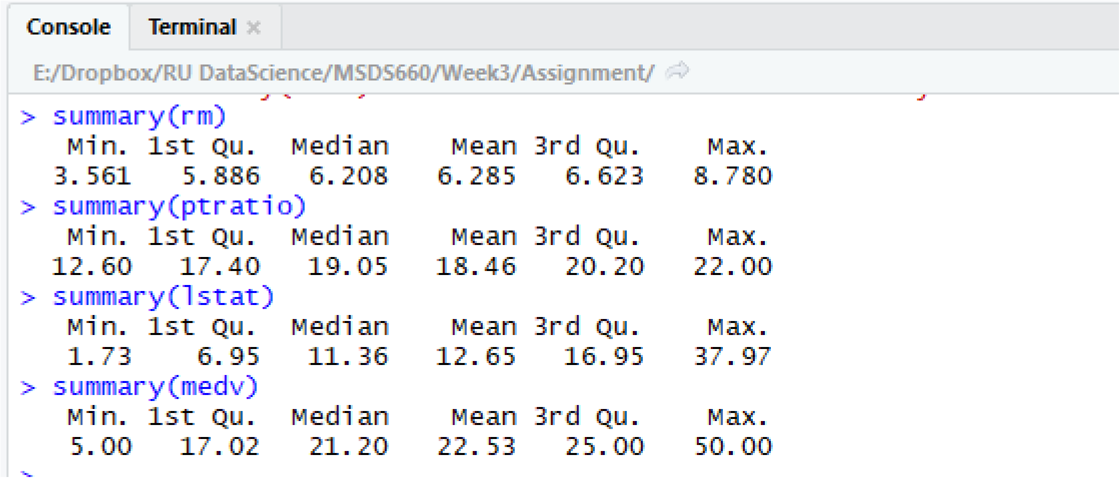
Contents

1. Choosing Predictor and Response Variables, Data Exploration

Continuing with my exploration of the Boston dataset, I chose to improve my simple linear regression model that I worked on last week by including more predictor variables into analysis. So, I focused on the same response variable **medv** (median value of owner-occupied homes in thousand of $) and three predictor variables: **rm** (average number of rooms per dwelling), **ptratio** (pupil-teacher ratio by town) and **lstat** (lower status of the population in percent).

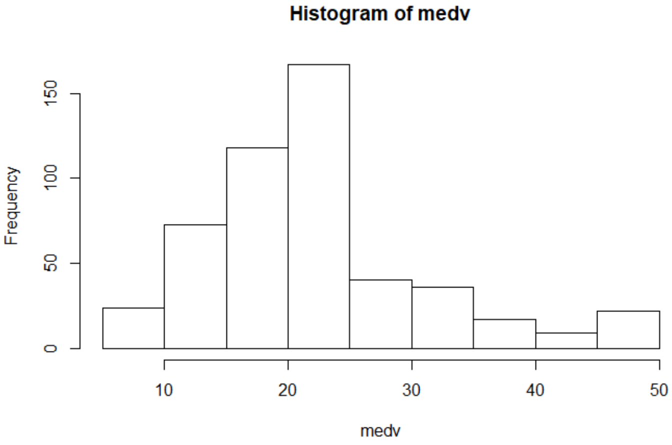
All four of these variables are continuous numeric variables. I used **summary()** function to look at the summary statistics of these four variables with the following output:



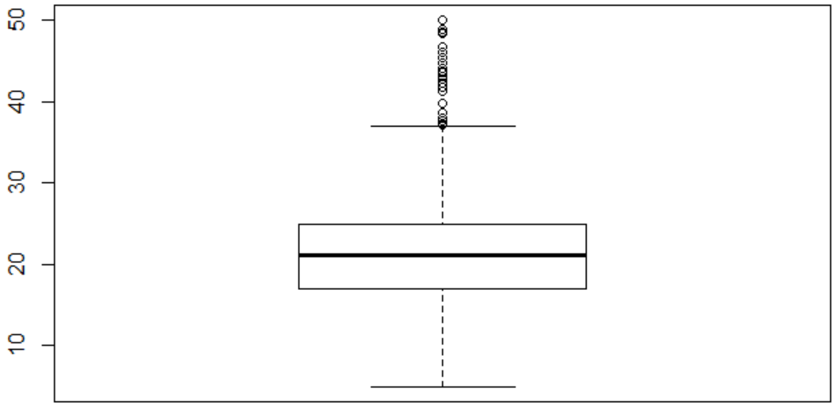
```
> summary(rm)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.561   5.886   6.208   6.285   6.623   8.780 
> summary(ptratio)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.60   17.40   19.05   18.46   20.20   22.00 
> summary(lstat)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   1.73    6.95   11.36   12.65   16.95   37.97 
> summary(medv)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.00   17.02   21.20   22.53   25.00   50.00 
```

The chosen response variable ***medv*** 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.

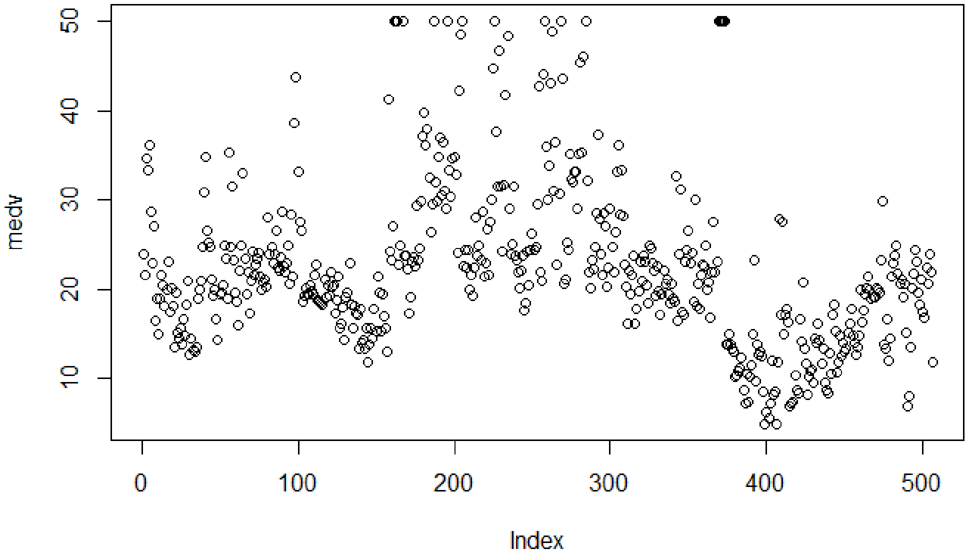


Creating a boxplot for median house values using ***boxplot(medv)*** only reinforced this conclusion:



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.

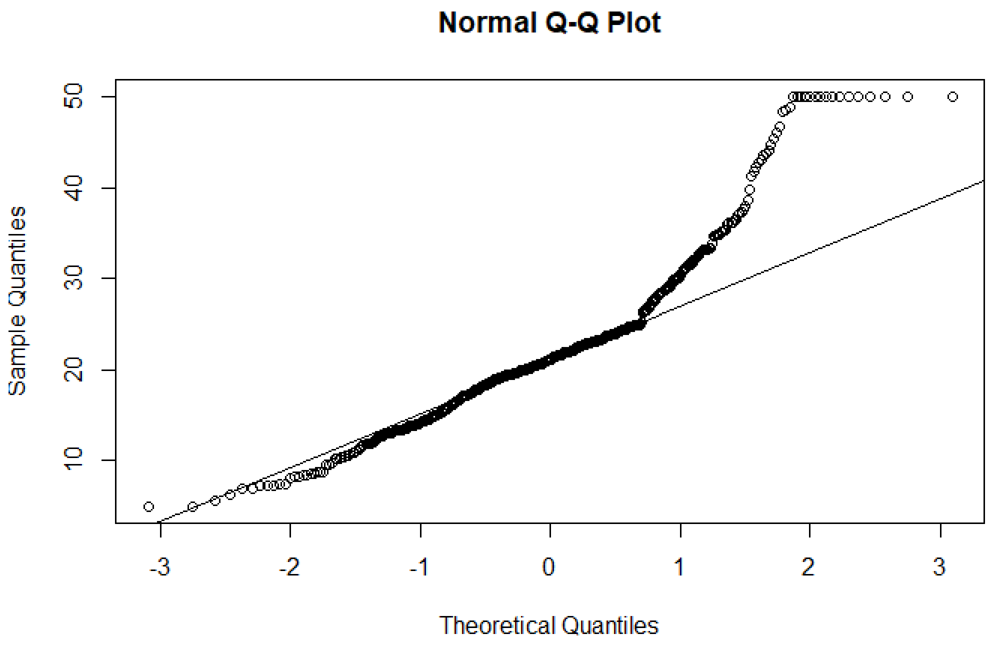


The plot showed a relatively high variability of data, and the standard deviation calculations showed 9.197 thousand dollars.

```
> 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)
```



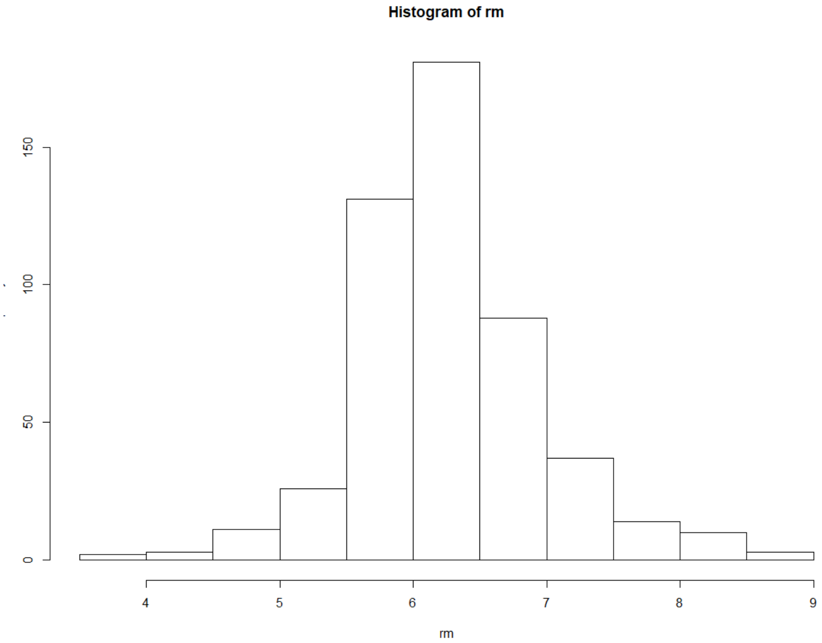
It showed that majority of the datapoints loosely follow theoretical Q-Q line, but the sample deviates to the top on the right side of the graph, which is typical for right-skewed distributions.

```
> shapiro.test(medv)  #Shapiro-Wilk normality test

        Shapiro-Wilk normality test

data:  medv
W = 0.91718, p-value = 4.941e-16
```

