Setup:

In this problem set you will need, at minimum, the following R packages.

```
# Load standard libraries
library(tidyverse)
```

```
## Warning: package 'dplyr' was built under R version 3.3.2
```

```
library(MASS) # Modern applied statistics functions
library(dplyr)
library(ggplot2)
```

Housing Values in Suburbs of Boston

In this problem we will use the Boston dataset that is available in the `MASS` package. This dataset contains information about median house value for 506 neighborhoods in Boston, MA. Load this data and use it to answer the following questions.

1. Describe the data and variables that are part of the `Boston` dataset. Tidy data as necessary.

```
#The following function gives the structure, data types and the data observations for each variable
str(Boston)
```

```
## '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 ...
```

```
#View the first few records of the dataset to get a glimse of the data
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

View the statistical summary of all the variables

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
```

```
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127
##       rad              tax          ptratio          black
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90
##      lstat            medv
##  Min.   : 1.73   Min.   : 5.00
##  1st Qu.: 6.95   1st Qu.:17.02
##  Median :11.36   Median :21.20
##  Mean   :12.65   Mean   :22.53
##  3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :37.97   Max.   :50.00
```

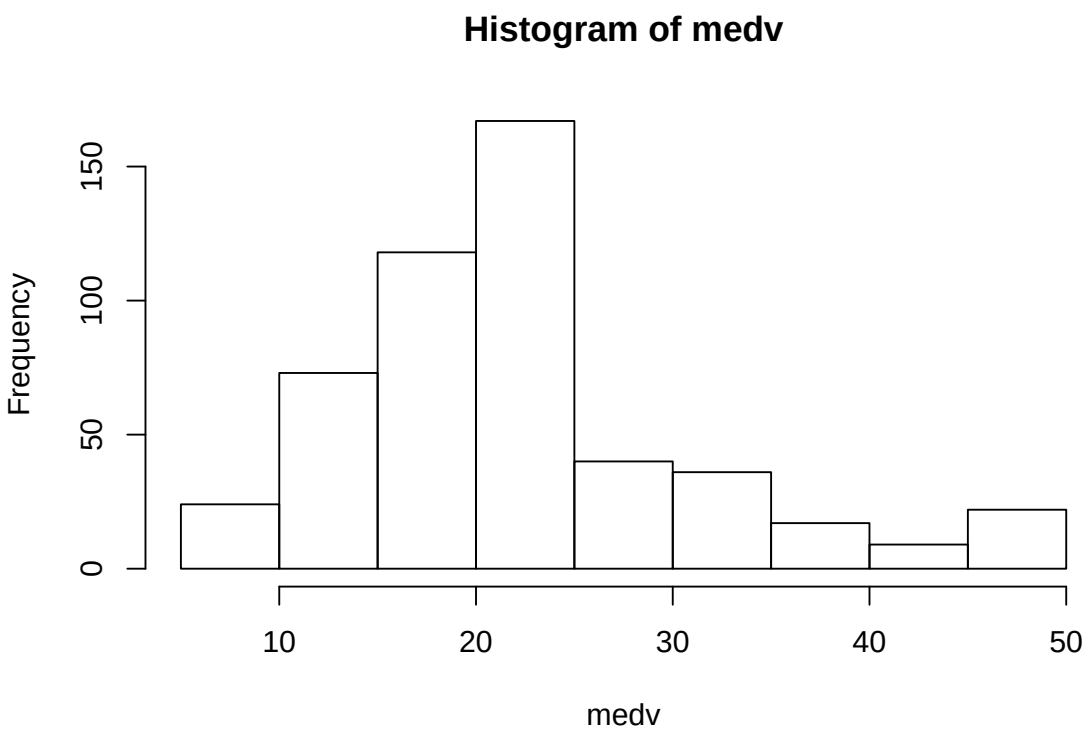
There are 14 variables in the Boston dataset. They are as follows.

crim - per capita crime rate / town zn - proportion of residential land zoned for lots over 25,000 sq.ft. indus - proportion of non-retail business acres per town. chas - dummy variable if tract bounds river nox - nitric 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 centres rad - index of accessibility to radial highways tax - full-value property-tax rate per $10,000 ptratio - pupil-teacher ratio by town black - proportion of blacks by town lstat - lower status of the population in % medv - Median value of owner-occupied homes in $1000's

2. Consider this data in context, what is the response variable of interest? Discuss how you think some of the possible predictor variables might be associated with this response.

