Assignment 4

Group 14

2023-11-18

Name	Student ID
Pham, Quoc Huy	2299356
Hussain, Zakiuddin	2338350
Lee, Daeul Haven	2308018
Preetham	2288949
Jayanth	2288552
Srikavya	2311351


```
# load libraries ggplot2 and ggally
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggpubr)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(glue)
```

Q1 Load the data. Please download the Boston Dataset from canvas and read it in R

Use read.csv function to load the data, row.names = 1 means the first column will be used as row names and header = T means the first row will be used as variable names.

```
data <- read.csv('Boston.csv')#,row.names = 1,header = T)
head(data) #check the first few rows and columns of the dataset
```

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

```
 #check the number of columns(variables) in the dataset
print("Number of columns in the dataset")
```

```
## [1] "Number of columns in the dataset"
```

```
print(ncol(data))
```

```
## [1] 15
```

```
 #check the number of rows (samples) in the dataset
print("Number of rows in the dataset")
```

```
## [1] "Number of rows in the dataset"
```

```
print(nrow(data))
```

```
## [1] 506
```

How many variables in the dataset? What are they? Are they quantitative or qualitative variables?

```
names(data) #check the name of variables
```

```
##  [1] "X"       "crim"    "zn"      "indus"   "chas"    "nox"     "rm"
##  [8] "age"     "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"
## [15] "medv"
```

```
summary(data) #+statistics of each variable
```

```
##        X              crim                zn              indus
##  Min.   :  1.0   Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46
```

```
##  1st Qu.:127.2   1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19
##  Median :253.5   Median : 0.25651   Median :  0.00   Median : 9.69
##  Mean   :253.5   Mean   : 3.61352   Mean   : 11.36   Mean   :11.14
##  3rd Qu.:379.8   3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10
##  Max.   :506.0   Max.   :88.97620   Max.   :100.00   Max.   :27.74
##       chas              nox               rm             age
##  Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90
##  1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02
##  Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50
##  Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57
##  3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08
##  Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00
##       dis              rad              tax           ptratio
##  Min.   : 1.130   Min.   : 1.000   Min.   :187.0   Min.   :12.60
##  1st Qu.: 2.100   1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40
##  Median : 3.207   Median : 5.000   Median :330.0   Median :19.05
##  Mean   : 3.795   Mean   : 9.549   Mean   :408.2   Mean   :18.46
##  3rd Qu.: 5.188   3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20
##  Max.   :12.127   Max.   :24.000   Max.   :711.0   Max.   :22.00
##      black            lstat            medv
##  Min.   :  0.32   Min.   : 1.73   Min.   : 5.00
##  1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02
##  Median :391.44   Median :11.36   Median :21.20
##  Mean   :356.67   Mean   :12.65   Mean   :22.53
##  3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :396.90   Max.   :37.97   Max.   :50.00
```

There are 15 attributes in each case of the dataset, which comprises of

3 qualitative variables:

- X - index of sample
- chas - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- rad - index of accessibility to radial highways

12 quantitative variables:

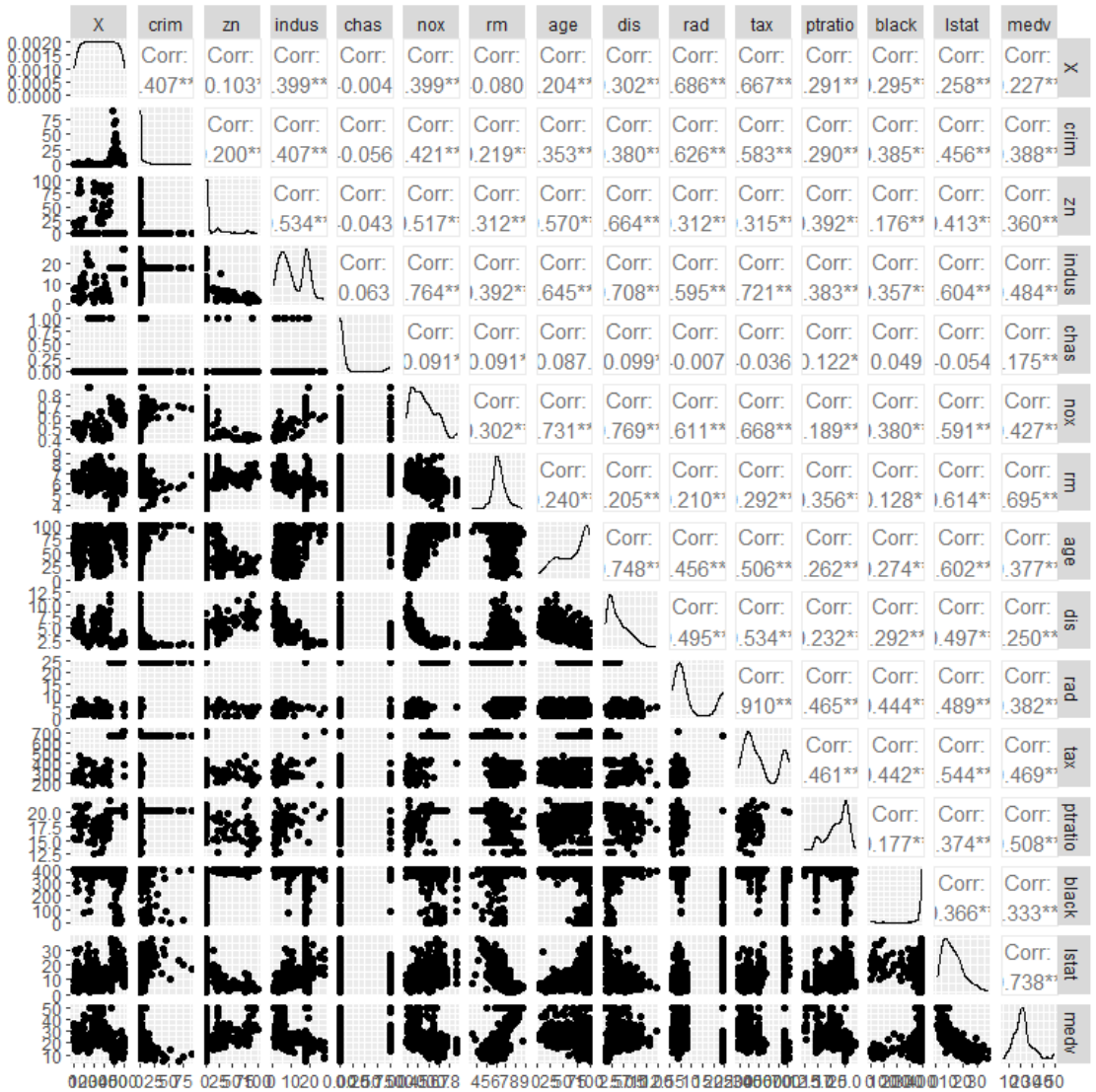
- crim - per capita crime rate by town
- zn - proportion of residential land zoned for lots over 25,000 sq.ft.
- indus - proportion of non-retail business acres per town.
- nox - nitric oxides concentration (parts per 10 million)
- rm - average number of rooms per dwelling
- age - proportion of owner-occupied units built prior to 1940
- dis - weighted distances to five Boston employment centers

- 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
- medv - Median value of owner-occupied homes in $1000's

Q2 Data visualization.

