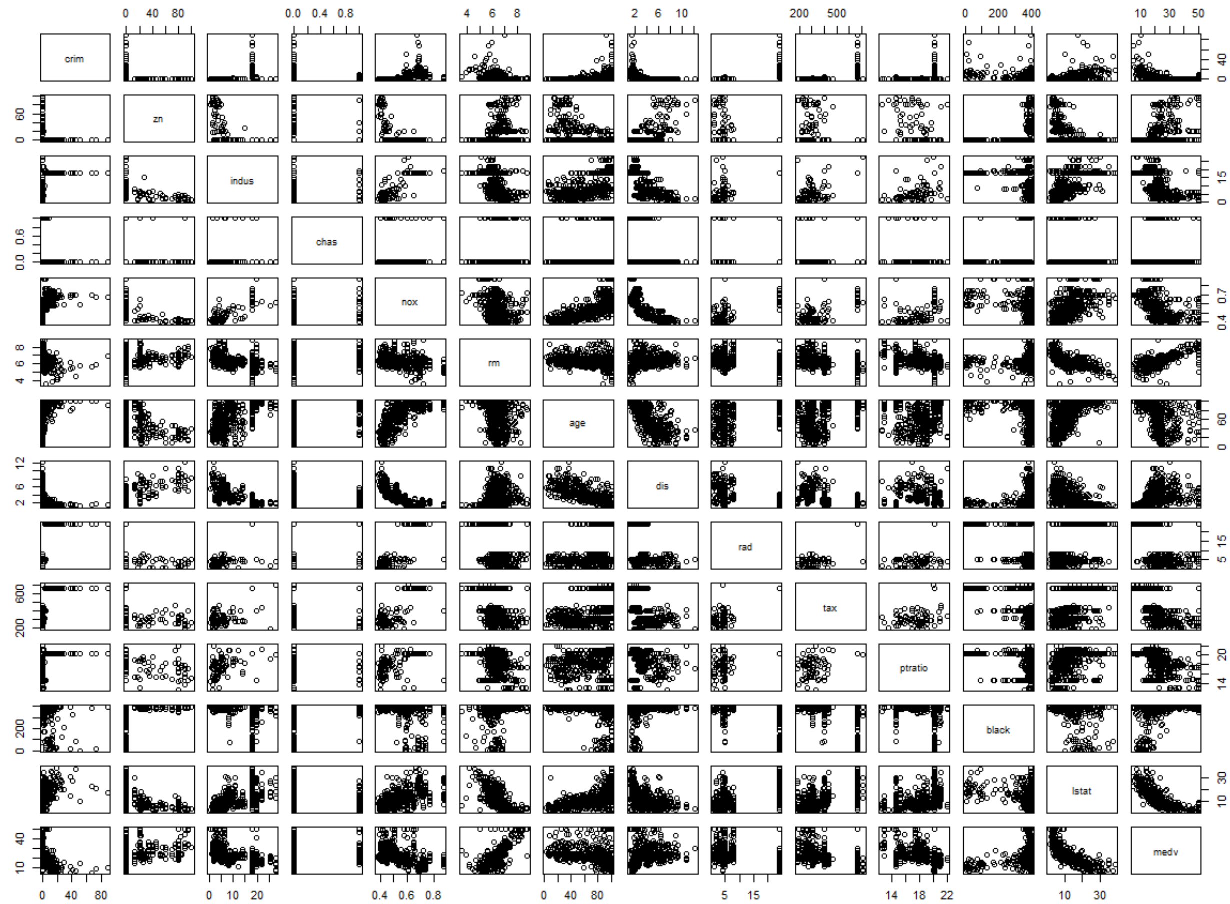


It showed that majority of the datapoints loosely follow theoretical Q-Q line, but the sample deviates to the top on the right side of the graph, which is typical for right-skewed distributions.

3. Pairwise Scatter Plots

Next, I created a matrix of pairwise scatter plots for all variables in the dataset.