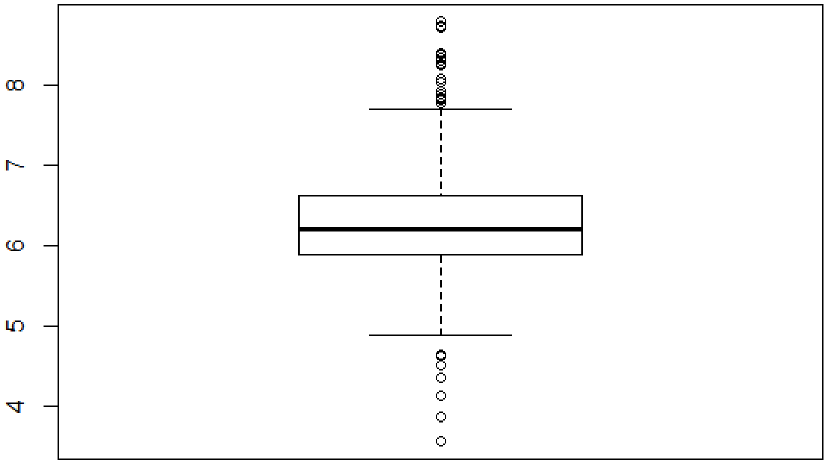
I used Shapiro-Wilk test to formally check for normality of the distribution, and it results presented above (very small p=value of 4.941e-16) show at this data is not strictly following the normal distribution.

I then explored my first potential explanatory variable **rm,** which represents the average number of rooms in a dwelling. It variates from a minimum of 3.561 rooms on average to a maximum of 8.78 average rooms with the median of 6.208 and close mean of 6.285. The middle 50% of dwellings have between 5.886 and 6.285 rooms on average.

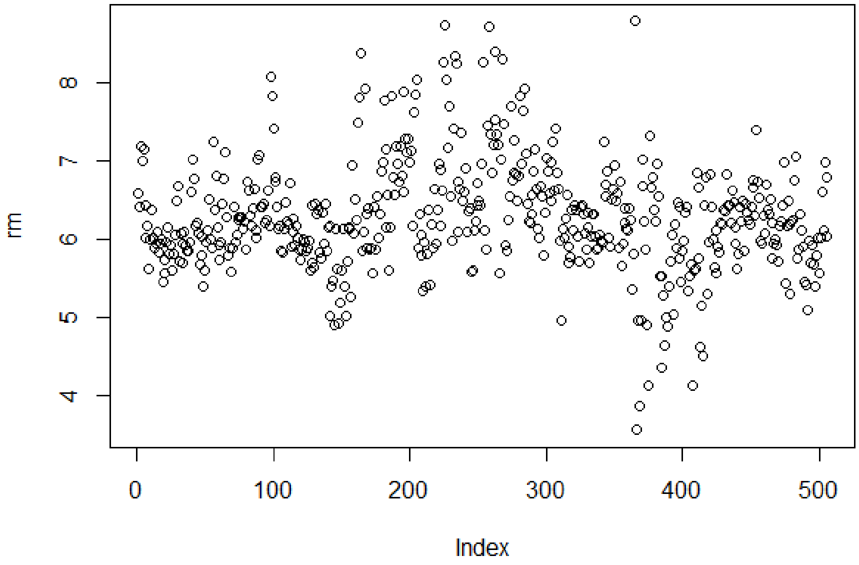
I first looked at the histogram, which showed an approximately bell-curved distribution with longer tails on both sides.



The box plot also showed that this data Is roughly symmetrical with narrow IQR and a few observations falling on both extremities – below 5 and above 7.5 rooms on average.

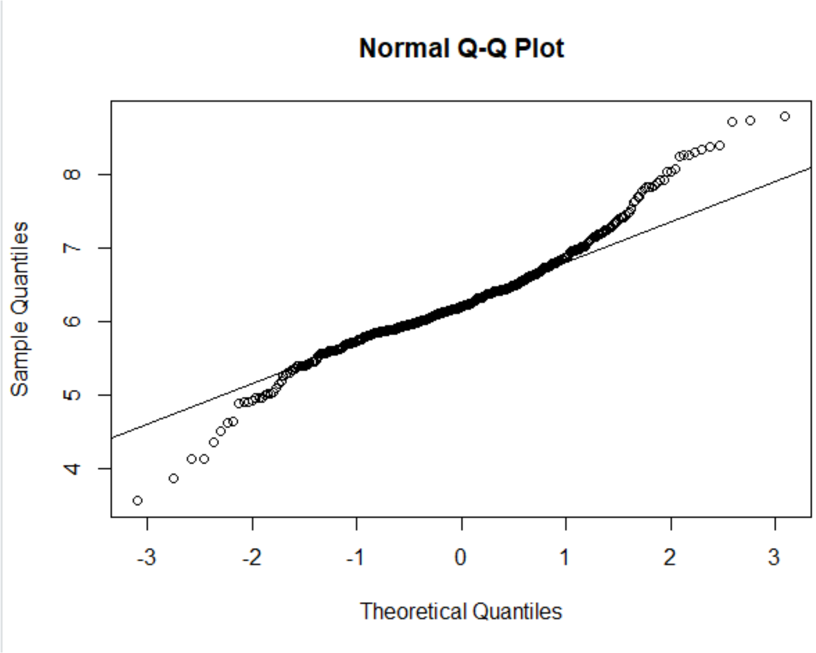


The following scatter plot helped visualize the observations mostly nesting in a band around the median with a few observation below and above.



```
> sd(rm) #Standard Deviation
[1] 0.7026171
```

The Q_Q plot below showed that the majority of the observations in the middle loosely follow the theoretical normal distribution, but the tails do not.



I used the Shapiro-Wilk test to check for normality:

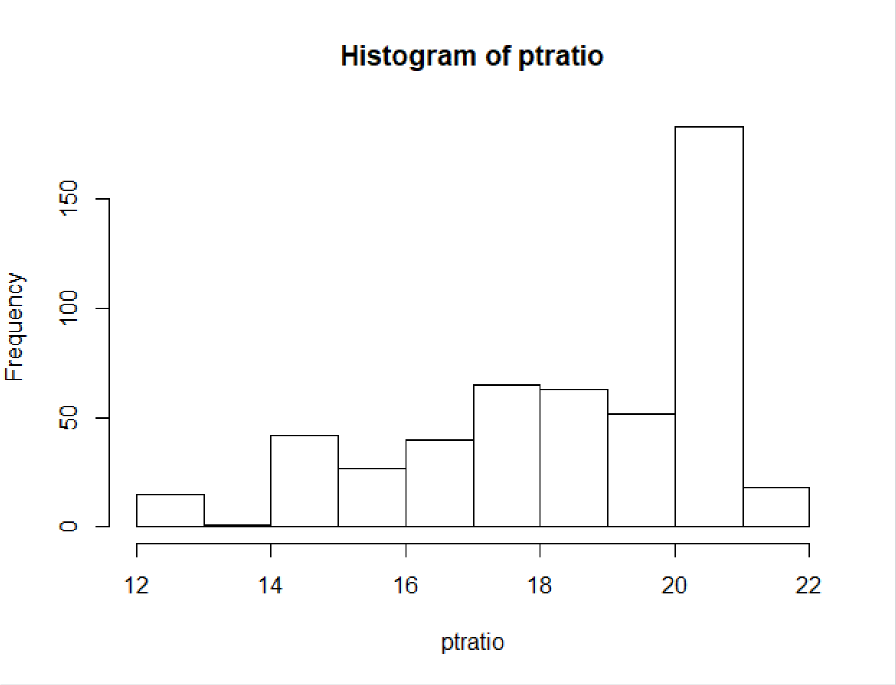
```
 > shapiro.test(rm)  #Shapiro-Wilk normality test
```

```
          Shapiro-Wilk normality test

data:  rm
W = 0.96087, p-value = 2.412e-10
```

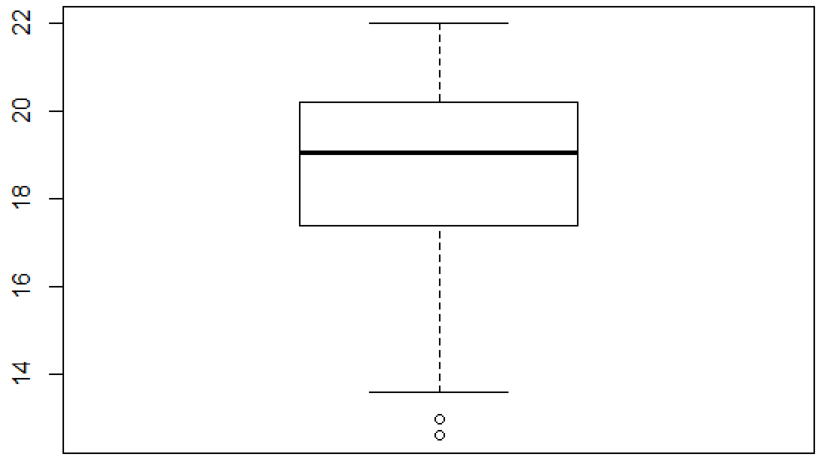
The p-values of 2.412e-10 is well above the significance level of 0.05, it means that we can reject the null hypothesis of normality. This data is not strictly following the normal distribution.

My next potential exploratory variable **ptratio** represents the pupils teacher ratio for each town. The minimum observation for this variable is 12.60 students per teacher and the maximum is 22.00 pupils per each teacher with IQR 17.40 – 20.20. The median value for this variable is 19.05 and the mean is 18.46 suggesting a skewed distribution.