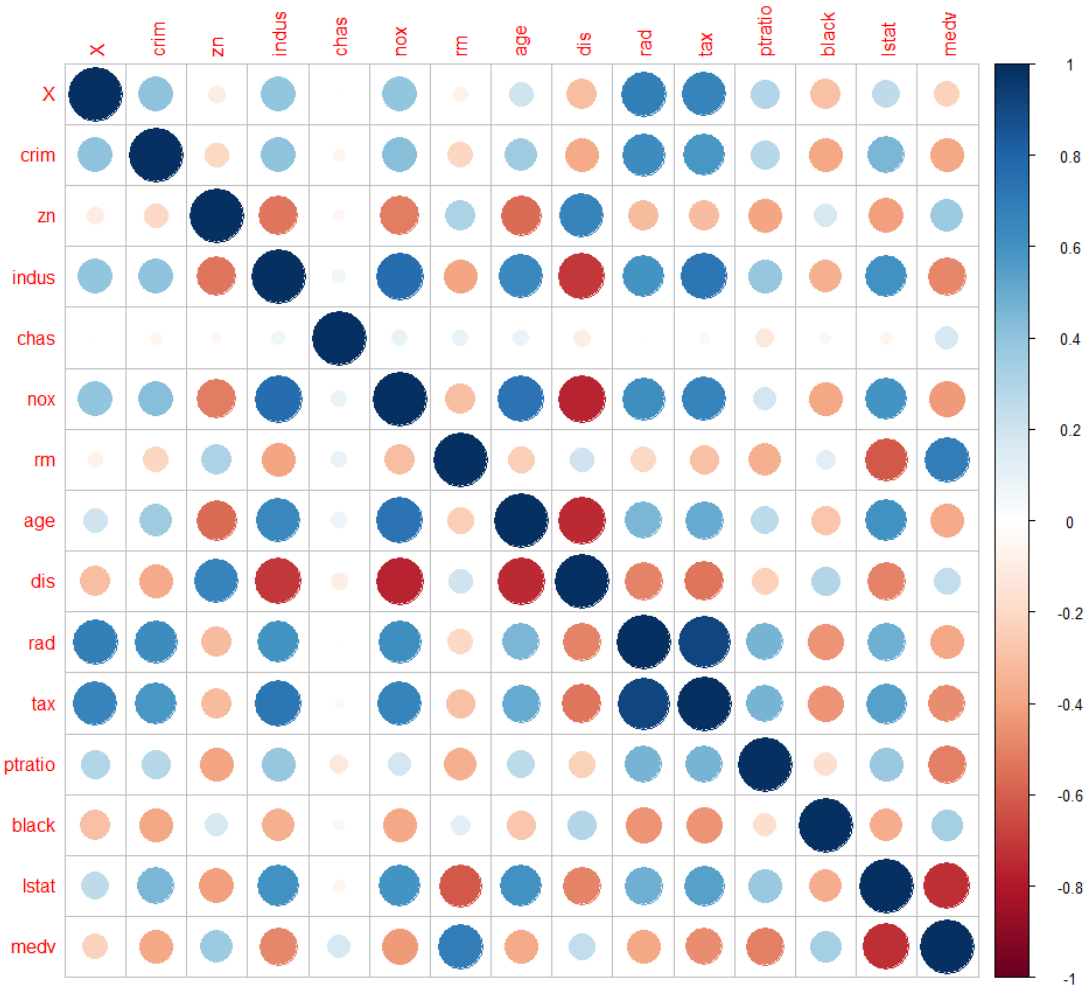
Please use the scatterplots to visualize this dataset.

```
#do a pair plot
ggpairs(data)
```

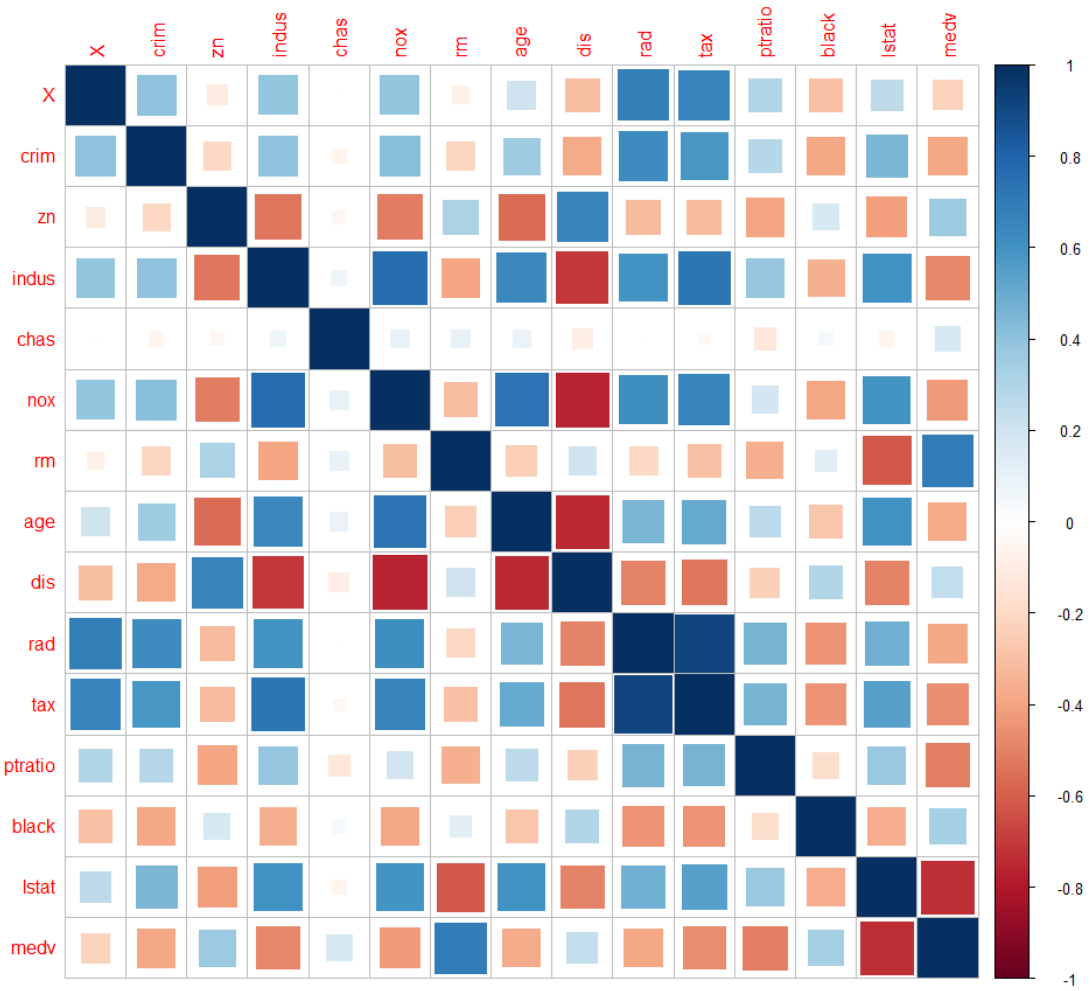


Correlation plot using package "corrplot"

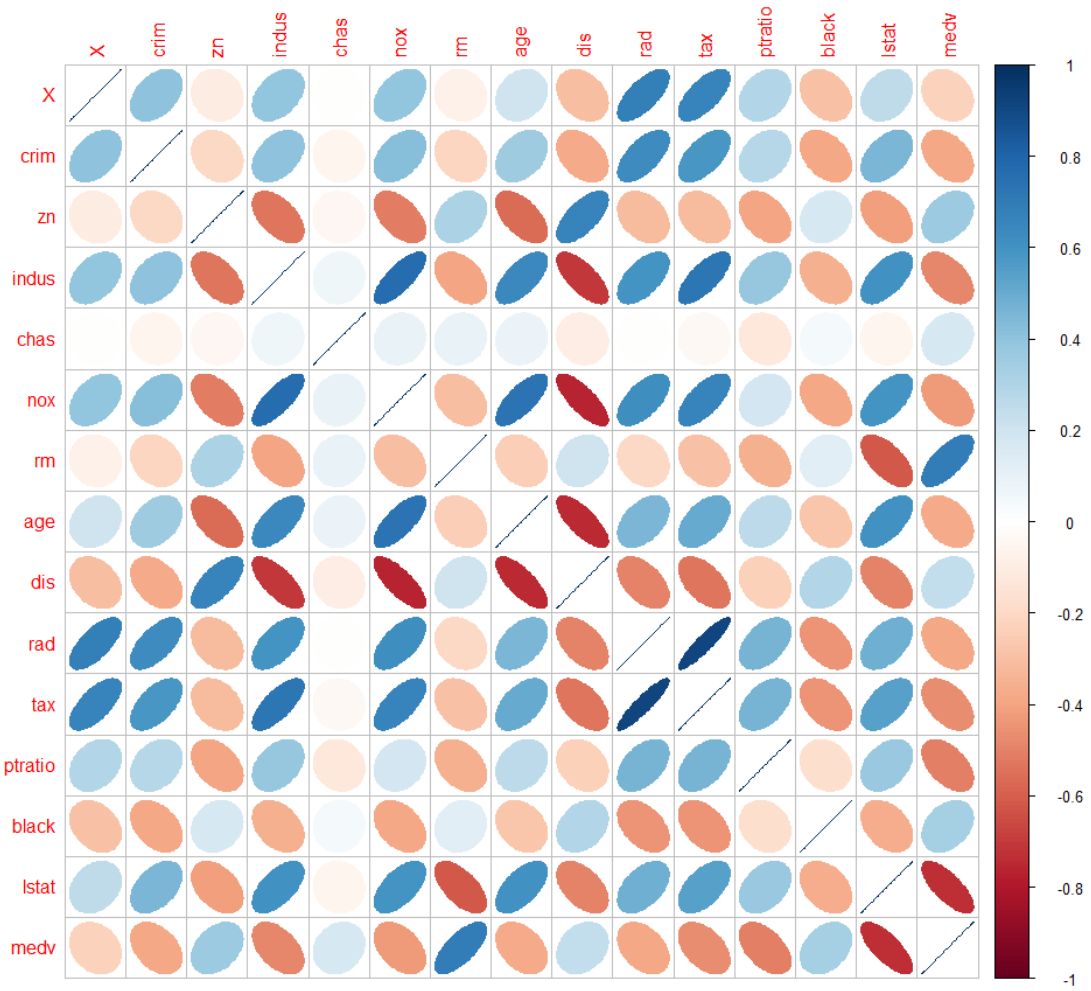
```
library(corrplot) #load the package to current environment

## corrplot 0.92 loaded

Corr <- cor(data) #calculate the correlation coefficient matrix of variables
# method=circle, square, ellipse, pie

corrplot(Corr, method = 'circle')
```