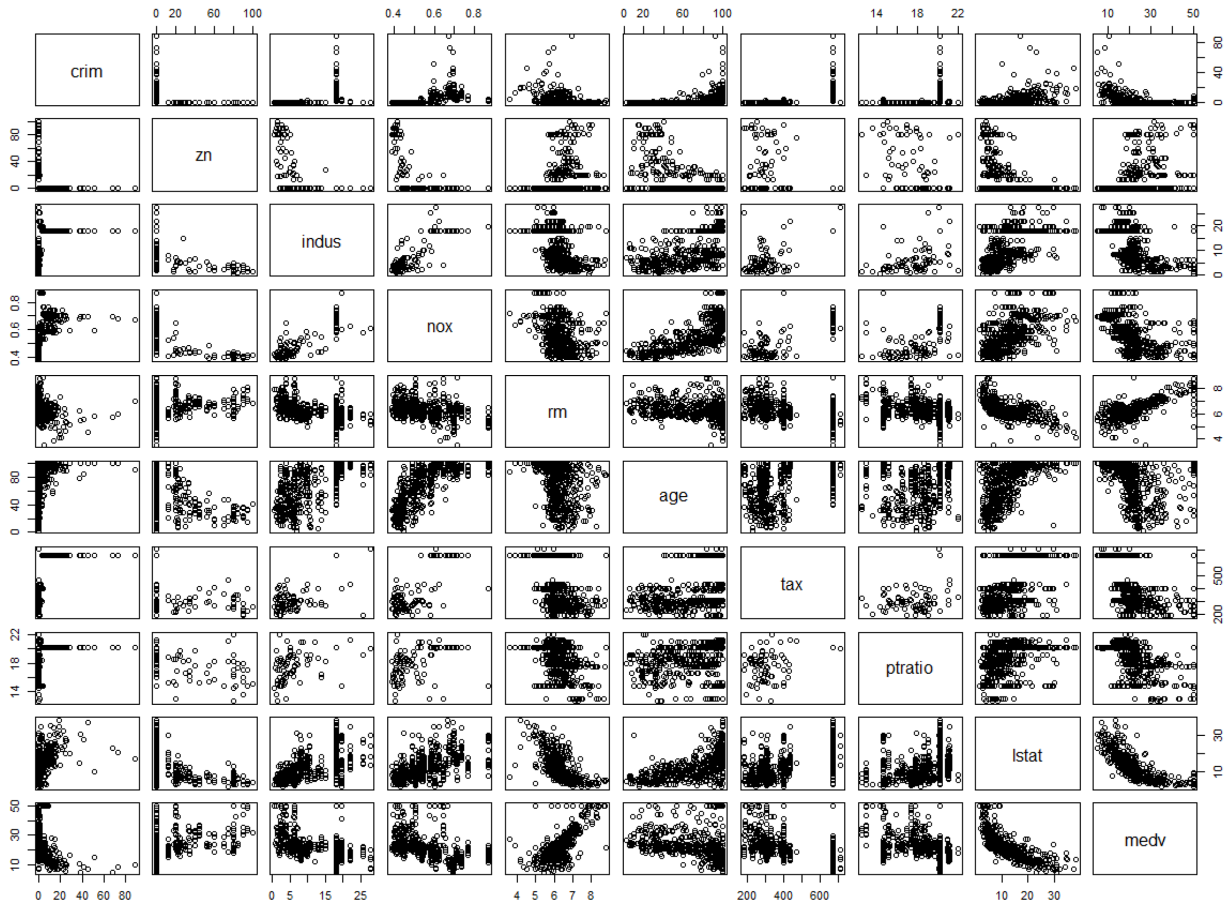
```
> pairs(~crim+zn+indus+chas+nox+rm+age+dis+rad+tax+ptratio+black+lstat+medv, data
=Boston) #Create scatterplot matrix
```



Due to the number of variables (14) this matrix was not very convenient for visual analysis, but it provided enough information allowing to eliminate four variables that were less significant for the purposes of the analysis – chas, dis, rad and black.

I displayed pairwise scatter plots for the remaining variables using the following command:

```
> pairs(~crim+zn+indus+nox+rm+age+tax+ptratio+lstat+medv, data=Boston) #Create sc
atterplot matrix without chas, dis, rad and black variables
```



Looking at the medv variable (the median value of owner-occupied homes), these plots suggest that there might be a positive correlation between the rm (average number of rooms per dwelling) and medv, negative correlation between ptratio (pupil-teacher ratio) and medv, and negative correlation between lstat (percentage of lower status of the population) and medv. The rest of the pairwise plots did not show definitive trends suggesting strong correlation with medv.