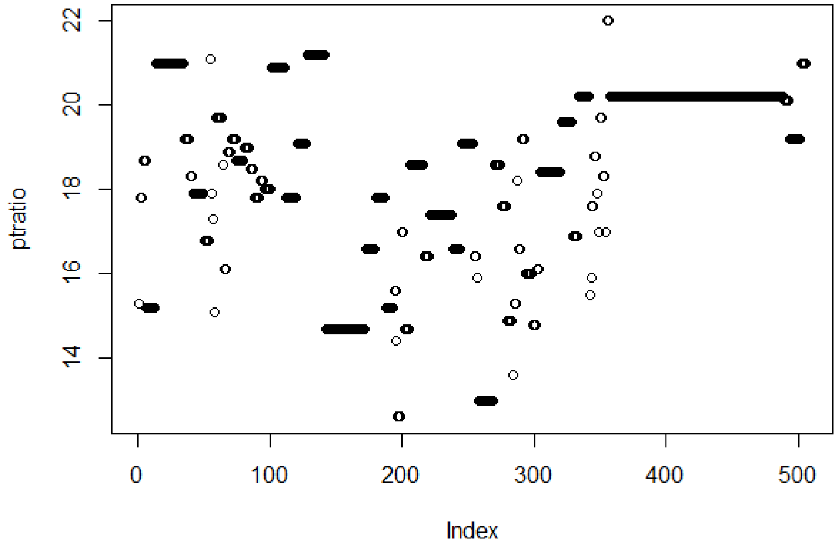
The histogram showed a unimodal distribution with a peak around 20-21 pupils per teacher.



The following boxplot confirmed that the data is not symmetrical, it is highly skewed to the left.



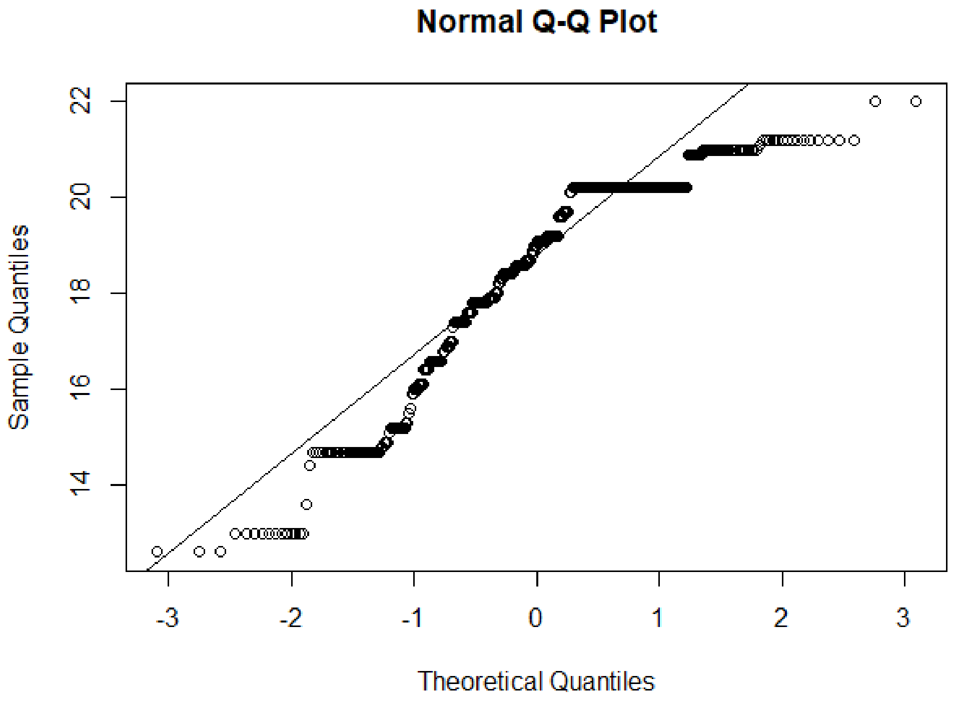
The following scatter plot suggests that some town might have a regulatory cap on the pupils per teacher ratios, affecting the distribution.



The standard deviation for ptratio is 2.164946.

```
> sd(ptratio) #Standard Deviation
[1] 2.164946
```

The following Q-Q plot shows considerable deviations of the normal distribution in the first and fourth quantiles



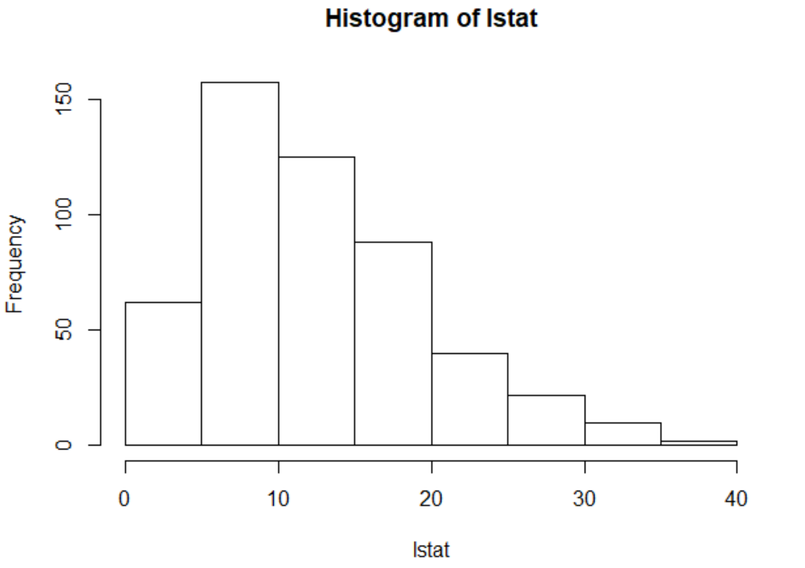
```
> shapiro.test(ptratio)

        Shapiro-Wilk normality test

data:  ptratio
W = 0.9036, p-value < 2.2e-16
```

P <0.05 Data not normal

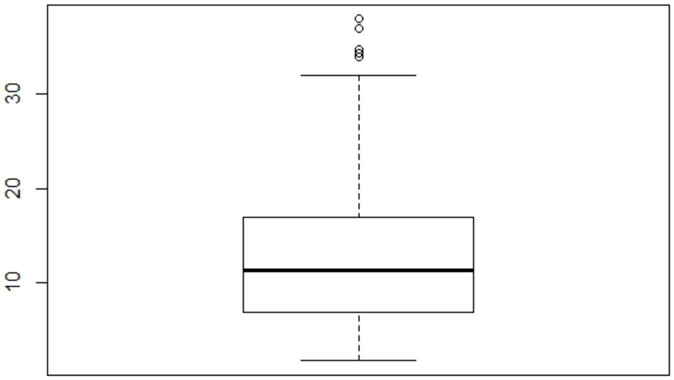
The above results of Shapiro-Wilk test confirms previous findings and returns p-value of 2.2e-16, which lets us reject the normality hypothesis for this data.

Lstat is the third explanatory variable that I would like to include in the model. Lstat is the percentage of people with low status in the area where median home prices information was collected. The Boston dataset includes observations ranging from 1.73 (min for lstat) to 37.97 (max for lastat) with the middle fifty percent of the observation in the range between 6.95 and 16.95 (IQR). The median of 11.36 is smaller than the mean 12.65 suggesting a skewed distribution.

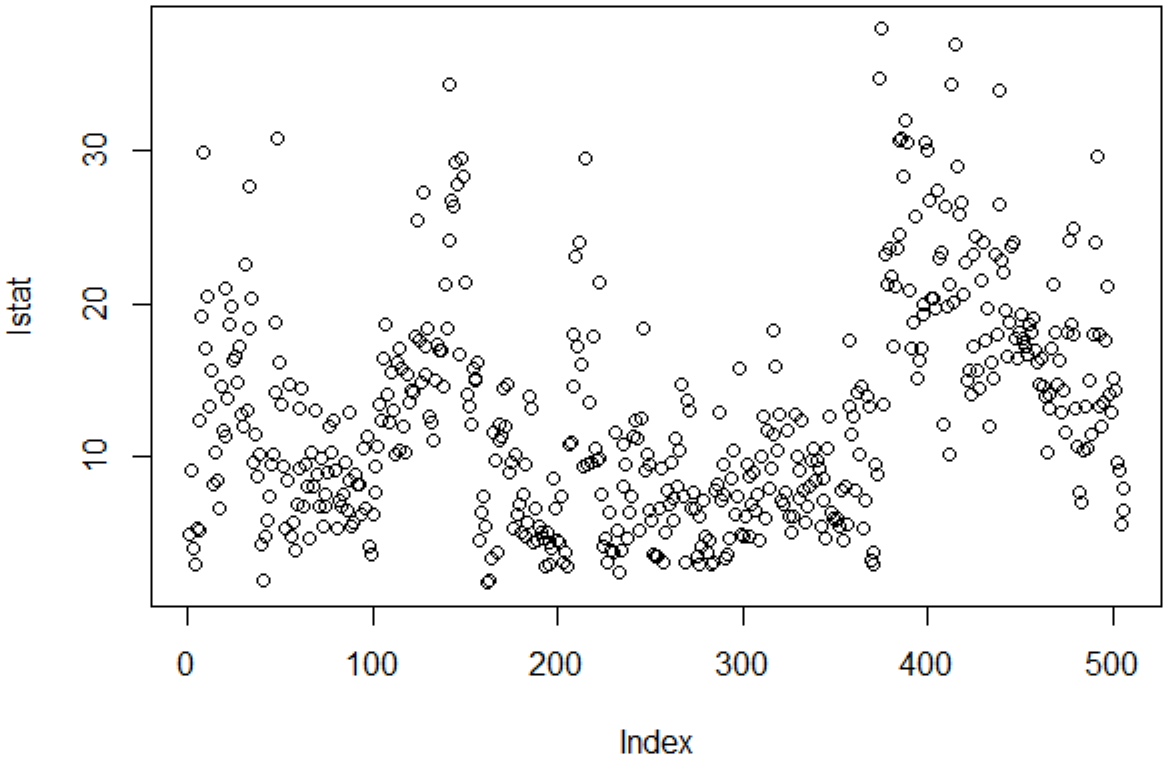
In fact, the following histogram confirms an approximately bell-shaped distribution highly skewed to the right because of the observation for towns with very high percentages of low status population.