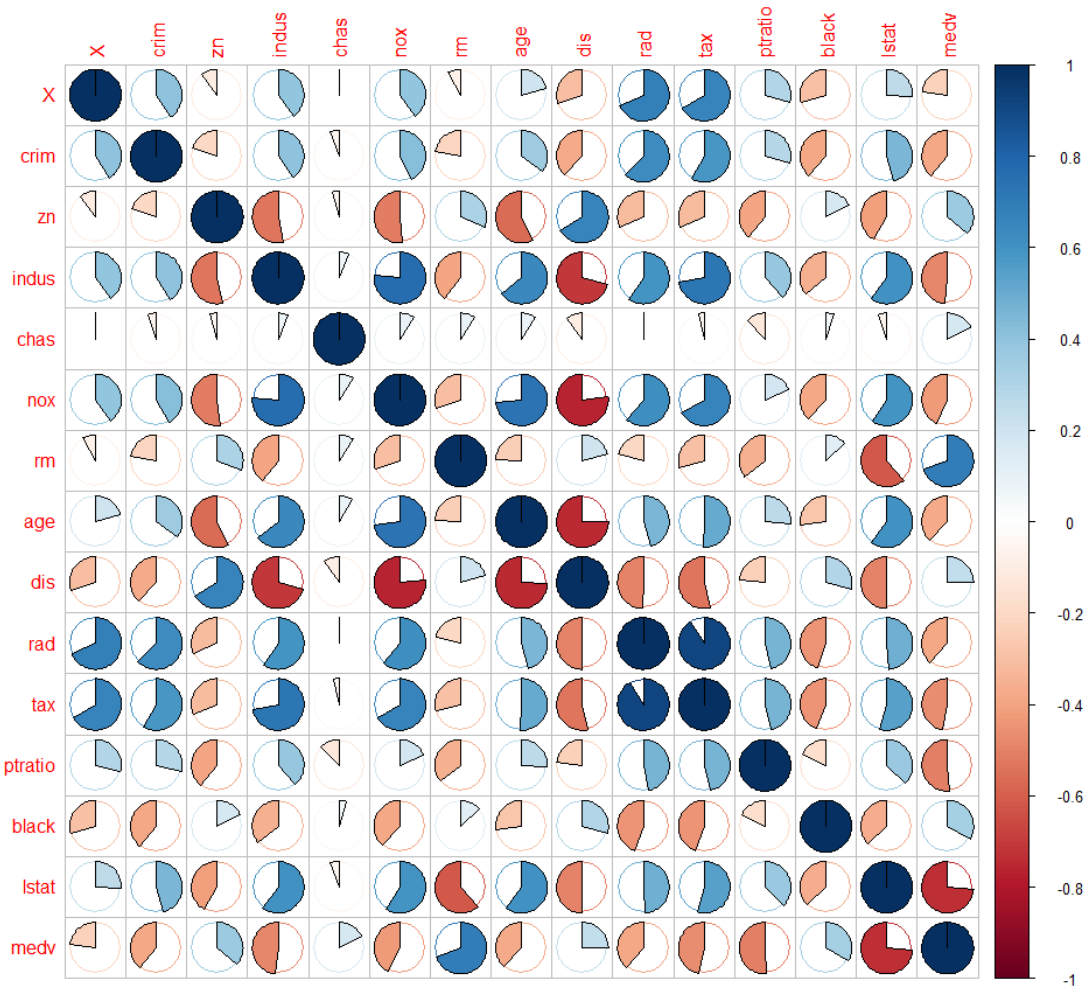

```
corrplot(Corr, method = 'square')
```




```
corrplot(Corr, method = 'ellipse')
```

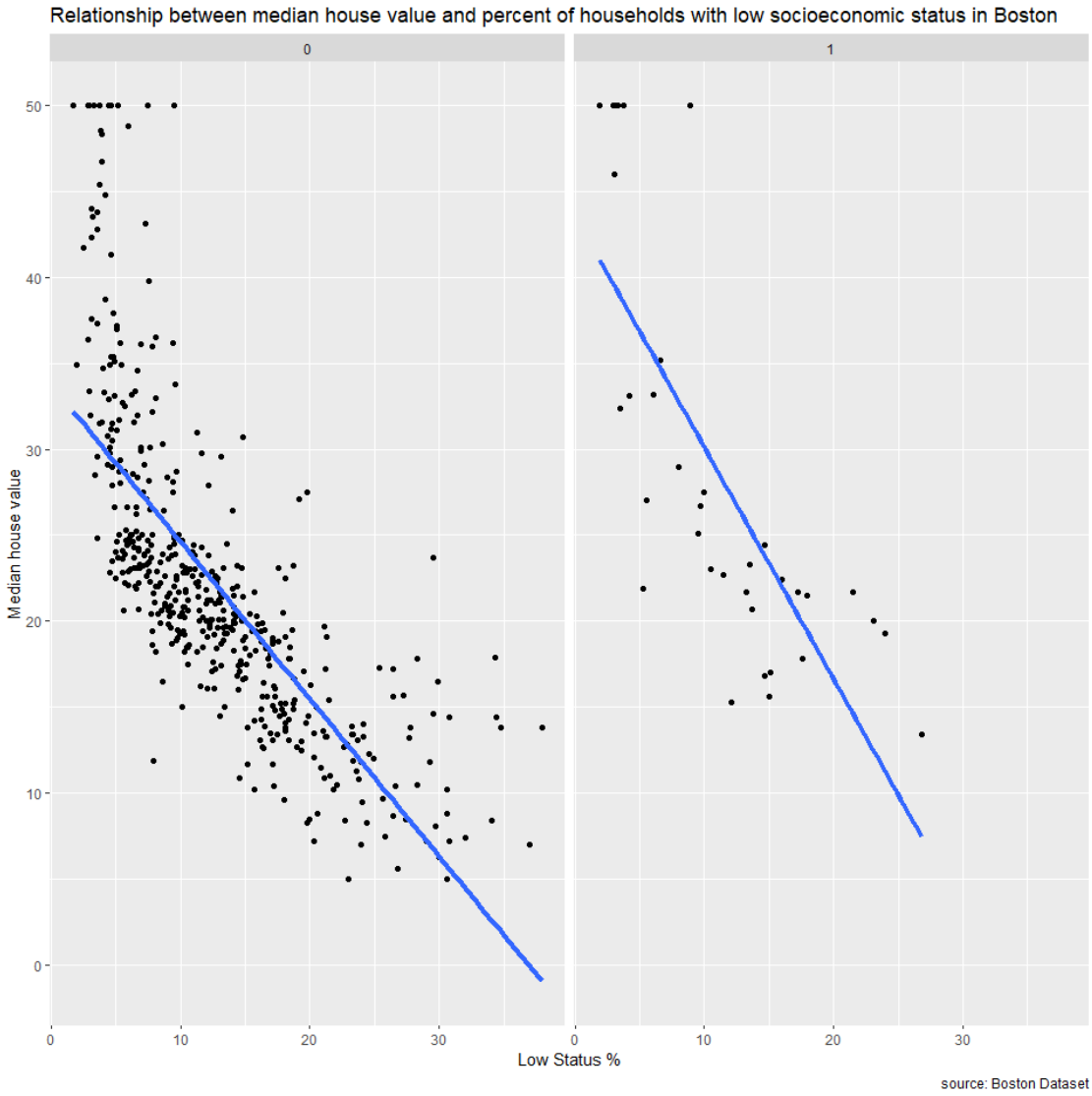



```
corrplot(Corr, method = 'pie')
```



add custom visualizations too

```
ggplot(data = data, mapping = aes(x = lstat, y = medv)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, linewidth = 1.5)+
  facet_wrap(~chas) +
  labs(title = "Relationship between median house value and percent of
households with low socioeconomic status in Boston",
       caption = "source: Boston Dataset",
       x = "Low Status %",
       y = "Median house value")

## `geom_smooth()` using formula = 'y ~ x'
```



Q3 Simple linear regression. Please fit a simple linear regression model between medv (median house value) and lstat (percent of households with low socioeconomic status).