In order to confirm this conclusion, I used ***cor(Boston)*** to display the correlation matrix.

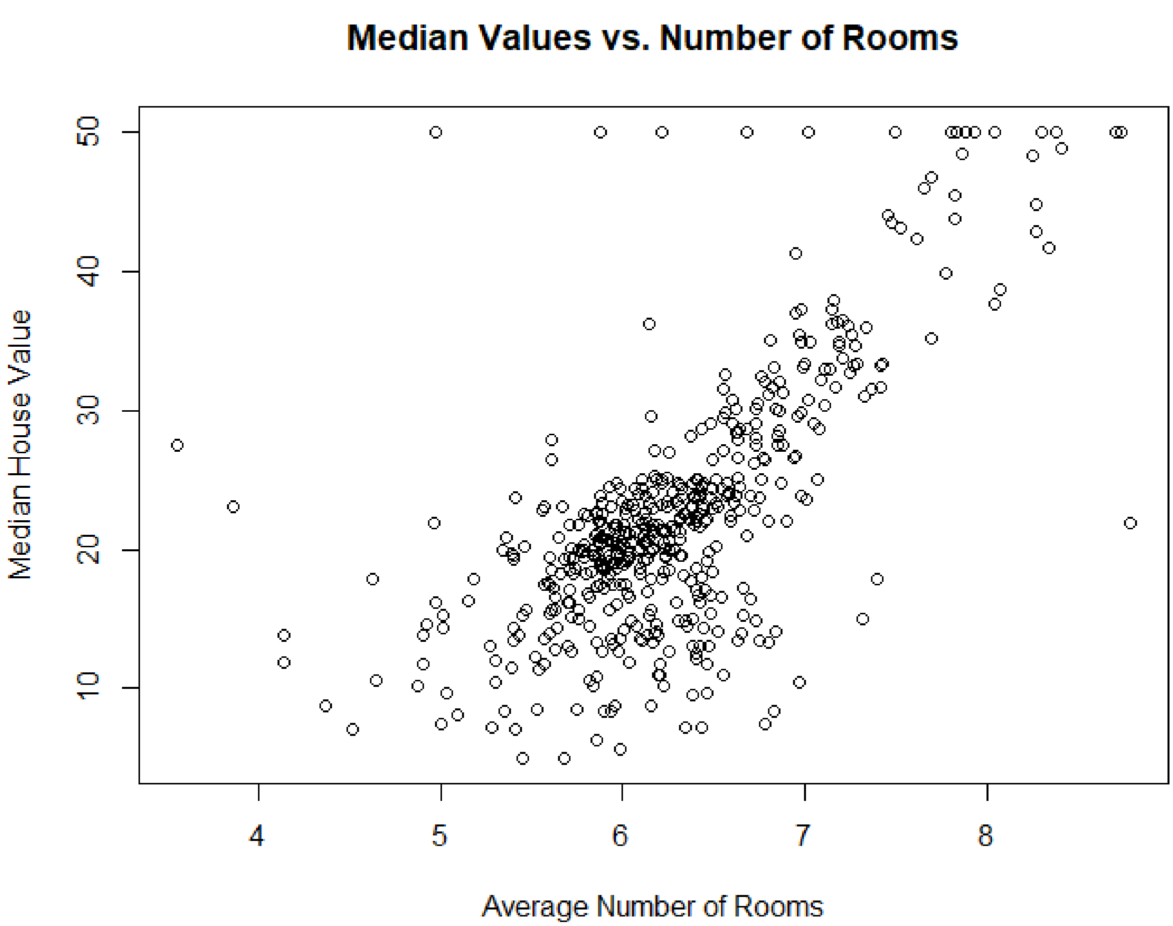
```
> cor(Boston) #Display correlation between variables
```

```
> cor(Boston) #Display correlation between variables
               crim          zn       indus         chas         nox          rm         age         dis          rad         tax
crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171 -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431
zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371  0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332
indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145 -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018
chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281  0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652
nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000 -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320
rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819  1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783
age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010 -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559
dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011  0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158
rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056 -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819
tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320 -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000
ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268 -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304
black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064  0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801
lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892 -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341
medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077  0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593
            ptratio       black      lstat       medv
crim      0.2899456 -0.38506394  0.4556215 -0.3883046
zn       -0.3916785  0.17552032 -0.4129946  0.3604453
indus     0.3832476 -0.35697654  0.6037997 -0.4837252
chas     -0.1215152  0.04878848 -0.0539293  0.1752602
nox       0.1889327 -0.38005064  0.5908789 -0.4273208
rm       -0.3555015  0.12806864 -0.6138083  0.6953599
age       0.2615150 -0.27353398  0.6023385 -0.3769546
dis      -0.2324705  0.29151167 -0.4969958  0.2499287
rad       0.4647412 -0.44441282  0.4886763 -0.3816262
tax       0.4608530 -0.44180801  0.5439934 -0.4685359
ptratio   1.0000000 -0.17738330  0.3740443 -0.5077867
black    -0.1773833  1.00000000 -0.3660869  0.3334608
lstat     0.3740443 -0.36608690  1.0000000 -0.7376627
medv     -0.5077867  0.33346082 -0.7376627  1.0000000
>
```