This lack of symmetry in the distribution is also apparent when looking at the boxplot for lstat. Majority of the datapoint are below approximately 19.0, with a long tail of observations with much higher percentages of low status population.



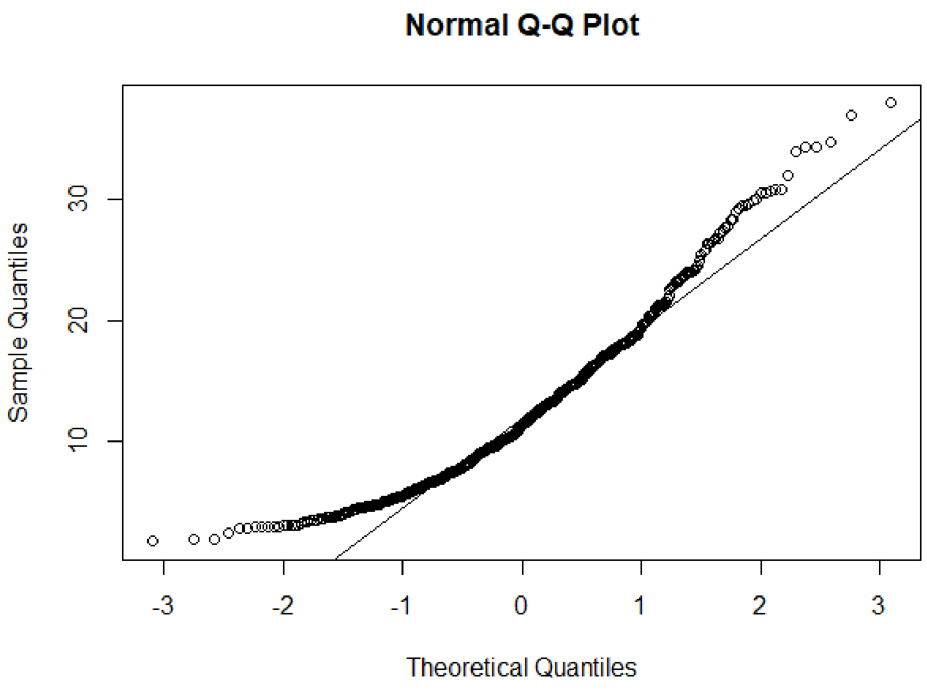
The scatter plot also shows a high number of observations for lstat above 19.



Overall standard deviation for this variable is 7.141062.

```
> sd(lstat) #Standard Deviation
[1] 7.141062
```

The following Q-Q plot shows that while the middle of the data has a tendency to follow the normal distribution, there is significant deviation in the first and fourth quantile.

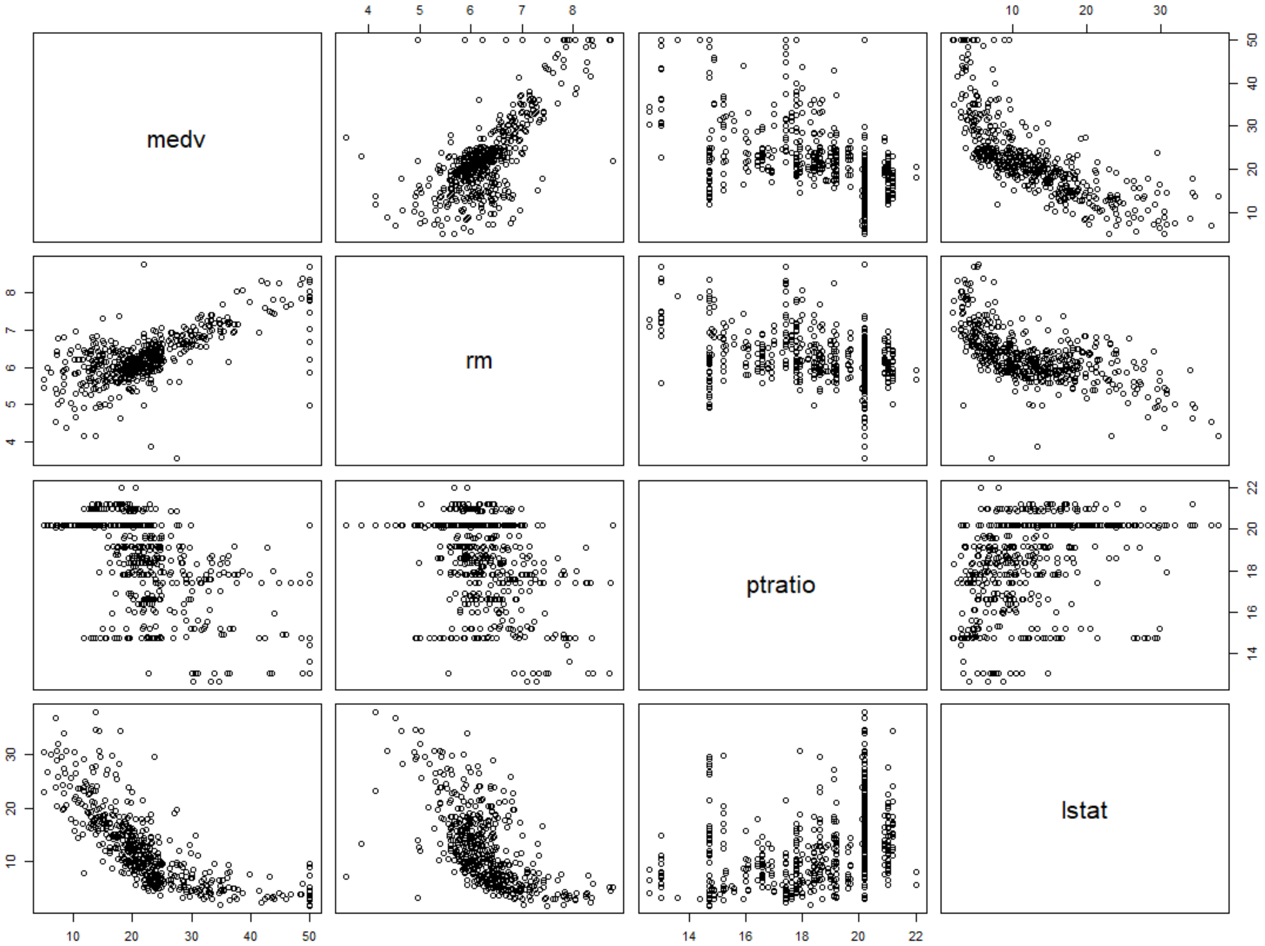


The Shapiro-Wilk test returns p-value of 8.287e-14 that allows us to reject the normality hypothesis for the overall distribution of lstat variable.

```
        Shapiro-Wilk normality test

data:  lstat
W = 0.93691, p-value = 8.287e-14
```

Next, I created a matrix of pairwise scatter plots for all four variables.

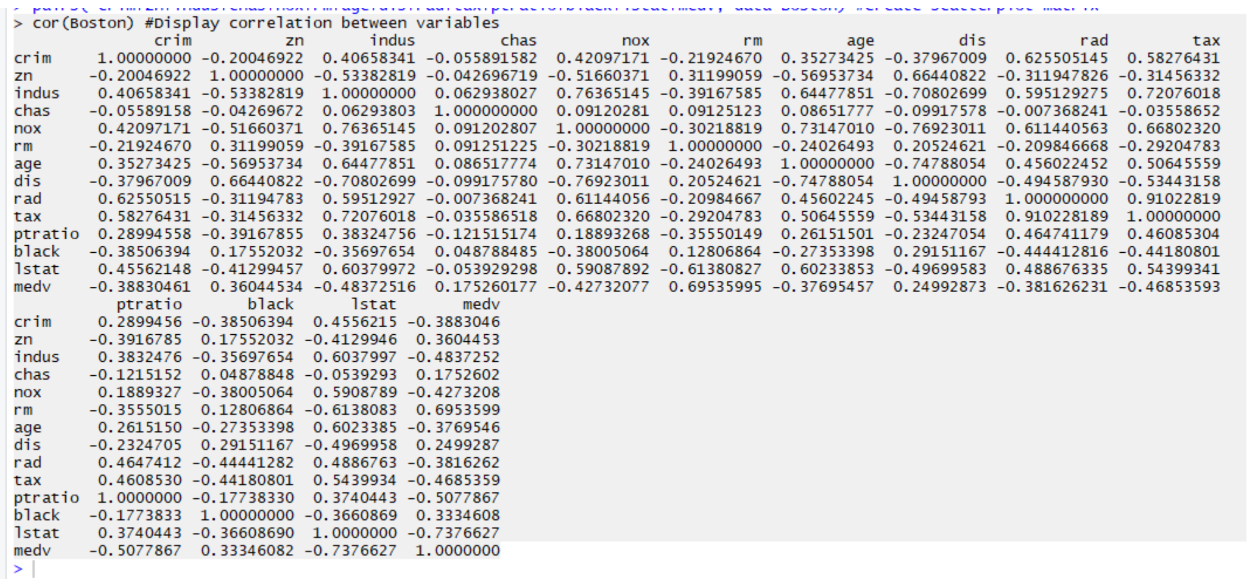
```
> pairs(~medv + rm + ptratio + lstat, data=Boston) #scatter plot matrix of four variables
```



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.