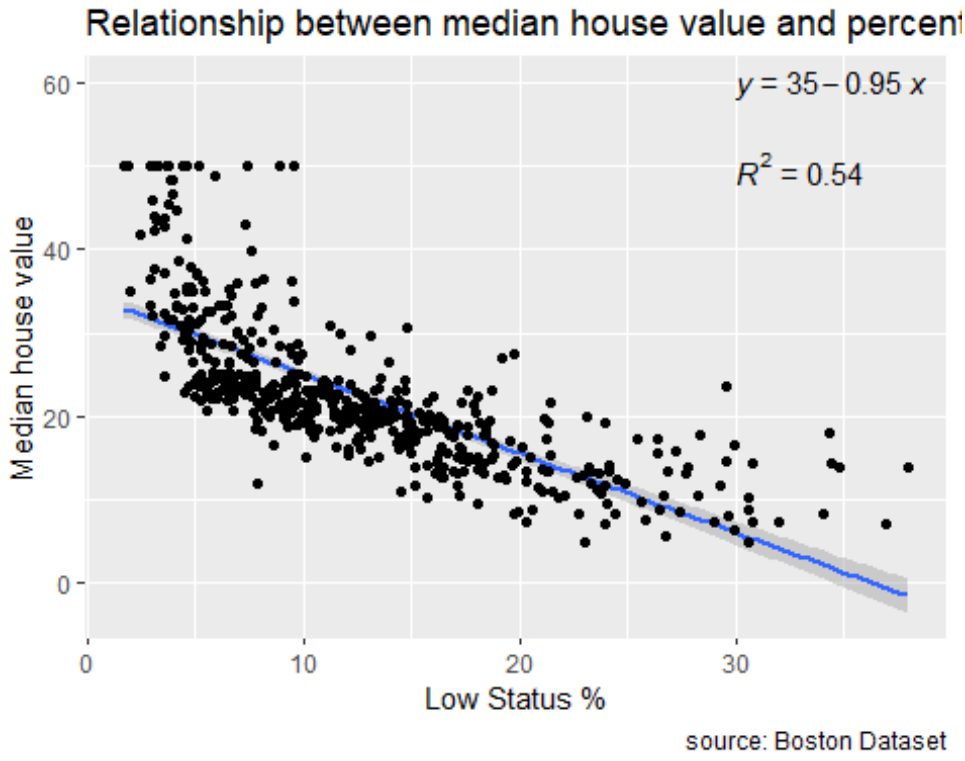
```
fit.simple <- lm(medv~lstat, data = data)
summary(fit.simple)
```

```
##
## Call:
## lm(formula = medv ~ lstat, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.168  -3.990  -1.318   2.034  24.500
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 34.55384    0.56263   61.41   <2e-16 ***
## lstat       -0.95005    0.03873  -24.53   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.216 on 504 degrees of freedom
## Multiple R-squared:  0.5441, Adjusted R-squared:  0.5432
## F-statistic: 601.6 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
ggplot(data = data, mapping = aes(x = lstat, y = medv)) +
        geom_smooth(method="lm") +
        geom_point() +
        stat_regline_equation(label.x=30, label.y=60) +
        stat_cor(aes(label=..rr.label..), label.x=30, label.y=50) +
        labs(title = "Relationship between median house value and percent of
households with low socioeconomic status in Boston",
       caption = "source: Boston Dataset",
       x = "Low Status %",
       y = "Median house value")
```

```
## Warning: The dot-dot notation (`..rr.label..`) was deprecated in ggplot2
3.4.0.
## ℹ Please use `after_stat(rr.label)` instead.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.

## `geom_smooth()` using formula = 'y ~ x'
```



a) Is there a relationship between median house value and percent of households with low socioeconomic status? Infer based on p-value of t-test on the coefficient.

Yes, there is a relationship between median house value and percent of households with low socioeconomic status. As the p-value of the t-test is smaller than 2.2e-16 , we can reject the null hypothesis with the significant value alpha of 0.001 and conclude that there is a statistically significant relationship between these two variables.

b) How large is the effect of percent of households with low socioeconomic status on median house value?

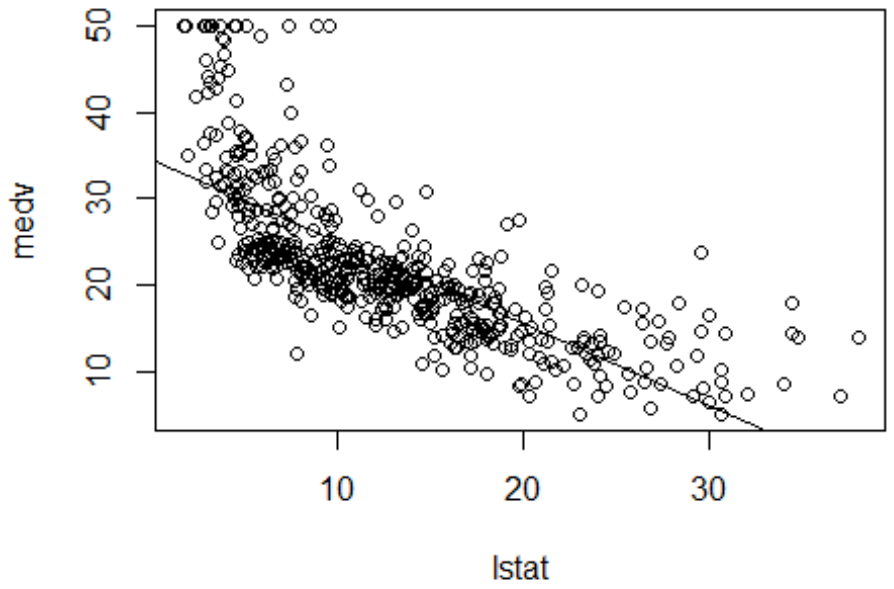
The coefficient of the percent of households with low socioeconomic status is -0.95 which means that for every one unit increase in the percent of households with low socioeconomic status, the median house value decreases by 0.95 units. This indicates a strong negative relationship between these two variables

c) How good this model fits the data? from the summary write the Residual standard error and variance what do they mean

From the regression report, Residual standard error: 6.216 on 504 degrees of freedom ,Multiple R-squared: 0.5441 and Adjusted R-squared: 0.5432 which indicates that the model explains about 54.41% of the variability in the median house value. The adjusted R-squared value suggests that approximately 54.32% of the variability is accounted for by the independent variable, low socioeconomic status. Overall, this indicates that the model fits the data moderately well, but there may be other factors influencing the median house value that are not captured by this model.

d) Visualize the fitted line.

```
plot(data$lstat, data$medv, xlab = 'lstat', ylab = 'medv')
abline(fit.simple)
```



e) If the percent of households with low socioeconomic status for three new neighborhoods are 5, 10 and 15, what will be the predictions of their median house value?

```
lstat.new <- data.frame( lstat = c(5,10,15))
medv.new <- predict(fit.simple, lstat.new)
medv.new

##        1        2        3
## 29.80359 25.05335 20.30310
```

f) What are the 95% confidence intervals of your predictions?

```
medv.new.conf <- predict(fit.simple, lstat.new, interval = "confidence")# use as a parameter interval = "confidence"
medv.new.conf

##        fit      lwr      upr
## 1 29.80359 29.00741 30.59978
## 2 25.05335 24.47413 25.63256
## 3 20.30310 19.73159 20.87461
```

g) If the true median house values for three new neighborhoods are 33, 20, 50 respectively, what are residuals, what are the prediction errors? Which prediction is more accurate?