4. Choosing Variables

As I was interested to see what factors and how strongly influence the median house values, I chose ***rm*** variable (average number of rooms) as a predictor variable and ***medv*** (median house values) as the response variable. The previous step demonstrated that the rm variable had the strongest positive correlation with the median house values. To confirm my choice, I constructed a scatter plot of these two variables:

```
> plot(rm, medv, main="Median Values vs. Number of Rooms", xlab="Number of rooms"
, ylab="Median House Value")
```



This plot suggests potential linear relationship between the average number of rooms and the median house values, however it also shows some outliers that might influence the regression analysis. There are several data points along the maximum 50.0 mark looking as if there was a limit on possible house values, or if 50.0 was used as a default value during data collection or consolidation.

5. Simple Linear Regression Model and Its Discussion

The exploratory data analysis led me to the decision to choose ***rm*** and ***medv*** as independent and dependent variables respectively. So, my hypotheses are as follow:

> H_0: There is no relationship between the average number or rooms (rm) and the median values of the owner-occupied homes (medv).
>
> H_1: There is some relationship between the average number or rooms (rm) and the median values of the owner-occupied homes (medv).

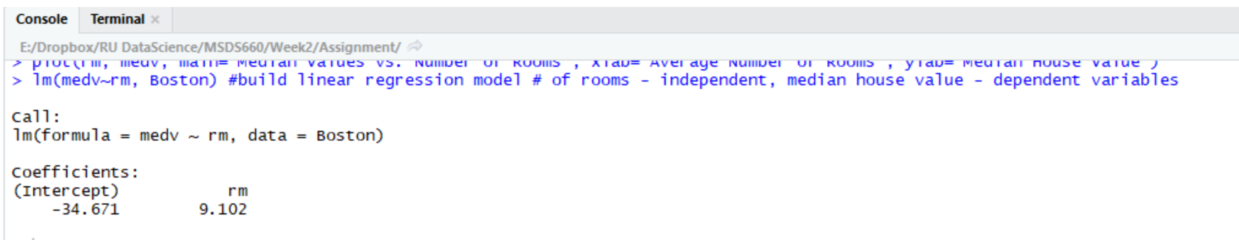
Simple Linear Regression

I used lm() command to build a simple linear regression model using these two variables.

```
> lm(medv~rm, Boston) #build linear regression model # of rooms - independent, me
dian house value - dependent variables

Call:
lm(formula = medv ~ rm, data = Boston)

Coefficients:
(Intercept)           rm
    -34.671        9.102
```



```
Console   Terminal
E:/Dropbox/RU DataScience/MSDS660/Week2/Assignment/
> lm(medv~rm, Boston) #build linear regression model # of rooms - independent, median house value - dependent variables

Call:
lm(formula = medv ~ rm, data = Boston)

Coefficients:
(Intercept)           rm
    -34.671        9.102
```