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
> 
```

Conclusion:

Overall, the above exploratory analysis allowed me to inspect data for possible errors in the dataset, missing values, outliers, possible clustering, unusual trends and asymmetric distributions etc. None of the chosen variable perfectly follows classical normal distribution, however my analysis did not find any unexpected patterns (ex. non-linear bivariate relationship, obvious signs of strong correlation between independent variables etc.) that would prevent using these four variables in a multiple linear regression model. At the same time, a high number of outliers could potentially decrease the predictive power of a model built using this data set.

This conclusion allowed me to proceed with my **main research question** – to find linear relationship between three independent variables (average number of rooms, pupils-teacher ratio and percentage of low status population) and the dependent variable (median house prices) allowing to predict house prices in the area. This research question is formally stated as a hypothesis in the following paragraph.

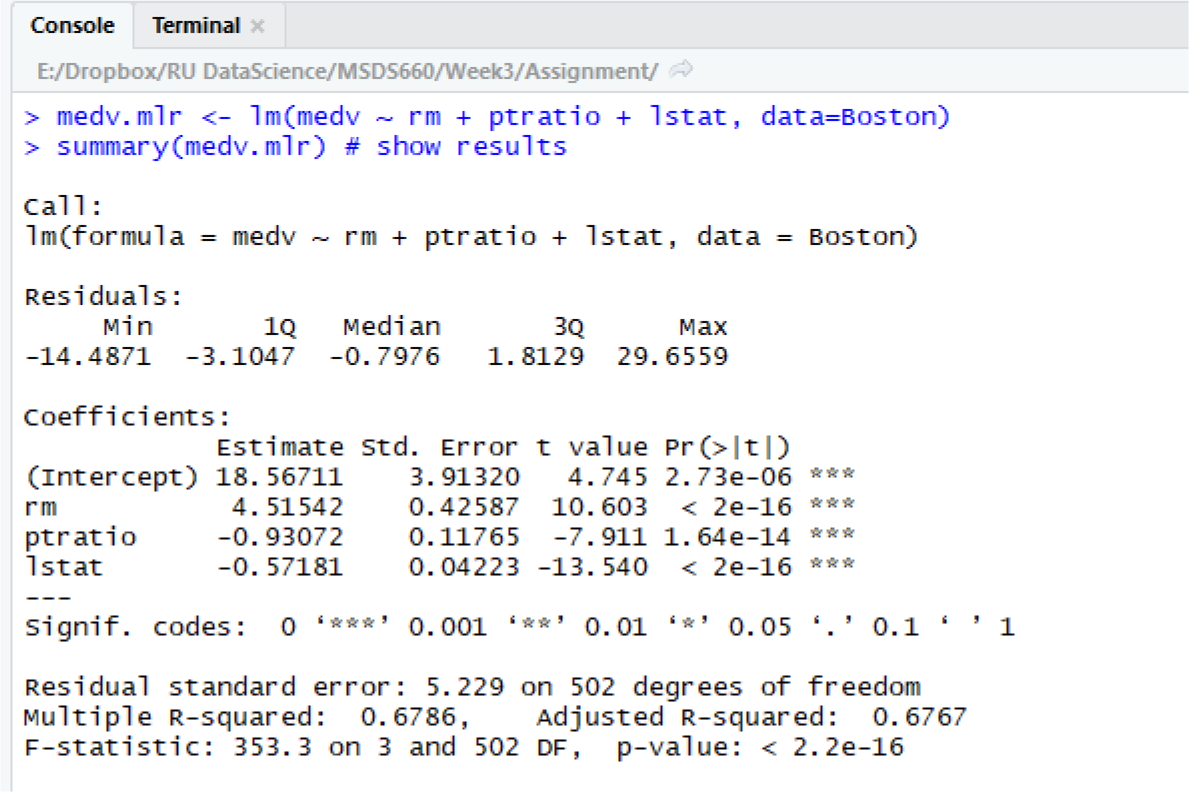
2. Multiple Linear Regression Model and Its Significance Testing

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

> H_0: There is no linear relationship between the average number or rooms (rm), pupil-teacher ratio (ptratio) percentage of lower status of population (lstat) and the median values of the owner-occupied homes (medv).
>
> H_1: There is a linear relationship between the average number or rooms (rm), pupil-teacher ratio (ptratio) percentage of lower status of population (lstat) and the median values of the owner-occupied homes (medv).

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

```
> medv.mlr <- lm(medv ~ rm + ptratio + lstat, data=Boston)
> summary(medv.mlr) # show results
```



```
> medv.mlr <- lm(medv ~ rm + ptratio + lstat, data=Boston)
> summary(medv.mlr) # show results

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

Residuals:
     Min       1Q   Median       3Q      Max 
-14.4871  -3.1047  -0.7976   1.8129  29.6559 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
rm           4.51542    0.42587  10.603  < 2e-16 ***
ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.229 on 502 degrees of freedom
Multiple R-squared:  0.6786,	Adjusted R-squared:  0.6767 
F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

It resulted in a model that fit the data with a liner equation with intercept coefficient 18.56711 and one positive coefficient for the rm variable (rm coefficient of 4.51542) and two negative coefficients (ptratio coefficient -0.93072 and lstat coefficient -0.57181).

medv= 18.56711 + 4.51542 x ***rm* - *0.93072 x ptratio* - *0.57181 x lstat***

The coefficient for the rm variable in this model is 4.51542, which means that according to the model, on average every additional room will add 4.51542 thousand to the median house values. An increase in pupil-teacher ratio in local schools by one student per teacher will decrease median house values by almost a thousand dollars (0.93072 thousand). Percentage of the population with low status also has a negative effect on median house values. Every additional percent will diminish median house values by 0.57181 thousand (or approximately 572 dollars).

Summary(medv.mlr) provided more detailed information about the model with the following output:

```
> medv.mlr <- lm(medv ~ rm + ptratio + lstat, data=Boston)
> summary(medv.mlr) # show results

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

Residuals:
     Min       1Q   Median       3Q      Max 
-14.4871  -3.1047  -0.7976   1.8129  29.6559 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 18.56711    3.91320   4.745 2.73e-06 ***
rm           4.51542    0.42587  10.603  < 2e-16 ***
ptratio     -0.93072    0.11765  -7.911 1.64e-14 ***
lstat       -0.57181    0.04223 -13.540  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.229 on 502 degrees of freedom
Multiple R-squared:  0.6786,	Adjusted R-squared:  0.6767 
F-statistic: 353.3 on 3 and 502 DF,  p-value: < 2.2e-16
```

The output shows that overall F statistic is 353.3 on 3 and 502 degrees of freedom, and p-value is significantly smaller than the significance level of 0.05 (p-value < 2.2e-16). This high value of F-statistic indicates that we can clearly reject the null hypothesis stating that there is no linear dependence between the explanatory variables and the median house prices. In fact, the alternative hypothesis is true: the explanatory variables in our model (rm, ptratio, lstat) collectively have a linear effect of median house values (dependent variable).

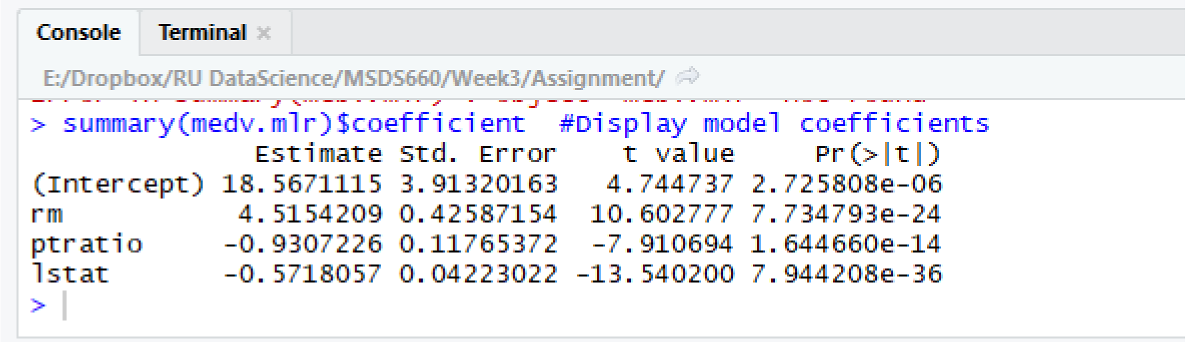
R-squared, which measures how well the model is fitting existing data, for the model is 0.6786. It means that roughly 68% of changes in the response variable (median home values) can be explained by the collective changes in the predictor variables (average number of rooms, pupil-teacher ratios, and % of low status population). The adjusted R-square is 0.6767. It means that after taking into consideration the degree of freedom, we can explain about 68% of variations in the response variable using the variations in our predictors.

The overall residuals are relatively high, ranging from -14.4871 to 29.6559, however the median residual is 0.7976 with IQR -3.1047 to 1.8129. It suggests, that high residuals of the current model might be caused by some outliers in the training data, which needs to be researched further to increase the predictive accuracy of the model. Residual standard error, which measures quality of the regression fit, is 5.229 on 502 degrees of freedom.

F-statistic of 353.3 told us that our predictor variables have an effect on median house values, however, the F statistics does not tell us which one of the predictor variables has a statistically significant effect. We need to test which individual predictor variables are important, so we need to look at t-statistics testing whether the corresponding regression coefficient is different from 0, and associated p-values for each variable:

Multiple Linear Regression

```
> summary(medv.mlr)$coefficient  #Display model coefficients
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 18.5671115 3.91320163   4.744737 2.725808e-06
rm           4.5154209 0.42587154  10.602777 7.734793e-24
ptratio     -0.9307226 0.11765372  -7.910694 1.644660e-14
lstat       -0.5718057 0.04223022 -13.540200 7.944208e-36
```



```
> summary(medv.mlr)$coefficient  #Display model coefficients
              Estimate Std. Error    t value     Pr(>|t|)
(Intercept) 18.5671115 3.91320163   4.744737 2.725808e-06
rm           4.5154209 0.42587154  10.602777 7.734793e-24
ptratio     -0.9307226 0.11765372  -7.910694 1.644660e-14
lstat       -0.5718057 0.04223022 -13.540200 7.944208e-36
> 
```

Hypotheses:

H0: rm coefficient = 0 (no linear relationship between rm and medv)

H1: rm coefficient ≠ 0 (linear relationship between rm and medv

The above result show that we can reject the null hypothesis in favor of the alternative hypothesis. P-value for t-test performed on rm variable is considerably smaller than the significance level (p-value=7.734793e-24 <0.05). It proves that there is a significant linear relationship between rm and medv variables.

Hypotheses:

H0: ptratio coefficient = 0 (no linear relationship between ptratio and medv)

H1: ptratio coefficient ≠ 0 (linear relationship between ptratio and medv)

The p-value associated with t-test results presented above is 1.644660e-14 which is much smaller that the significance level of 0.05. It means that we can reject the null hypothesis in favor of the alternative hypothesis. There is a significant linear relationship between ptratio and medv variables.