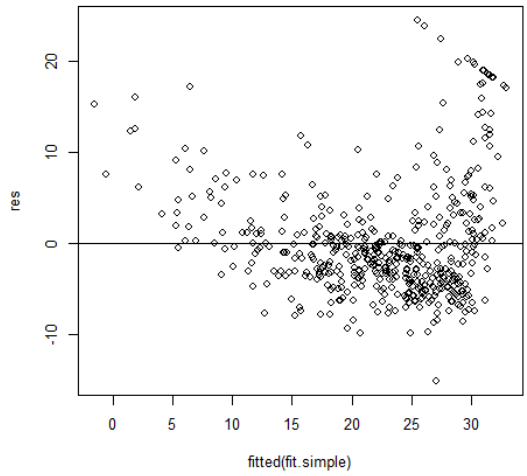
residual = observation - prediction

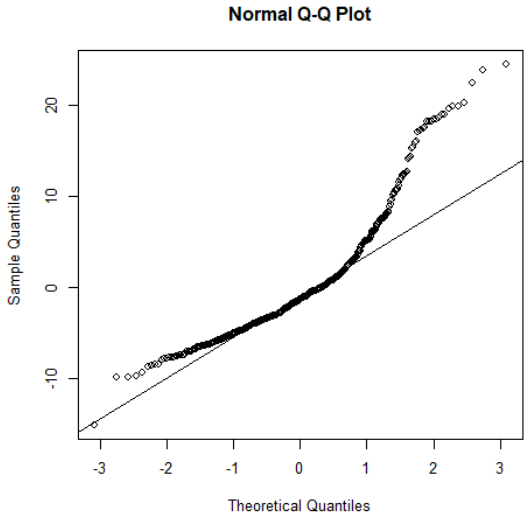
```r
#observations as a set - medv.new
medv.true = c(33,20,50)
medv.residual = medv.new- medv.true
medv.residual
```

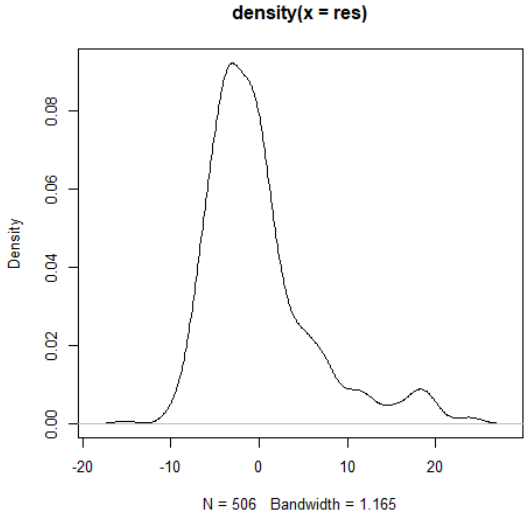
```
##          1          2          3
##  -3.196406   5.053347 -29.696899
```

Q4 Residual plot. Please plot the residual plots of simple linear regression model fitted in Problem 3 and answer the following question.

```r
par(mfrow =c(2,2))
res <- resid(fit.simple)
plot(fitted(fit.simple), res)
abline(0,0)

qqnorm(res)
qqline(res)

plot(density(res))
```







a) Is there a nonlinear relationship between medv and lstat?

In the residual plot, we observed that the residuals follow a parabolic curve , indicating a possible nonlinear relationship between medv and lstat. This suggests that the relationship between these two variables may not be adequately captured by a linear model. Further analysis using nonlinear regression techniques may be necessary to better understand this relationship.

b) Is there correlation between error terms?

```
cor.test(fitted(fit.simple), res, method="pearson")

##
##  Pearson's product-moment correlation
##
## data:  fitted(fit.simple) and res
## t = -1.0144e-15, df = 504, p-value = 1
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.08716868  0.08716868
## sample estimates:
##           cor
## -4.518703e-17
```

Using Pearson's correlation test we have p-value = 1, thus we do not have evidence to reject the null hypothesis of no correlation between the error terms. This suggests that the error terms are independent and there is no correlation present among them.

c) Is there heteroscedasticity between error terms?

```
bptest(fit.simple)

##
##  studentized Breusch-Pagan test
##
## data:  fit.simple
## BP = 15.497, df = 1, p-value = 8.262e-05
```

Using the Breusch-Pagan test, we obtain the p-value of 8.262e-5 which is less than 0.01. Thus, we reject Ho (The variance of the errors is constant -homoscedasticity) and conclude that there is heteroscedasticity between error terms.

d) Are there outliers?

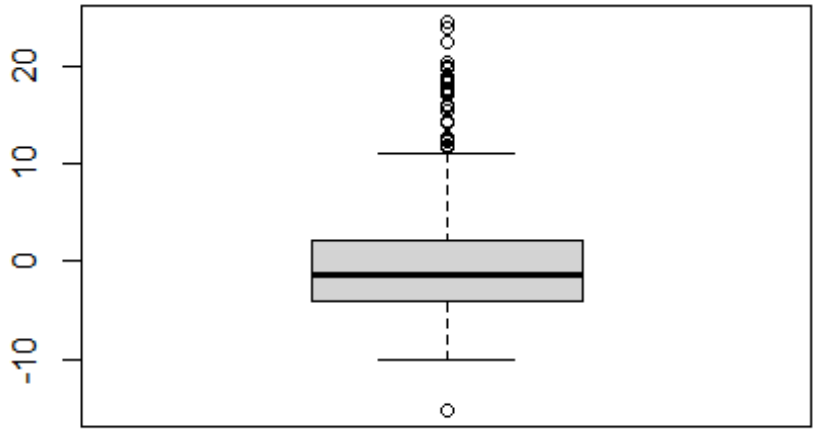
Cook's distance is a measure of the influence of each observation on the fitted values. Large values of Cook's distance indicate potential outliers.

```
cooksd <- cooks.distance(fit.simple)
potential.outliers <- which(cooksd > 4 / length(cooksd))

print(res[potential.outliers])

##          9         49         99        142        148        149
162
##  10.381136   9.117180  12.637835  12.537357   8.101117  10.151557
17.089745
##        163        164        167        187        196        203
204
##  17.270254  18.600323  18.961342  19.673879  18.267806  10.700813
17.565847
##        205        215        225        226        229        234
254
##  18.182301  17.220118  14.179363  19.844888  15.870353  17.498854
11.609334
##        257        258        262        263        268        269
281
##  12.400813  20.310412  15.443517  19.860951  22.514526  11.948315
14.418345
##        283        284        369        370        371        372
```

```
373
##  14.305808  18.448315  18.543320  18.989843  18.258305  24.500129
23.882597
##        374        375        413        415        439        506
##  12.279375  15.319533  15.999355   7.578984   6.166838 -15.167452
```

Using Cook's distance we found 41 outliers represent in the residual plot. We can confirm the finding using boxplot on the residuals

```
boxplot(res)
```



The boxplot show consistent result with the Cook's distance test. Thus, there are outliers.

Q5 Multiple linear regression. Please fit a multiple linear regression model between medv and the other variables.

```
fit.multiple <- lm(medv~., data = data)
summary(fit.multiple)

##
## Call:
## lm(formula = medv ~ ., data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.8948  -2.7585  -0.4663   1.7963  26.0911
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.461352   5.100994   7.148 3.21e-12 ***
## X            -0.002526   0.002080  -1.215 0.225046
## crim         -0.108762   0.032855  -3.310 0.001000 **
## zn            0.048031   0.013785   3.484 0.000538 ***
## indus         0.019932   0.061468   0.324 0.745871
## chas          2.705245   0.861298   3.141 0.001786 **
## nox         -17.541602   3.822390  -4.589 5.66e-06 ***
## rm            3.839225   0.418422   9.175  < 2e-16 ***
## age          -0.001938   0.013380  -0.145 0.884866
## dis          -1.493304   0.199892  -7.471 3.68e-13 ***
## rad           0.324925   0.068111   4.771 2.43e-06 ***
## tax          -0.011598   0.003807  -3.046 0.002443 **
## ptratio      -0.947985   0.130822  -7.246 1.67e-12 ***
## black         0.009357   0.002685   3.485 0.000536 ***
## lstat        -0.526184   0.050704 -10.377  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.743 on 491 degrees of freedom
## Multiple R-squared:  0.7414, Adjusted R-squared:  0.734
## F-statistic: 100.6 on 14 and 491 DF,  p-value: < 2.2e-16
```

a) Is there a relationship between median house value and the other variables?