   The response variable of interest in the dataset is "medv" which is the median value of owner-occupied homes in $1000's. By obervation this can be deducted but we will go ahead and check the normalcy of the Median value variable to check if it has a normalcy behcaviour overe the set of obervations

```
attach(Boston)
hist(medv)
```



From the histogram it is evident that the variable does not have a normalcy behaviour and hence can be considered as a response variable.

3. For each predictor, fit a simple linear regression model to predict the response. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

   First we will find the correlation between all the variables in the data set to have a glimpse of correlation between all the variables.

```
cor(Boston)
```

```
##                crim          zn       indus         chas         nox
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## black   -0.38506394  0.17552032 -0.35697654  0.048788485 -0.38005064
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                rm         age         dis          rad         tax
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431
```

```
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304
## black    0.12806864 -0.27353398  0.29151167 -0.444412816 -0.44180801
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341
## medv     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
```

From the results we can see that the lstat (lower status of the population in %), rm (average number of rooms per dwelling) and ptratio (pupil teacher reation per town) have the highest correlation to the response variable of medv. We will use these three variables to conduct further analysis and establish if there is any statistically significant association between them and the response variable.

```
#Fit a linear regresssion model for the 3 most significant variables
lmLstat <- lm(data = Boston, medv ~ lstat, na.action = na.exclude)
lmRm <- lm(data = Boston, medv ~ rm, na.action = na.exclude)
lmPtratio <- lm(data = Boston, medv ~ ptratio, na.action = na.exclude)

#Also fit linear models for other variables.
lmCrim <- lm(data = Boston, medv ~ crim, na.action = na.exclude)
lmZn <- lm(data = Boston, medv ~ zn, na.action = na.exclude)
lmIndus <- lm(data = Boston, medv ~ indus, na.ac = na.exclude)
lmchas <- lm(data = Boston, medv ~ chas, na.action= na.exclude)
lmNox <- lm(data = Boston, medv ~ nox, na.action= na.exclude)
lmAge <- lm(data = Boston, medv ~ age, na.action= na.exclude)
lmDis <- lm(data = Boston, medv ~ dis, na.action= na.exclude)
lmRad <- lm(data = Boston, medv ~ rad, na.action= na.exclude)
lmTax <- lm(data = Boston, medv ~ tax, na.action= na.exclude)
lmBlack <- lm(data = Boston, medv ~ black, na.action= na.exclude)
```

4. Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis $H_0 : \beta_j = 0$?

```
bostonRegModel <- lm(medv ~  rm +  ptratio + lstat + crim + zn + indus + chas + nox + age + dis + r

bostonRegModel
```

```
##
## Call:
## lm(formula = medv ~ rm + ptratio + lstat + crim + zn + indus +
##     chas + nox + age + dis + rad + tax + black, data = Boston)
##
## Coefficients:
## (Intercept)           rm      ptratio        lstat         crim
##   3.646e+01    3.810e+00   -9.527e-01   -5.248e-01   -1.080e-01
##          zn        indus         chas          nox          age
##   4.642e-02    2.056e-02    2.687e+00   -1.777e+01    6.922e-04
##         dis          rad          tax        black
##  -1.476e+00    3.060e-01   -1.233e-02    9.312e-03
```

```
summary(bostonRegModel)
```

```
##
## Call:
## lm(formula = medv ~ rm + ptratio + lstat + crim + zn + indus +
##     chas + nox + age + dis + rad + tax + black, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas         2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## black        9.312e-03  2.686e-03   3.467 0.000573 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

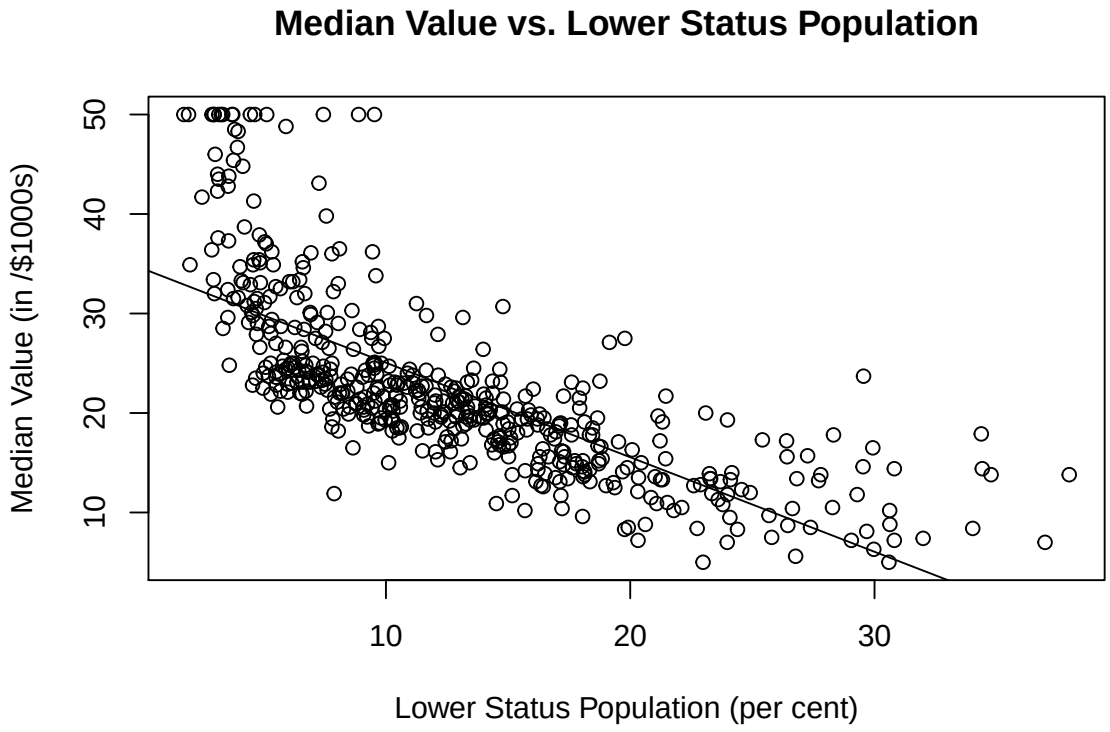
From the result we can see that the three variables of lstat, rm and ptraion have extremely low p values.

The variables zn and rad have the highest t statistic values and thus we will eliminate them or we can say for these two variables we can reject the null hypothesis.

We will plot the linear models for the three most statistically significant variables to show a graphical representation of the predictor and response variable

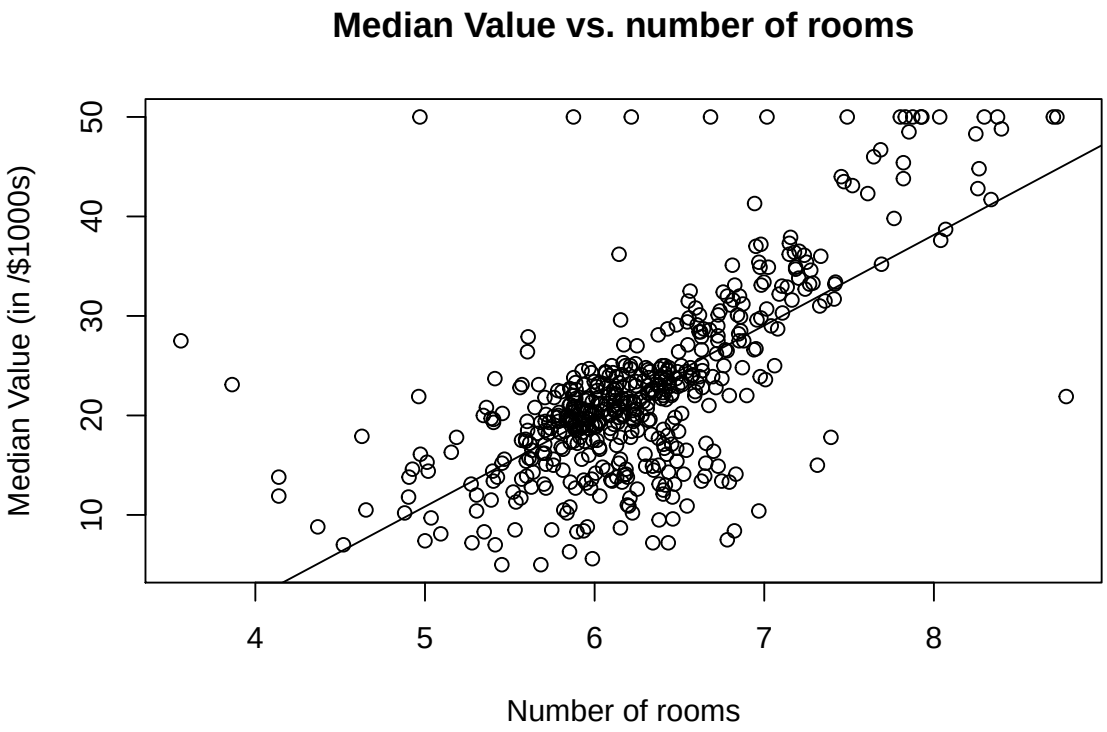
```
#Create the linear model for lmstat and medv variables
lmLstat <- lm(data = Boston, medv ~ lstat, na.action = na.exclude)