Hypotheses:

H0: lstat coefficient = 0 (no linear relationship between lstat and medv)

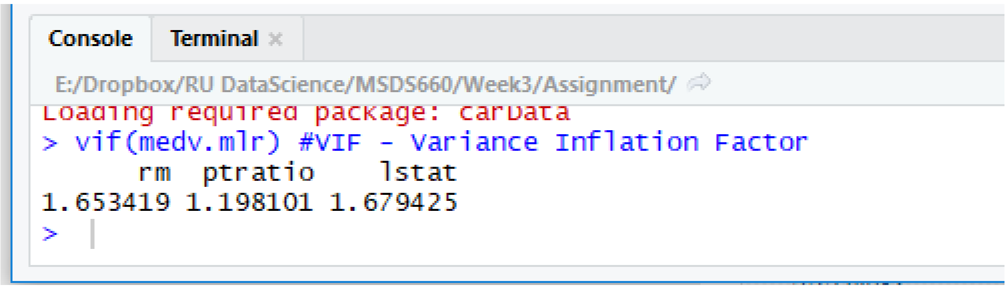
H1: lstat coefficient ≠ 0 (linear relationship between lstat and medv)

T-test performed on lstat variable yielded p-value of 7.944208e-36, which is well below the significance level of 0.05. It suggests that we can reject the null hypothesis in favor of the alternative hypothesis, which states that there is a significant linear relationship between lstat and medv variables.

So, the above results show that all three explanatory variables in our regression model have statistically significant linear relationship with the dependent variable medv.

However, we need to test our model for multicollinearity, which appears when one predictor variable in multiple regression model can be linearly predicted from the others with a substantial degree of accuracy.

```
> vif(medv.mlr) #VIF - Variance Inflation Factor
      rm  ptratio    lstat 
1.653419 1.198101 1.679425 
```



```
Loading required package: carData
> vif(medv.mlr) #VIF - Variance Inflation Factor
      rm  ptratio    lstat 
1.653419 1.198101 1.679425 
> 
```

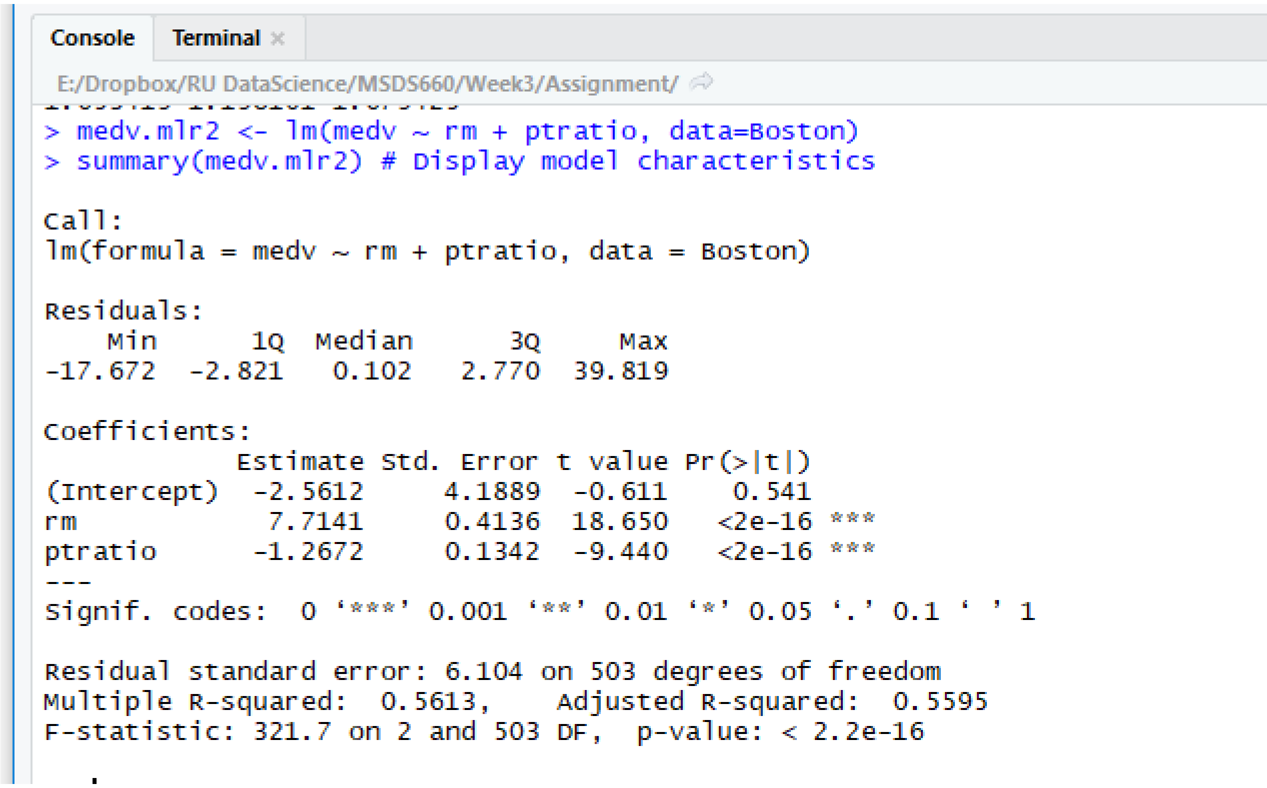
Variance Inflation Factor calculations using **vif(**) function presented above show that VIF for rm variable equals 1.653419, for ptratio 1.198101 and 1.679425 for lstat variable. VIF measures how much the variance of the estimated regression coefficient is inflated by the existence of correlation among the predictor variables in the model. VIFs close to 1 for all three model coefficients in our case mean that there is no significant correlation among the predictor variables in the model.

3. Comparing two multiple regression models

Next, I created a second multiple regression model using just two explanatory variables – rm (average room number) and ptratio (pupil-teacher ratio) and the same dependent variable -medv (median house values).

A summary of the second model is presented below:

```
> medv.mlr2 <- lm(medv ~ rm + ptratio, data=Boston)
> summary(medv.mlr2) # Display model characteristics

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

Residuals:
    Min      1Q  Median      3Q     Max 
-17.672  -2.821   0.102   2.770  39.819 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.5612     4.1889  -0.611    0.541    
rm            7.7141     0.4136  18.650   <2e-16 ***
ptratio      -1.2672     0.1342  -9.440   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.104 on 503 degrees of freedom
Multiple R-squared:  0.5613,	Adjusted R-squared:  0.5595 
F-statistic: 321.7 on 2 and 503 DF,  p-value: < 2.2e-16
```



```
> medv.mlr2 <- lm(medv ~ rm + ptratio, data=Boston)
> summary(medv.mlr2) # Display model characteristics

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

Residuals:
    Min      1Q  Median      3Q     Max 
-17.672  -2.821   0.102   2.770  39.819 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -2.5612     4.1889  -0.611    0.541    
rm            7.7141     0.4136  18.650   <2e-16 ***
ptratio      -1.2672     0.1342  -9.440   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.104 on 503 degrees of freedom
Multiple R-squared:  0.5613,	Adjusted R-squared:  0.5595 
F-statistic: 321.7 on 2 and 503 DF,  p-value: < 2.2e-16
```

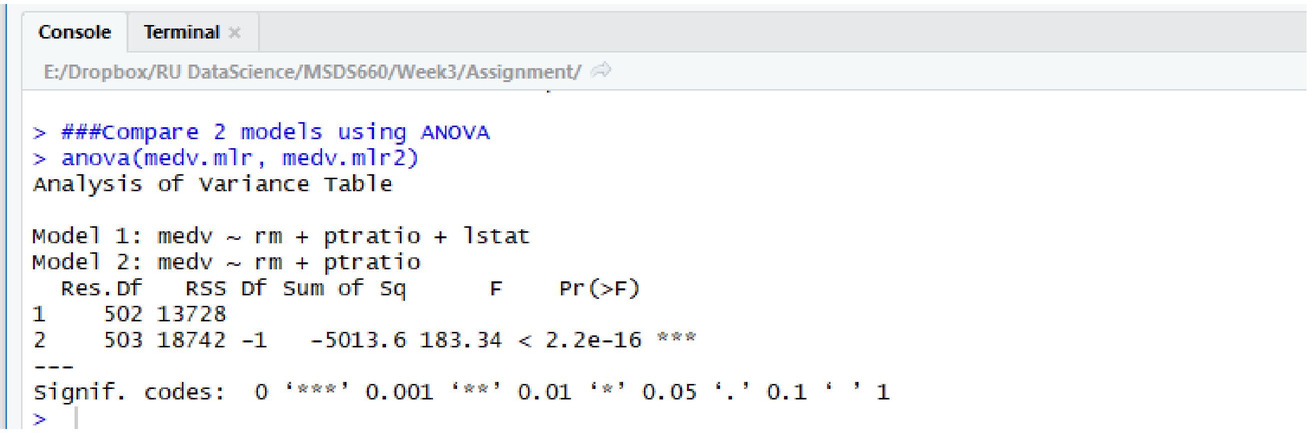
The second model with two predictor variables has an intercept coefficient of -2.5612, positive coefficient rm variable (7.7141) and a negative coefficient doe ptratio (-1.2672) and can be expressed using the following equation:

medv= -2.5612 +7.7141 x rm - 1.2672 x ptratio

It means that according to this model, an increase by one room on average is going to add 7.7141 thousand to the median house values in the area, and an increase in pupil to teacher ration by one student will diminish the median house values by 1.2672 thousand.