Yes, there is a relationship between median house value and the other variables. As the p-value of the t-test is smaller than 2.2e-16 , we can reject the null hypothesis with the significant value alpha of 0.001 and conclude that there is a statistically significant relationship between median house value and the other variables.

b) Which variables are significant and how large are the effect?

Looking at the coefficients, we found that variables dis, rm, nox, chas has large value of coefficients, which indicates that these variables have a significant effect on the outcome. The larger the coefficient, the larger the impact of that variable on the outcome variable.

The value of coefficients of the aforementioned variables is as follow: * chas: 2.705245 * nox:-17.541602 * rm: 3.839225 * dis: -1.493304

This means a unit increase of chas will result in a 2.705245 unit increase in the outcome variable, while a unit increase of nox will lead to a decrease of 17.541602 units in the outcome variable. Similarly, a unit increase in rm will result in a 3.839225 unit increase in the outcome variable, while a unit increase in dis will lead to a decrease of 1.493304 units in the outcome variable. These coefficients provide valuable insights into the relationship between these variables and the outcome

c) How good this model fits the data?

From the regression report, Residual standard error: 4.743 on 491 degrees of freedom ,Multiple R-squared of 0.7414 and Adjusted R-squared of 0.734 which indicates that the

model explains about 74.14% of the variability in the median house value. The adjusted R-squared value suggests that approximately 73.4% of the variability is accounted for by the independent variables. Overall, this indicates that the model fits the data better than the fit.simple model with the Multiple R-squared of only 0.5441, but there may be other factors influencing the median house value that are not captured by this model.

d) Select the best subset of variables using forward selection, backward selection and mixed selection with AIC criteria. (write what they are and what is AIC criteria)

The Akaike Information Criterion (AIC) is a statistical measure used for model selection. It was developed by Hirotugu Akaike, a Japanese statistician, and is based on information theory. AIC provides a way to balance the goodness of fit of a model to the data and its complexity.

AIC takes into account two main factors: goodness of fit and model complexity penalty. Goodness of fit refers to how well a model explains the observed data. A model that fits the data well will have a lower AIC value. On the other hand, model complexity penalty refers to the idea that simpler models are preferred unless the added complexity significantly improves the fit to the data. AIC penalizes models for being too complex, with a greater penalty for models with more parameters.

The formula for AIC is given by The formula for AIC is given by **AIC=−2×log-likelihood+2×number of parameters**. The log-likelihood measures how well the model explains the observed data, while the penalty term discourages overly complex models.

In the context of model selection, the goal is to find the model that minimizes the AIC value. This involves a trade-off between goodness of fit and model simplicity. A lower AIC suggests a better balance between explaining the data and avoiding overfitting.

Forward selection

```
fit.null <- lm(medv~1,data = data)
select.forward <- step(fit.null, scope=list(lower=fit.null,
upper=fit.multiple), direction="forward")
```

```
## Start:  AIC=2246.51
## medv ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + lstat    1   23243.9 19472 1851.0
## + rm       1   20654.4 22062 1914.2
## + ptratio  1   11014.3 31702 2097.6
## + indus    1    9995.2 32721 2113.6
## + tax      1    9377.3 33339 2123.1
## + nox      1    7800.1 34916 2146.5
## + crim     1    6440.8 36276 2165.8
## + rad      1    6221.1 36495 2168.9
## + age      1    6069.8 36647 2171.0
## + zn       1    5549.7 37167 2178.1
## + black    1    4749.9 37966 2188.9
```

```
## + dis      1    2668.2 40048 2215.9
## + X        1    2193.4 40523 2221.8
## + chas     1    1312.1 41404 2232.7
## <none>                 42716 2246.5
##
## Step:  AIC=1851.01
## medv ~ lstat
##
##           Df Sum of Sq   RSS    AIC
## + rm       1    4033.1 15439 1735.6
## + ptratio  1    2670.1 16802 1778.4
## + chas     1     786.3 18686 1832.2
## + dis      1     772.4 18700 1832.5
## + age      1     304.3 19168 1845.0
## + tax      1     274.4 19198 1845.8
## + black    1     198.3 19274 1847.8
## + zn       1     160.3 19312 1848.8
## + crim     1     146.9 19325 1849.2
## + indus    1      98.7 19374 1850.4
## <none>                 19472 1851.0
## + X        1      59.1 19413 1851.5
## + rad      1      25.1 19447 1852.4
## + nox      1       4.8 19468 1852.9
##
## Step:  AIC=1735.58
## medv ~ lstat + rm
##
##           Df Sum of Sq   RSS    AIC
## + ptratio  1   1711.32 13728 1678.1
## + chas     1    548.53 14891 1719.3
## + black    1    512.31 14927 1720.5
## + tax      1    425.16 15014 1723.5
## + dis      1    351.15 15088 1725.9
## + crim     1    311.42 15128 1727.3
## + X        1    205.01 15234 1730.8
## + rad      1    180.45 15259 1731.6
## + indus    1     61.09 15378 1735.6
## <none>                 15439 1735.6
## + zn       1     56.56 15383 1735.7
## + age      1     20.18 15419 1736.9
## + nox      1     14.90 15424 1737.1
##
## Step:  AIC=1678.13
## medv ~ lstat + rm + ptratio
##
##         Df Sum of Sq   RSS    AIC
## + dis    1    499.08 13229 1661.4
## + black  1    389.68 13338 1665.6
## + chas   1    377.96 13350 1666.0
## + crim   1    122.52 13606 1675.6
```

```
## + age    1     66.24 13662 1677.7
## <none>               13728 1678.1
## + tax    1     44.36 13684 1678.5
## + nox    1     24.81 13703 1679.2
## + X      1     20.62 13707 1679.4
## + zn     1     14.96 13713 1679.6
## + rad    1      6.07 13722 1679.9
## + indus  1      0.83 13727 1680.1
##
## Step:  AIC=1661.39
## medv ~ lstat + rm + ptratio + dis
##
##         Df Sum of Sq   RSS    AIC
## + nox    1    759.56 12469 1633.5
## + black  1    502.64 12726 1643.8
## + chas   1    267.43 12962 1653.1
## + indus  1    242.65 12986 1654.0
## + tax    1    240.34 12989 1654.1
## + crim   1    233.54 12995 1654.4
## + zn     1    144.81 13084 1657.8
## + X      1     76.13 13153 1660.5
## + age    1     61.36 13168 1661.0
## <none>               13229 1661.4
## + rad    1     22.40 13206 1662.5
##
## Step:  AIC=1633.47
## medv ~ lstat + rm + ptratio + dis + nox
##
##         Df Sum of Sq   RSS    AIC
## + chas   1    328.27 12141 1622.0
## + black  1    311.83 12158 1622.7
## + zn     1    151.71 12318 1629.3
## + crim   1    141.43 12328 1629.7
## + rad    1     53.48 12416 1633.3
## <none>               12469 1633.5
## + indus  1     17.10 12452 1634.8
## + tax    1     10.50 12459 1635.0
## + X      1      1.09 12468 1635.4
## + age    1      0.25 12469 1635.5
##
## Step:  AIC=1621.97
## medv ~ lstat + rm + ptratio + dis + nox + chas
##
##         Df Sum of Sq   RSS    AIC
## + black  1   272.837 11868 1612.5
## + zn     1   164.406 11977 1617.1
## + crim   1   116.330 12025 1619.1
## + rad    1    58.556 12082 1621.5
## <none>               12141 1622.0
## + indus  1    26.274 12115 1622.9
```

```
## + tax    1     4.187 12137 1623.8
## + age    1     2.331 12139 1623.9
## + X      1     0.540 12140 1624.0
##
## Step:  AIC=1612.47
## medv ~ lstat + rm + ptratio + dis + nox + chas + black
##
##         Df Sum of Sq   RSS    AIC
## + zn     1   189.936 11678 1606.3
## + rad    1   144.320 11724 1608.3
## + crim   1    55.633 11813 1612.1
## <none>               11868 1612.5
## + indus  1    15.584 11853 1613.8
## + age    1     9.446 11859 1614.1
## + tax    1     2.703 11866 1614.4
## + X      1     2.601 11866 1614.4
##
## Step:  AIC=1606.31
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn
##
##         Df Sum of Sq   RSS    AIC
## + crim   1    94.712 11584 1604.2
## + rad    1    93.614 11585 1604.2
## <none>               11678 1606.3
## + indus  1    16.048 11662 1607.6
## + tax    1     3.952 11674 1608.1
## + X      1     2.199 11676 1608.2
## + age    1     1.491 11677 1608.2
##
## Step:  AIC=1604.19
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim
##
##         Df Sum of Sq   RSS    AIC
## + rad    1   228.604 11355 1596.1
## <none>               11584 1604.2
## + indus  1    15.773 11568 1605.5
## + age    1     2.470 11581 1606.1
## + tax    1     1.305 11582 1606.1
## + X      1     0.317 11583 1606.2
##
## Step:  AIC=1596.1
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim + rad
##
##         Df Sum of Sq   RSS    AIC
## + tax    1   273.619 11081 1585.8
## + X      1    70.508 11284 1595.0
## <none>               11355 1596.1
## + indus  1    33.894 11321 1596.6
```

```
## + age    1     0.096 11355 1598.1
##
## Step:  AIC=1585.76
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim + rad + tax
##
##         Df Sum of Sq   RSS    AIC
## <none>               11081 1585.8
## + X      1    32.937 11048 1586.2
## + indus  1     2.518 11079 1587.7
## + age    1     0.063 11081 1587.8
```

```
# summary of forward
summary(select.forward)
```

```
##
## Call:
## lm(formula = medv ~ lstat + rm + ptratio + dis + nox + chas +
##     black + zn + crim + rad + tax, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## chas          2.718716   0.854240   3.183 0.001551 **
## black         0.009291   0.002674   3.475 0.000557 ***
## zn            0.045845   0.013523   3.390 0.000754 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## rad           0.299608   0.063402   4.726 3.00e-06 ***
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Using forward selection results in the following variables: lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad and tax #### Backward selection