#Plot the graph for lmstat and medv
plot(Boston$medv ~ Boston$lstat
     , main = "Median Value vs. Lower Status Population"
     , ylab = "Median Value (in /$1000s)"
     , xlab = "Lower Status Population (per cent)")
#Plotting the regression line
abline(lmLstat)
```



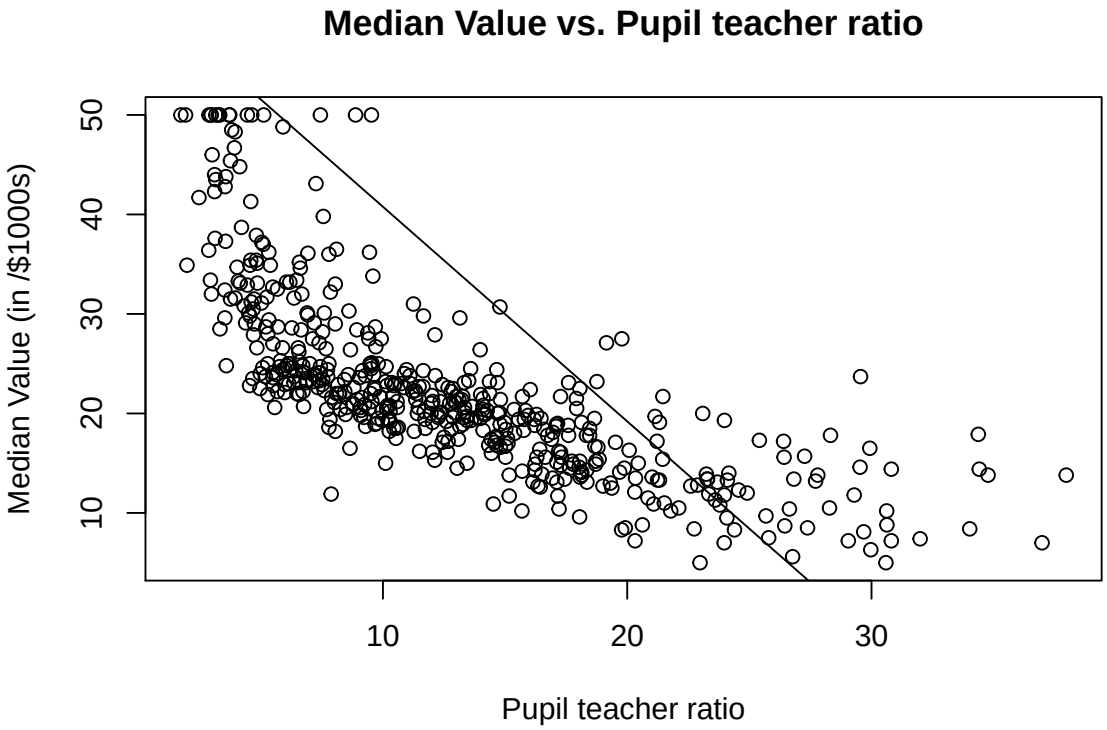

```
#Create the linear model for rm and medv variables
lmRm <- lm(data = Boston, medv ~ rm, na.action = na.exclude)

#Plot the graph for lmstat and medv
plot(Boston$medv ~ Boston$rm
     , main = "Median Value vs. number of rooms"
     , ylab = "Median Value (in /$1000s)"
     , xlab = "Number of rooms")
#Plotting the regression line
abline(lmRm)
```




```
#Create the linear model for lmstat and medv variables
lmPtratio <- lm(data = Boston, medv ~ ptratio, na.action = na.exclude)