In order to compare those two models, I used ANOVA with the following output:

```
> ###Compare 2 models using ANOVA
> anova(medv.mlr, medv.mlr2)
Analysis of Variance Table

Model 1: medv ~ rm + ptratio + lstat
Model 2: medv ~ rm + ptratio
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1    502 13728                                  
2    503 18742 -1   -5013.6 183.34 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```



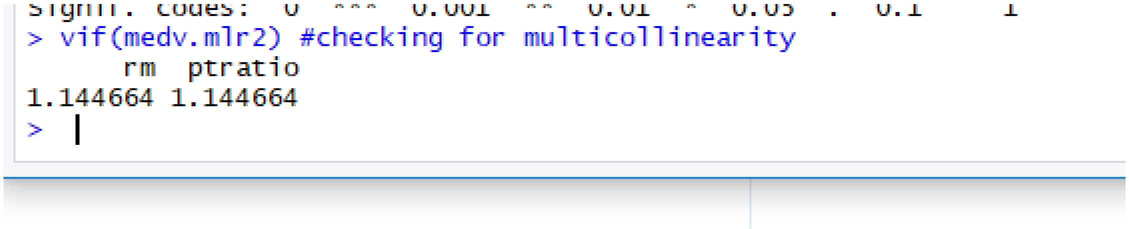
```
> ###Compare 2 models using ANOVA
> anova(medv.mlr, medv.mlr2)
Analysis of Variance Table

Model 1: medv ~ rm + ptratio + lstat
Model 2: medv ~ rm + ptratio
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)    
1    502 13728                                  
2    503 18742 -1   -5013.6 183.34 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

Anova() function compares the models to find the one which provides the best fit for the data. A very small p-value of 2.2e-16 shows that a simpler model with two predictor variables actually led to an improved fit over the model one for the existing data set and should be chosen over the more complicated model.

4. Model Diagnostic

Next, I checked the second model for potential multicollinearity using **vit**() function:

```
> vif(medv.mlr2) #checking for multicollinearity
      rm  ptratio 
1.144664 1.144664 
```



```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> vif(medv.mlr2) #checking for multicollinearity
      rm  ptratio 
1.144664 1.144664 
> 
```

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

a) Linearity:

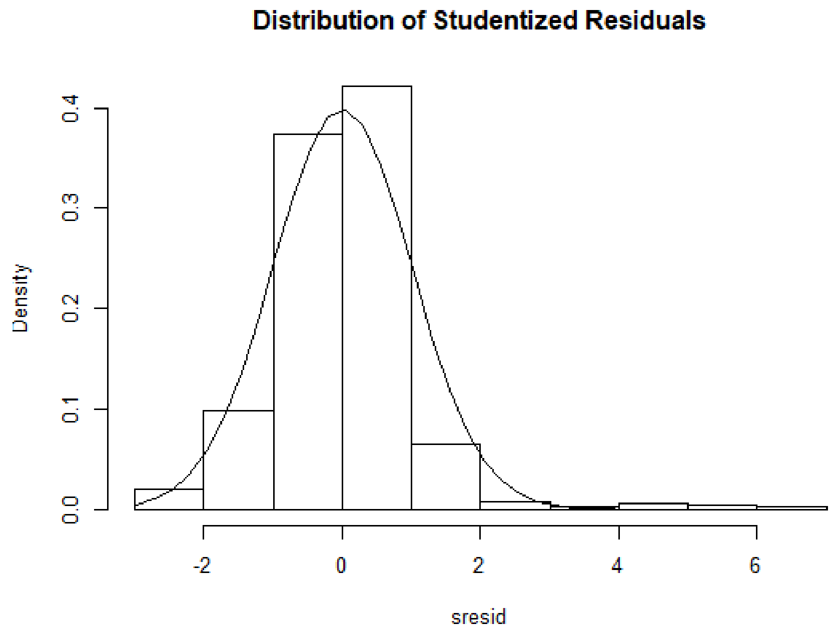
medv= -2.5612 +7.7141 x rm - 1.2672 x ptratio

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 graphs.

```
> # distribution of studentized residuals
> library(MASS)
> sresid <- studres(medv.mlr2)
> hist(sresid, freq=FALSE,
+      main="Distribution of Studentized Residuals")
> xfit<-seq(min(sresid),max(sresid),length=40)
> yfit<-dnorm(xfit)
> lines(xfit, yfit)
```

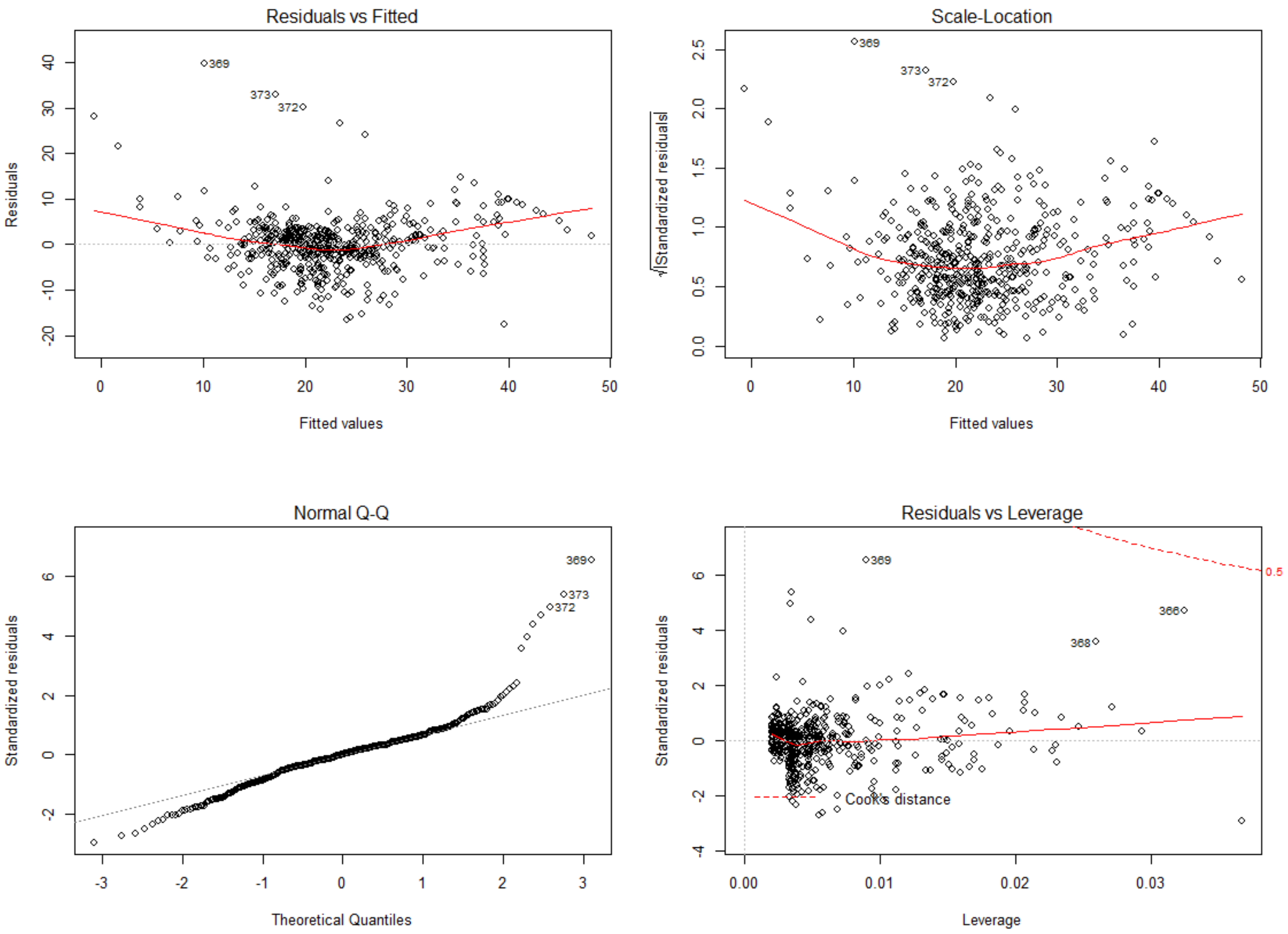


This histogram suggests that the distribution of the residuals loosely is close enough to normal distribution to justify the use of the model, however it has a longer right tail, which can decrease its predictive power. The reason might need to be researched further and the model might need to be adjusted.

c) Homoscedasticity of residuals or equal variance

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

```
> # diagnostic plots
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(medv.mlr2)
```



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.

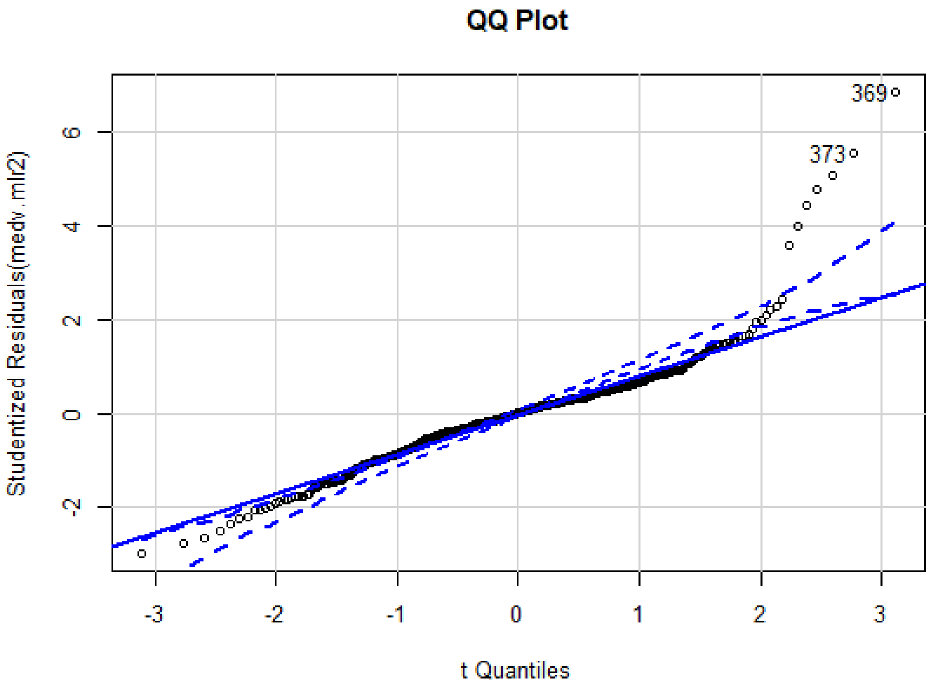
Overall, the assumption of the homoscedasticity of residuals or equal variance is not completely met.

d) Influential observations, outliers

To evaluate the model, it is useful to look at the points that that influence the model too much.