```
select.backward <- step(fit.multiple, scope=list(lower=fit.null,
upper=fit.multiple), direction="backward")
```

```
## Start:  AIC=1590.12
## medv ~ X + crim + zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat
##
##           Df Sum of Sq   RSS    AIC
## - age      1      0.47 11046 1588.2
## - indus    1      2.37 11048 1588.2
## - X        1     33.20 11079 1589.6
## <none>                 11046 1590.1
## - tax      1    208.73 11254 1597.6
## - chas     1    221.93 11268 1598.2
## - crim     1    246.53 11292 1599.3
## - zn       1    273.12 11319 1600.5
## - black    1    273.20 11319 1600.5
## - nox      1    473.78 11519 1609.4
## - rad      1    511.97 11558 1611.0
## - ptratio  1   1181.26 12227 1639.5
## - dis      1   1255.48 12301 1642.6
## - rm       1   1893.94 12940 1668.2
## - lstat    1   2422.66 13468 1688.5
##
## Step:  AIC=1588.15
## medv ~ X + crim + zn + indus + chas + nox + rm + dis + rad +
##     tax + ptratio + black + lstat
##
##           Df Sum of Sq   RSS    AIC
## - indus    1      2.37 11048 1586.2
## - X        1     32.79 11079 1587.7
## <none>                 11046 1588.2
## - tax      1    210.46 11256 1595.7
## - chas     1    221.47 11268 1596.2
## - crim     1    246.53 11293 1597.3
## - black    1    272.88 11319 1598.5
## - zn       1    278.41 11324 1598.7
## - rad      1    513.75 11560 1609.2
## - nox      1    519.51 11566 1609.4
## - ptratio  1   1193.16 12239 1638.0
## - dis      1   1362.40 12408 1645.0
## - rm       1   1970.67 13017 1669.2
## - lstat    1   2749.73 13796 1698.6
##
## Step:  AIC=1586.25
## medv ~ X + crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     black + lstat
##
##           Df Sum of Sq   RSS    AIC
## - X        1     32.94 11081 1585.8
## <none>                 11048 1586.2
## - chas     1    228.66 11277 1594.6
## - tax      1    236.05 11284 1595.0
```

```
## - crim     1    248.68 11297 1595.5
## - black    1    271.50 11320 1596.5
## - zn       1    276.10 11324 1596.7
## - rad      1    533.86 11582 1608.1
## - nox      1    541.12 11590 1608.5
## - ptratio  1   1199.77 12248 1636.4
## - dis      1   1458.98 12507 1647.0
## - rm       1   1975.19 13024 1667.5
## - lstat    1   2754.60 13803 1696.9
##
## Step:  AIC=1585.76
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     black + lstat
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## - chas     1    227.21 11309 1594.0
## - crim     1    245.37 11327 1594.8
## - zn       1    257.82 11339 1595.4
## - black    1    270.82 11352 1596.0
## - tax      1    273.62 11355 1596.1
## - rad      1    500.92 11582 1606.1
## - nox      1    541.91 11623 1607.9
## - ptratio  1   1206.45 12288 1636.0
## - dis      1   1448.94 12530 1645.9
## - rm       1   1963.66 13045 1666.3
## - lstat    1   2723.48 13805 1695.0
```

```
#summary of backward
summary(select.backward)
```

```
##
## Call:
## lm(formula = medv ~ crim + zn + chas + nox + rm + dis + rad +
##     tax + ptratio + black + lstat, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## zn            0.045845   0.013523   3.390 0.000754 ***
## chas          2.718716   0.854240   3.183 0.001551 **
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## rad           0.299608   0.063402   4.726 3.00e-06 ***
```

```
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## black         0.009291   0.002674   3.475 0.000557 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Using forward selection results in the following variables: lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad and tax