#Plot the graph for lmstat and medv
plot(Boston$medv ~ Boston$lstat
     , main = "Median Value vs. Pupil teacher ratio"
     , ylab = "Median Value (in /$1000s)"
     , xlab = "Pupil teacher ratio")
#Plotting the regression line
abline(lmPtratio)
```



5. How do your results from (3) compare to your results from (4)? Create a plot displaying the univariate regression coefficients from (3) on the x-axis and the multiple regression coefficients from part (4) on the y-axis. Use this visualization to support your response.

   From the statistical analysis of (3) andb from (4) we have been able to establish the relationship between 3 significant variables. lstat, em and ptratio. In (3) these variables showed a higher correlation and the results in (4) supported those findings.

```
#Create list of coeffecients from (c)
univariateCoeff <- c(summary(lmLstat)$coefficient[2,1],
                         summary(lmRm)$coefficient[2,1],
                     summary(lmPtratio)$coefficient[2,1],
                     summary(lmCrim)$coefficient[2,1],
                     summary(lmZn)$coefficient[2,1],
                     summary(lmIndus)$coefficient[2,1],
                     summary(lmchas)$coefficient[2,1],
                     summary(lmNox)$coefficient[2,1],
                     summary(lmAge)$coefficient[2,1],
                     summary(lmDis)$coefficient[2,1],
                     summary(lmRad)$coefficient[2,1],
                     summary(lmTax)$coefficient[2,1],
                     summary(lmBlack)$coefficient[2,1])

multiVariateCoeff <- bostonRegModel$coefficients[2:14]

univariateCoeff
```

```
##  [1]  -0.95004935   9.10210898  -2.15717530  -0.41519028   0.14213999
```

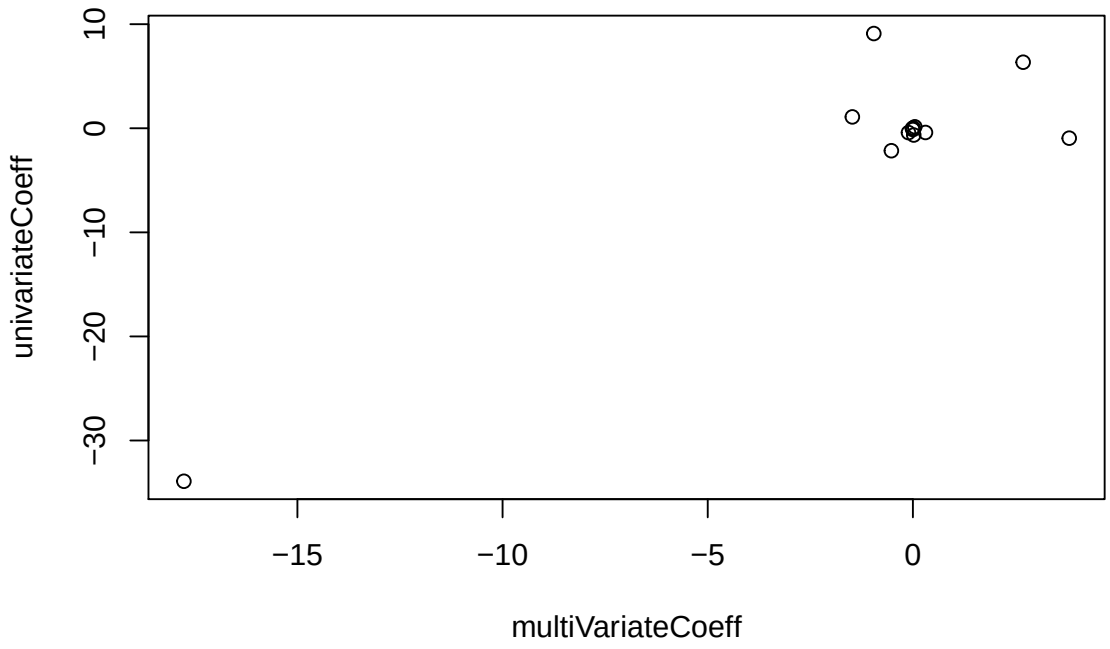
```
##  [6]  -0.64849005   6.34615711 -33.91605501  -0.12316272   1.09161302
## [11]  -0.40309540  -0.02556810   0.03359306
```

```
multiVariateCoeff
```

```
##           rm       ptratio         lstat          crim            zn
##  3.809865e+00 -9.527472e-01 -5.247584e-01 -1.080114e-01  4.642046e-02
##         indus          chas           nox           age           dis
##  2.055863e-02  2.686734e+00 -1.776661e+01  6.922246e-04 -1.475567e+00
##           rad           tax         black
##  3.060495e-01 -1.233459e-02  9.311683e-03
```

The coeffecients are different. we will now plot the two list of coefficients.