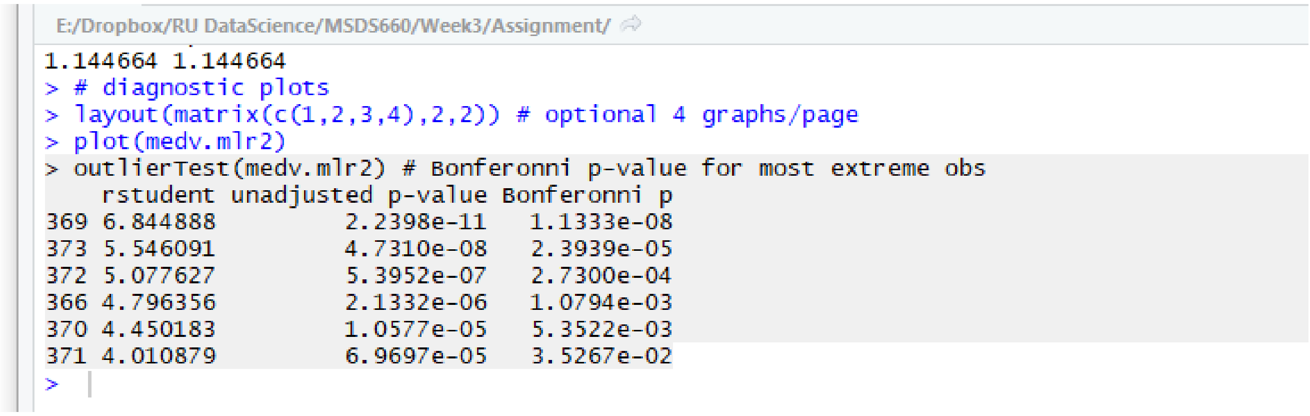
The following shows residuals for each variable and two observation that substantially deviate from the normal - 369 and 373.

```
> qqPlot(medv.mlr2, main="QQ Plot") #qq plot for studentized residuals
[1] 369 373
```



Assessing outliers:

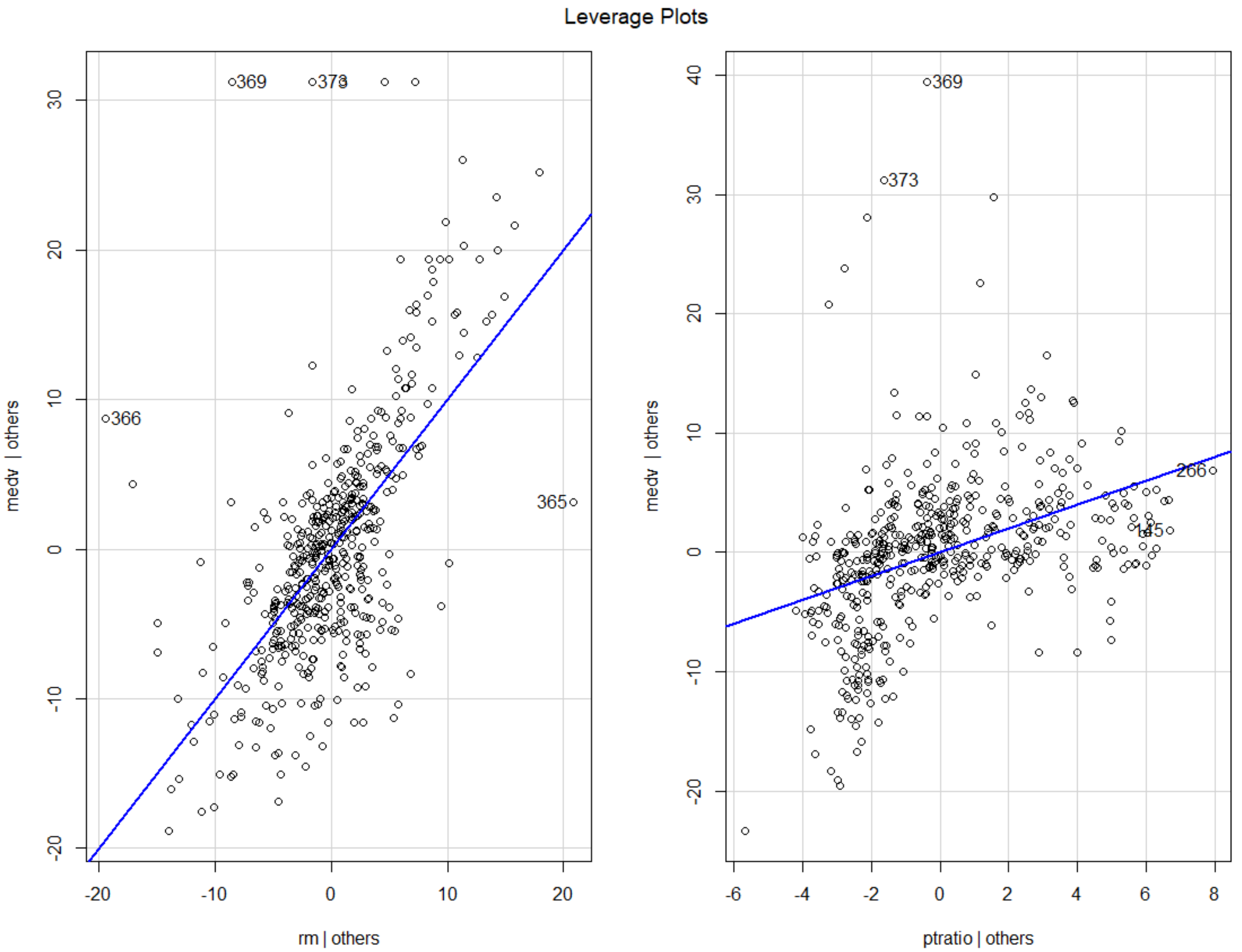
```
> outlierTest(medv.mlr2) # Bonferonni p-value for most extreme observations
    rstudent unadjusted p-value Bonferonni p
369 6.844888         2.2398e-11   1.1333e-08
373 5.546091         4.7310e-08   2.3939e-05
372 5.077627         5.3952e-07   2.7300e-04
366 4.796356         2.1332e-06   1.0794e-03
370 4.450183         1.0577e-05   5.3522e-03
371 4.010879         6.9697e-05   3.5267e-02
```



```
1.144664 1.144664
> # diagnostic plots
> layout(matrix(c(1,2,3,4),2,2)) # optional 4 graphs/page
> plot(medv.mlr2)
> outlierTest(medv.mlr2) # Bonferonni p-value for most extreme obs
    rstudent unadjusted p-value Bonferonni p
369 6.844888         2.2398e-11   1.1333e-08
373 5.546091         4.7310e-08   2.3939e-05
372 5.077627         5.3952e-07   2.7300e-04
366 4.796356         2.1332e-06   1.0794e-03
370 4.450183         1.0577e-05   5.3522e-03
371 4.010879         6.9697e-05   3.5267e-02
> 
```

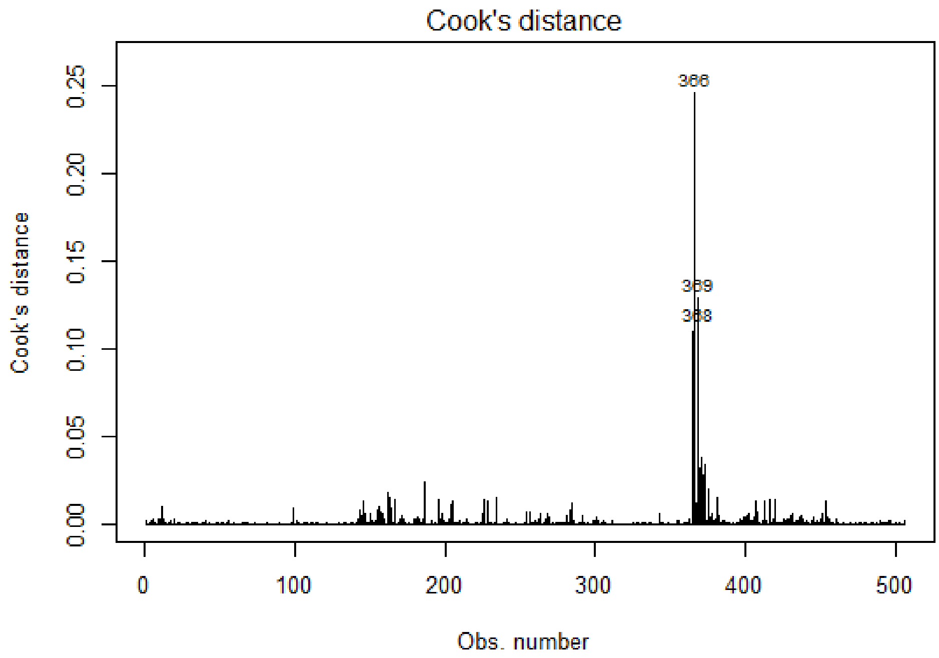
The following leverage plots visualize the effect of the outliers:

```
> leveragePlots(medv.mlr2) # leverage plots
```



I also created Cook's distance plot that estimates the influence of data points.

```
> # Influential Observations
> # Cook's D plot
> # identify D values > 4/(n-k-1)
> cutoff <- 4/((nrow(mtcars)-length(medv.mlr2$coefficients)-2))
> plot(medv.mlr2, which=4, cook.levels=cutoff)
```



Again, it clearly showed that at least three additional data points (366, 368, 369) need to be checked for validity.

Overall, the dataset contains several data points that have strong influence on the resulting multiple regression model. These points need to be further investigated and either eliminated, if they are found to be erroneous, or otherwise addressed in the modelling process. This can potentially normalize the residuals and increase predictive power of the model.

<u>Conclusions:</u>

- The multiple linear regression model has definite benefits over the single linear regression model built last week as it allowed to better explain variations in the dependent variables using variations in the independent variables.
- To my surprise according to ANOVA, a simpler model with two independent variables (rm and ptratio) showed a better fit result on the Boston dataset compared to my initial MLR model using three explanatory variables (rm, ptratio and lstat).
- The resulting multiple linear regression model is helpful in approximating the relationship between the average number of rooms in the dwellings in the area, average student to teacher ratios in local

schools and the median house values, but its predictive power could potentially be improved by including other variables into analysis, for example the average age of the housing in the area.

- Another potential way to improve the model is to look at the outliers in the data set that seem to have high influence on the resulting model. Excluding these data points from the training set could potentially improve precision and normalize residuals for the model.