It resulted in a model that fit the data with a liner equation with intercept coefficient -34.671 and rm coefficient of 9.102 :

$$\textbf{\textit{medv}} = -34.671 + 9.102 \times \textbf{\textit{rm}}$$

Summary(lm_model) provided more detailed information about the model with the following output:

```
> summary(lm_model)#Display detailed info about the model

Call:
lm(formula = medv ~ rm, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max
-23.346  -2.547   0.090   2.986  39.433

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,	Adjusted R-squared:  0.4825
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

Simple Linear Regression

```
Console   Terminal ×
E:/Dropbox/RU DataScience/MSDS660/Week2/Assignment/
> lm_model<-lm(medv~rm, Boston) #Save the model output as an object
> summary(lm_model)#Display detailed info about the model

Call:
lm(formula = medv ~ rm, data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-23.346  -2.547   0.090   2.986  39.433 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -34.671      2.650  -13.08   <2e-16 ***
rm             9.102      0.419   21.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.616 on 504 degrees of freedom
Multiple R-squared:  0.4835,	Adjusted R-squared:  0.4825 
F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16

> |
```

The overall residuals are quite high, ranging from -23.346 to 39.433, however the median residual is 0.090 with IQR -2.547 to 2.986. It suggests, that the overall low accuracy of the current model is mostly due to some extreme outliers in the training data. Residual standard error, which measures quality of the linear regression fit, is 6.616 on 504 degrees of freedom. So, on average, the actual median values deviate 6.616 from the regression line.

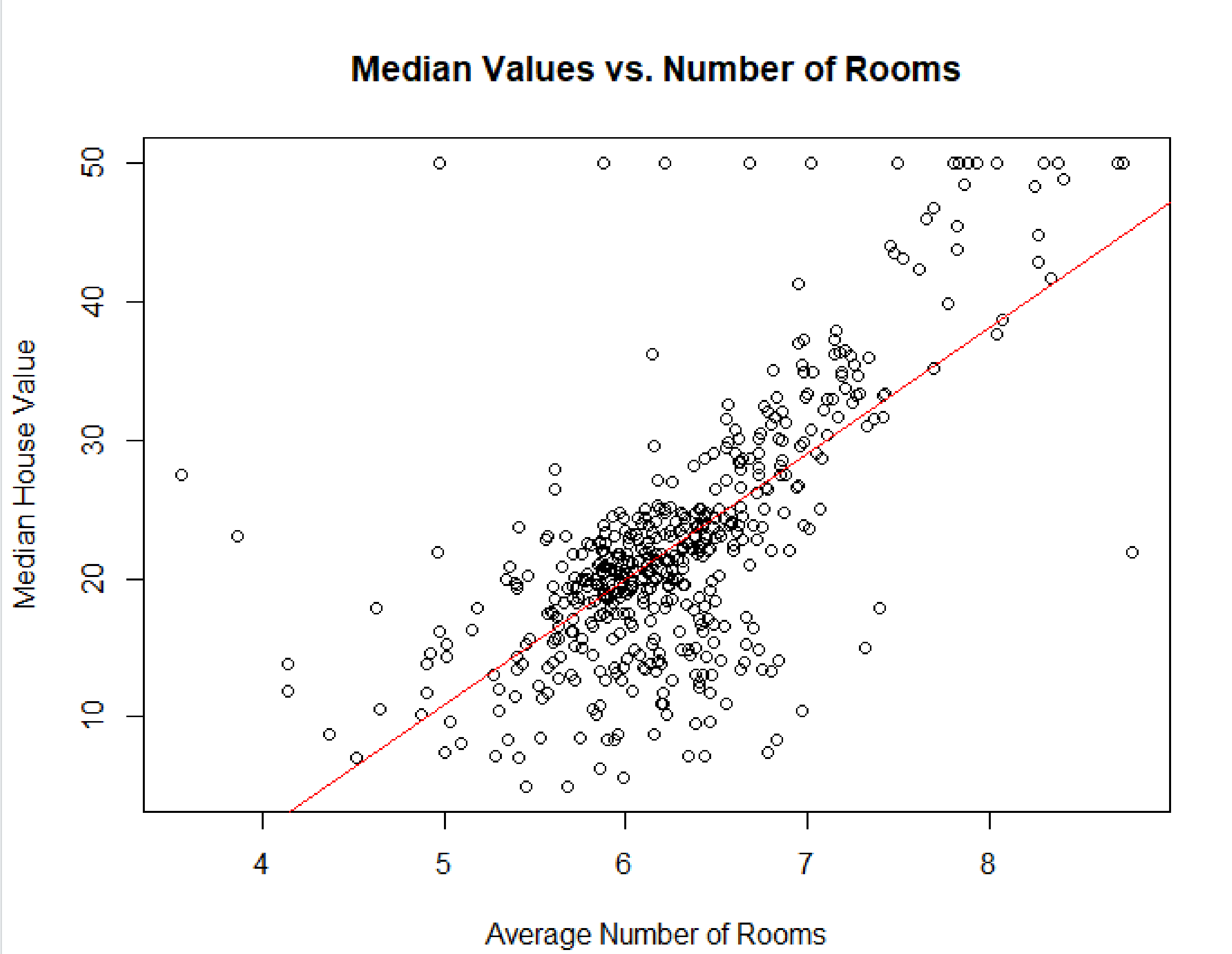
The coefficient in our model is equal to 9.102 with the standard error of 0.419. It means that, according to the model, every additional room in the house adds about 9.102 thousand to the median value of the house with a standard error of +/- 0.419 thousand.

The t-value is relatively large (21.72) and much larger than the standard error. The resulting p-value is small 2e-16 and allows us to **reject the null hypothesis in favor of the alternative hypothesis**. In fact, there is a linear relationship between the average number of rooms per dwelling (rm) and the median house values (medv). The F-statistic is also significantly large 471.8 on 1 and 504 degrees of freedom to reject the null hypothesis.

R-squared, which measures how well the model is fitting existing data, for the model is 0.4835. It means that roughly 48% of changes in the response variable (median home values) can be explained by the changes in the predictor variable (average number of rooms). The adjusted R-square is 0.4825. It means that after taking into consideration the degree of freedom, we still can explain only about 48% of variations in the response variable using the variations in our predictor.

To visually present the regression model, I added the regression line on top on the scatter plot using the following command:

```
> abline(lm(medv~rm), col="red") #Add regression line on top of the scatter plot
```

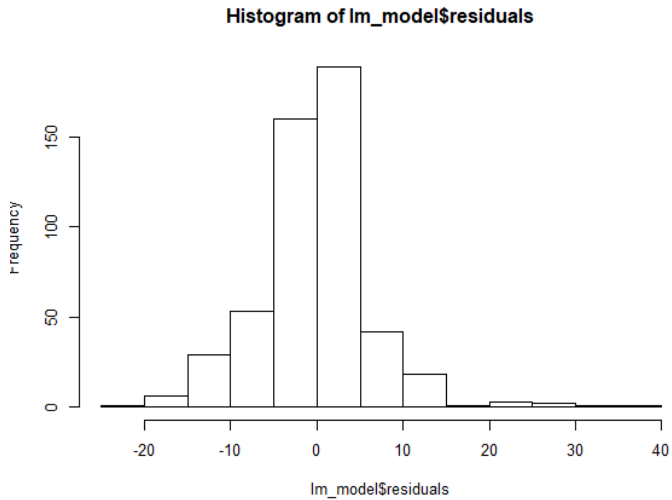


6. Model Diagnostic

In order for the resulting model to be useful, it needs to conform to the assumptions of linear regression:

a) Linearity: ***medv***=- 34.671 + 9.102 x ***rm.*** The resulting regression model is linear in parameters. The assumption holds.
b) Normal distribution of the residuals.
   The shape of the distribution of the residuals can be checked using a histogram.