```
#Plot the results from univariate and multivariate analysis.
plot(univariateCoeff ~ multiVariateCoeff)
```



6. Is there evidence of a non-linear association between any of the predictors and the response? To answer this question, for each predictor X fit a model of the form:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$$

7. Consider performing a stepwise model selection procedure to determine the bets fit model. Discuss your results. How is this model different from the model in (4)?

```r
#Fit a model of form $$ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon $$ for lstat
model.lstat <- lm(medv ~ lstat + I(lstat^2) + I(lstat^3), data = Boston)

# Summarize model to check linearity
summary(model.lstat)
```

```
##
## Call:
## lm(formula = medv ~ lstat + I(lstat^2) + I(lstat^3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.5441  -3.7122  -0.5145   2.4846  26.4153
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 48.6496253  1.4347240  33.909  < 2e-16 ***
## lstat       -3.8655928  0.3287861 -11.757  < 2e-16 ***
## I(lstat^2)   0.1487385  0.0212987   6.983 9.18e-12 ***
## I(lstat^3)  -0.0020039  0.0003997  -5.013 7.43e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.396 on 502 degrees of freedom
## Multiple R-squared:  0.6578, Adjusted R-squared:  0.6558
## F-statistic: 321.7 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since the p - value is significantly less than 0.05, we can say that the association is non linear.

```r
#Fit a model of form $$ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon $$ for lstat
model.rm <- lm(medv ~ rm + I(rm^2) + I(rm^3), data = Boston)

# Summarize model to check linearity
summary(model.rm)
```

```
##
## Call:
## lm(formula = medv ~ rm + I(rm^2) + I(rm^3), data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.102  -2.674   0.569   3.011  35.911
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  241.3108    47.3275   5.099 4.85e-07 ***
## rm          -109.3906    22.9690  -4.763 2.51e-06 ***
## I(rm^2)       16.4910     3.6750   4.487 8.95e-06 ***
## I(rm^3)       -0.7404     0.1935  -3.827 0.000146 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.11 on 502 degrees of freedom
```

```
## Multiple R-squared:  0.5612, Adjusted R-squared:  0.5586
## F-statistic:   214 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since the p - value is significantly less than 0.05, we can say that the association is non linear.

```
#Fit a model of form $$ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon $$ for lstat
model.ptratio <- lm(medv ~ ptratio + I(ptratio^2) + I(ptratio^3), data = Boston)