```
> hist(lm_model$residuals) #histogram of residuals to check for normality
```



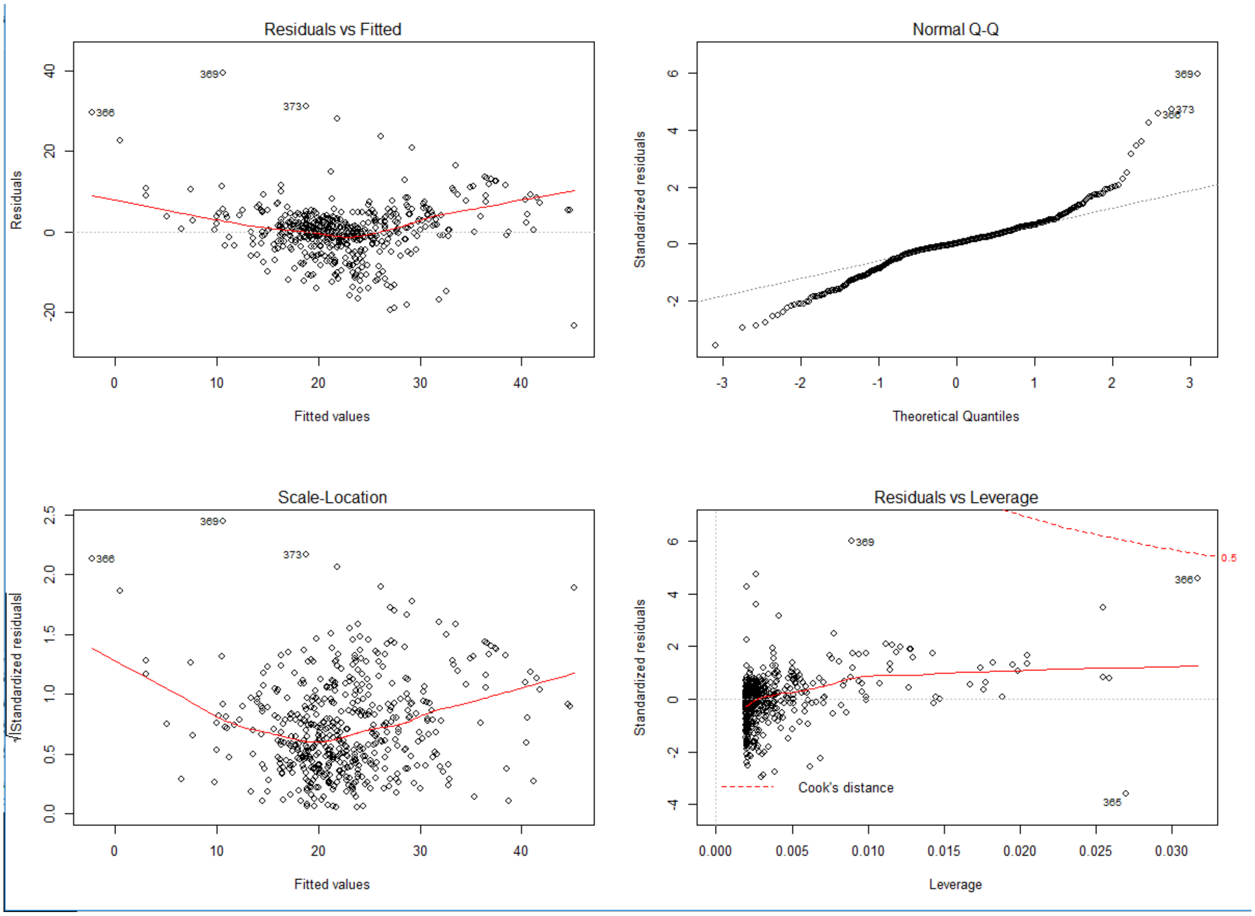
This histogram suggests that the distribution is close enough to normal to justify the use of the model, however it is right skewed, which can decrease its predictive power. The model might need to be adjusted.

c) Homoscedasticity of residuals or equal variance

The following plots are also useful for model evaluation – Residuals vs. Fitted, Normal Q-Q plot, Scale Location and Residuals vs. Leverage.

```
> plot(lm_model) #Fit information
```

Simple Linear Regression



Residuals vs. Fitted checks for homogeneity of the variance and the linear relation. The red pattern line shows that as fitted values increase, the residuals first slightly decrease, then increase again. The second graph checks for the normal distribution of the residuals, and it shows that it is more of an S-curve, than a straight line. The fourth graph shows the points that have too big impact on the regression coefficient and should be removed.

Overall, the assumption of the homoscedasticity of residuals or equal variance is not completely met.

<u>Conclusions:</u>

The model would benefit from recalibrating after making adjustments to the training dataset. Some outliers in the data have too much leverage and should be excluded.

The simple linear regression model is helpful in approximating the relationship between the average number of rooms in the dwellings in the area and the median house values, but its predictive power could be improved by including other variables into analysis