# Summarize model to check linearity
summary(model.ptratio)
```

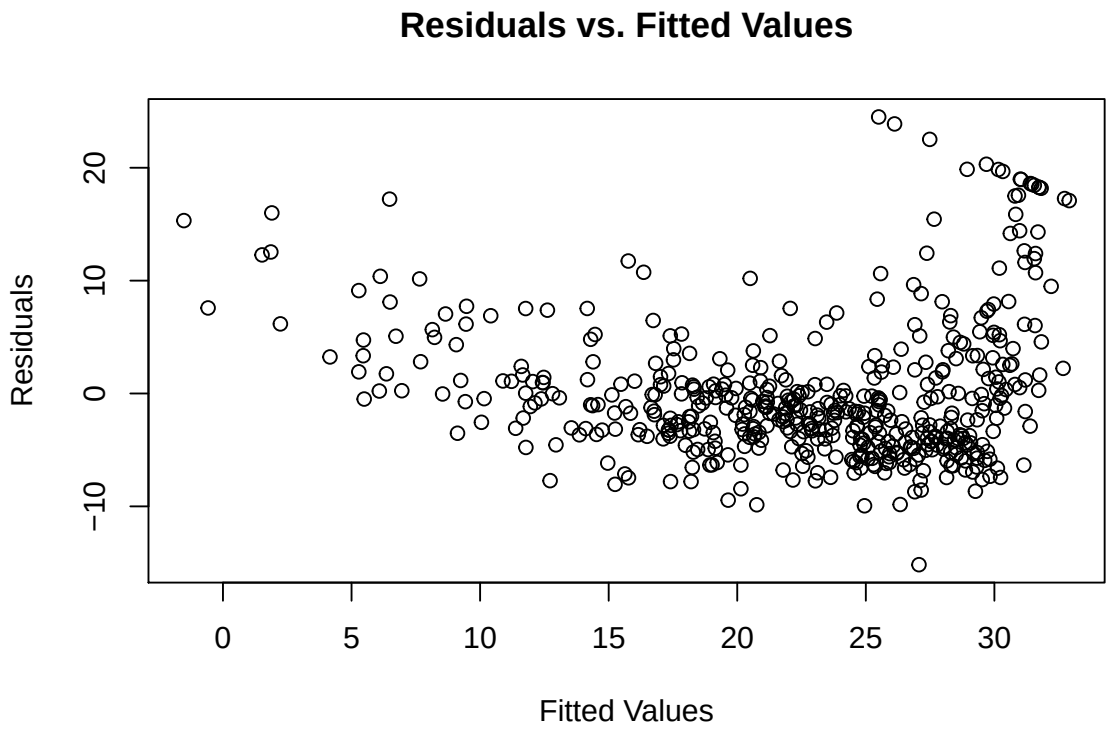
```
##
## Call:
## lm(formula = medv ~ ptratio + I(ptratio^2) + I(ptratio^3), data = Boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7795  -5.0364  -0.9778   3.4766  31.1636
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  312.28642  152.48693   2.048   0.0411 *
## ptratio      -48.69114   26.88441  -1.811   0.0707 .
## I(ptratio^2)   2.83995    1.56413   1.816   0.0700 .
## I(ptratio^3)  -0.05686    0.03005  -1.892   0.0590 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.898 on 502 degrees of freedom
## Multiple R-squared:  0.2669, Adjusted R-squared:  0.2625
## F-statistic: 60.91 on 3 and 502 DF,  p-value: < 2.2e-16
```

Since the p - value is significantly less than 0.05, we can say that the association is non linear.

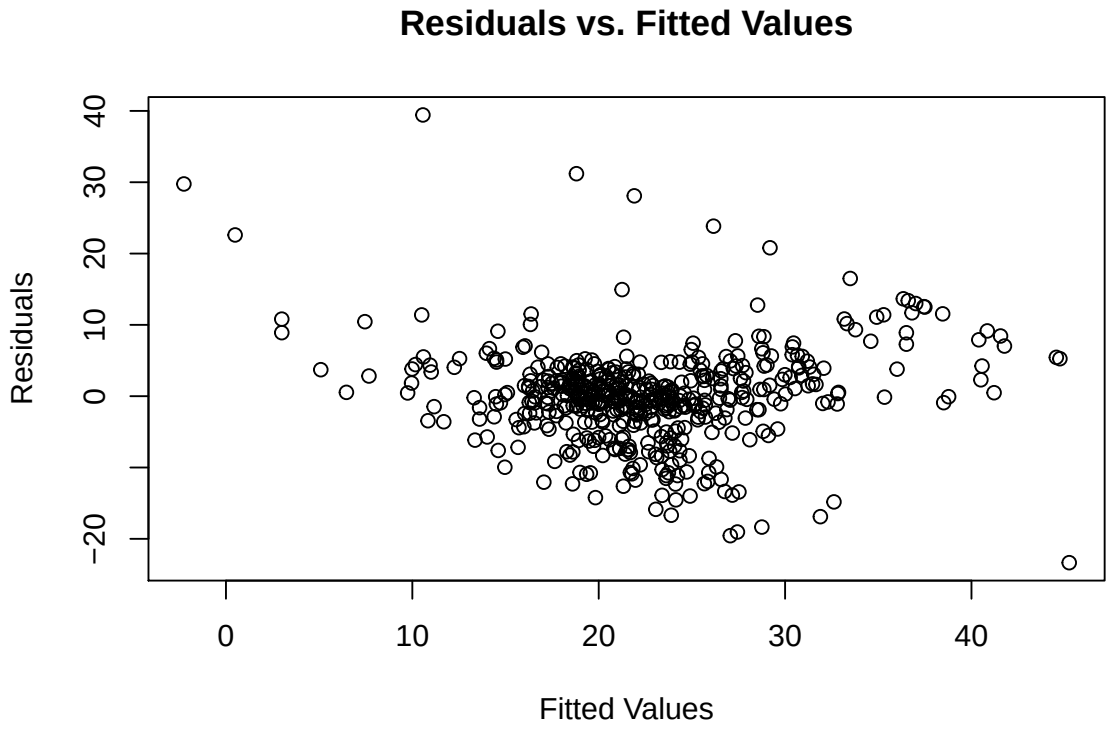
8. Evaluate the statistical assumptions in your regression analysis from (7) by performing a basic analysis of model residuals and any unusual observations. Discuss any concerns you have about your model.

   In addition to fitting the a model to check the linearity of variables we can check the residuals vs fitted values to better understand it

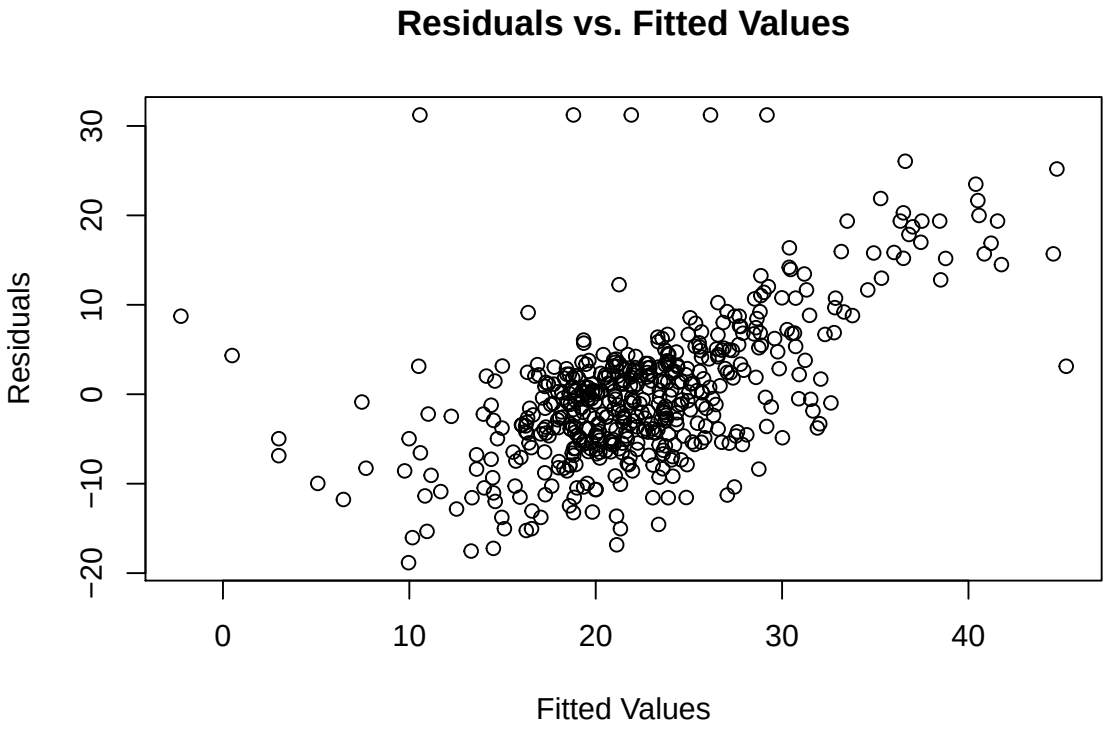
```
plot(lmLstat$residuals ~ lmLstat$fitted.values
     , main = "Residuals vs. Fitted Values"
     , ylab = "Residuals", xlab = "Fitted Values")
```




```
plot(lmRm$residuals ~ lmRm$fitted.values
     , main = "Residuals vs. Fitted Values"
     , ylab = "Residuals", xlab = "Fitted Values")
```




```
plot(lmPtratio$residuals ~ lmRm$fitted.values
     , main = "Residuals vs. Fitted Values"
     , ylab = "Residuals", xlab = "Fitted Values")
```



We cannot make any asumptions in our model. however looking at the plot for residuals and fitted values we see there is a trend in distribution much similar to the individual models we plotted earlier which is suggestive of a degree of significant statistical association. Since we were unable to prove any linearity in the predictor and response variables, the model is a difficult source of potential and accurate predictions.