Mixed selection

```
select.mixed <- step(fit.null, scope=list(lower=fit.null,
upper=fit.multiple), direction="both")
```

```
## Start:  AIC=2246.51
## medv ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + lstat    1   23243.9 19472 1851.0
## + rm       1   20654.4 22062 1914.2
## + ptratio  1   11014.3 31702 2097.6
## + indus    1    9995.2 32721 2113.6
## + tax      1    9377.3 33339 2123.1
## + nox      1    7800.1 34916 2146.5
## + crim     1    6440.8 36276 2165.8
## + rad      1    6221.1 36495 2168.9
## + age      1    6069.8 36647 2171.0
## + zn       1    5549.7 37167 2178.1
## + black    1    4749.9 37966 2188.9
## + dis      1    2668.2 40048 2215.9
## + X        1    2193.4 40523 2221.8
## + chas     1    1312.1 41404 2232.7
## <none>                 42716 2246.5
##
## Step:  AIC=1851.01
## medv ~ lstat
##
##           Df Sum of Sq   RSS    AIC
## + rm       1    4033.1 15439 1735.6
## + ptratio  1    2670.1 16802 1778.4
## + chas     1     786.3 18686 1832.2
## + dis      1     772.4 18700 1832.5
## + age      1     304.3 19168 1845.0
## + tax      1     274.4 19198 1845.8
## + black    1     198.3 19274 1847.8
## + zn       1     160.3 19312 1848.8
```

```
## + crim     1     146.9 19325 1849.2
## + indus    1      98.7 19374 1850.4
## <none>                 19472 1851.0
## + X        1      59.1 19413 1851.5
## + rad      1      25.1 19447 1852.4
## + nox      1       4.8 19468 1852.9
## - lstat    1   23243.9 42716 2246.5
##
## Step:  AIC=1735.58
## medv ~ lstat + rm
##
##           Df Sum of Sq   RSS    AIC
## + ptratio  1    1711.3 13728 1678.1
## + chas     1     548.5 14891 1719.3
## + black    1     512.3 14927 1720.5
## + tax      1     425.2 15014 1723.5
## + dis      1     351.2 15088 1725.9
## + crim     1     311.4 15128 1727.3
## + X        1     205.0 15234 1730.8
## + rad      1     180.5 15259 1731.6
## + indus    1      61.1 15378 1735.6
## <none>                 15439 1735.6
## + zn       1      56.6 15383 1735.7
## + age      1      20.2 15419 1736.9
## + nox      1      14.9 15424 1737.1
## - rm       1    4033.1 19472 1851.0
## - lstat    1    6622.6 22062 1914.2
##
## Step:  AIC=1678.13
## medv ~ lstat + rm + ptratio
##
##           Df Sum of Sq   RSS    AIC
## + dis      1     499.1 13229 1661.4
## + black    1     389.7 13338 1665.6
## + chas     1     378.0 13350 1666.0
## + crim     1     122.5 13606 1675.6
## + age      1      66.2 13662 1677.7
## <none>                 13728 1678.1
## + tax      1      44.4 13684 1678.5
## + nox      1      24.8 13703 1679.2
## + X        1      20.6 13707 1679.4
## + zn       1      15.0 13713 1679.6
## + rad      1       6.1 13722 1679.9
## + indus    1       0.8 13727 1680.1
## - ptratio  1    1711.3 15439 1735.6
## - rm       1    3074.3 16802 1778.4
## - lstat    1    5013.6 18742 1833.7
##
## Step:  AIC=1661.39
## medv ~ lstat + rm + ptratio + dis
```

```
##
##           Df Sum of Sq   RSS    AIC
## + nox      1     759.6 12469 1633.5
## + black    1     502.6 12726 1643.8
## + chas     1     267.4 12962 1653.1
## + indus    1     242.6 12986 1654.0
## + tax      1     240.3 12989 1654.1
## + crim     1     233.5 12995 1654.4
## + zn       1     144.8 13084 1657.8
## + X        1      76.1 13153 1660.5
## + age      1      61.4 13168 1661.0
## <none>                 13229 1661.4
## + rad      1      22.4 13206 1662.5
## - dis      1     499.1 13728 1678.1
## - ptratio  1    1859.3 15088 1725.9
## - rm       1    2622.6 15852 1750.9
## - lstat    1    5349.2 18578 1831.2
##
## Step:  AIC=1633.47
## medv ~ lstat + rm + ptratio + dis + nox
##
##           Df Sum of Sq   RSS    AIC
## + chas     1     328.3 12141 1622.0
## + black    1     311.8 12158 1622.7
## + zn       1     151.7 12318 1629.3
## + crim     1     141.4 12328 1629.7
## + rad      1      53.5 12416 1633.3
## <none>                 12469 1633.5
## + indus    1      17.1 12452 1634.8
## + tax      1      10.5 12459 1635.0
## + X        1       1.1 12468 1635.4
## + age      1       0.2 12469 1635.5
## - nox      1     759.6 13229 1661.4
## - dis      1    1233.8 13703 1679.2
## - ptratio  1    2116.5 14586 1710.8
## - rm       1    2546.2 15016 1725.5
## - lstat    1    3664.3 16134 1761.8
##
## Step:  AIC=1621.97
## medv ~ lstat + rm + ptratio + dis + nox + chas
##
##           Df Sum of Sq   RSS    AIC
## + black    1     272.8 11868 1612.5
## + zn       1     164.4 11977 1617.1
## + crim     1     116.3 12025 1619.1
## + rad      1      58.6 12082 1621.5
## <none>                 12141 1622.0
## + indus    1      26.3 12115 1622.9
## + tax      1       4.2 12137 1623.8
## + age      1       2.3 12139 1623.9
```

```
## + X        1       0.5 12140 1624.0
## - chas     1     328.3 12469 1633.5
## - nox      1     820.4 12962 1653.1
## - dis      1    1146.8 13288 1665.6
## - ptratio  1    1924.9 14066 1694.4
## - rm       1    2480.7 14622 1714.0
## - lstat    1    3509.3 15650 1748.5
##
## Step:  AIC=1612.47
## medv ~ lstat + rm + ptratio + dis + nox + chas + black
##
##           Df Sum of Sq   RSS    AIC
## + zn       1    189.94 11678 1606.3
## + rad      1    144.32 11724 1608.3
## + crim     1     55.63 11813 1612.1
## <none>                 11868 1612.5
## + indus    1     15.58 11853 1613.8
## + age      1      9.45 11859 1614.1
## + tax      1      2.70 11866 1614.4
## + X        1      2.60 11866 1614.4
## - black    1    272.84 12141 1622.0
## - chas     1    289.27 12158 1622.7
## - nox      1    626.85 12495 1636.5
## - dis      1   1103.33 12972 1655.5
## - ptratio  1   1804.30 13672 1682.1
## - rm       1   2658.21 14526 1712.7
## - lstat    1   2991.55 14860 1724.2
##
## Step:  AIC=1606.31
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn
##
##           Df Sum of Sq   RSS    AIC
## + crim     1     94.71 11584 1604.2
## + rad      1     93.61 11585 1604.2
## <none>                 11678 1606.3
## + indus    1     16.05 11662 1607.6
## + tax      1      3.95 11674 1608.1
## + X        1      2.20 11676 1608.2
## + age      1      1.49 11677 1608.2
## - zn       1    189.94 11868 1612.5
## - black    1    298.37 11977 1617.1
## - chas     1    300.42 11979 1617.2
## - nox      1    627.62 12306 1630.8
## - dis      1   1276.45 12955 1656.8
## - ptratio  1   1364.63 13043 1660.2
## - rm       1   2384.55 14063 1698.3
## - lstat    1   3052.50 14731 1721.8
##
## Step:  AIC=1604.19
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
```

```
##     crim
##
##           Df Sum of Sq   RSS    AIC
## + rad      1    228.60 11355 1596.1
## <none>                 11584 1604.2
## + indus    1     15.77 11568 1605.5
## + age      1      2.47 11581 1606.1
## + tax      1      1.31 11582 1606.1
## + X        1      0.32 11583 1606.2
## - crim     1     94.71 11678 1606.3
## - black    1    222.18 11806 1611.8
## - zn       1    229.02 11813 1612.1
## - chas     1    284.34 11868 1614.5
## - nox      1    578.44 12162 1626.8
## - ptratio  1   1192.90 12776 1651.8
## - dis      1   1345.70 12929 1657.8
## - rm       1   2419.57 14003 1698.2
## - lstat    1   2753.42 14337 1710.1
##
## Step:  AIC=1596.1
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim + rad
##
##           Df Sum of Sq   RSS    AIC
## + tax      1    273.62 11081 1585.8
## + X        1     70.51 11284 1595.0
## <none>                 11355 1596.1
## + indus    1     33.89 11321 1596.6
## + age      1      0.10 11355 1598.1
## - zn       1    171.14 11526 1601.7
## - rad      1    228.60 11584 1604.2
## - crim     1    229.70 11585 1604.2
## - chas     1    272.67 11628 1606.1
## - black    1    295.78 11651 1607.1
## - nox      1    785.16 12140 1627.9
## - dis      1   1341.37 12696 1650.6
## - ptratio  1   1419.77 12775 1653.7
## - rm       1   2182.57 13538 1683.1
## - lstat    1   2785.28 14140 1705.1
##
## Step:  AIC=1585.76
## medv ~ lstat + rm + ptratio + dis + nox + chas + black + zn +
##     crim + rad + tax
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## + X        1     32.94 11048 1586.2
## + indus    1      2.52 11079 1587.7
## + age      1      0.06 11081 1587.8
## - chas     1    227.21 11309 1594.0
```

```
## - crim     1    245.37 11327 1594.8
## - zn       1    257.82 11339 1595.4
## - black    1    270.82 11352 1596.0
## - tax      1    273.62 11355 1596.1
## - rad      1    500.92 11582 1606.1
## - nox      1    541.91 11623 1607.9
## - ptratio  1   1206.45 12288 1636.0
## - dis      1   1448.94 12530 1645.9
## - rm       1   1963.66 13045 1666.3
## - lstat    1   2723.48 13805 1695.0

#summary of mixed
summary(select.mixed)

##
## Call:
## lm(formula = medv ~ lstat + rm + ptratio + dis + nox + chas +
##     black + zn + crim + rad + tax, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## lstat        -0.522553   0.047424 -11.019  < 2e-16 ***
## rm            3.801579   0.406316   9.356  < 2e-16 ***
## ptratio      -0.946525   0.129066  -7.334 9.24e-13 ***
## dis          -1.492711   0.185731  -8.037 6.84e-15 ***
## nox         -17.376023   3.535243  -4.915 1.21e-06 ***
## chas          2.718716   0.854240   3.183 0.001551 **
## black         0.009291   0.002674   3.475 0.000557 ***
## zn            0.045845   0.013523   3.390 0.000754 ***
## crim         -0.108413   0.032779  -3.307 0.001010 **
## rad           0.299608   0.063402   4.726 3.00e-06 ***
## tax          -0.011778   0.003372  -3.493 0.000521 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

Using forward selection results in the following variables: lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad and tax

e) Do different selection algorithms find the same subset? Which variables are selected?

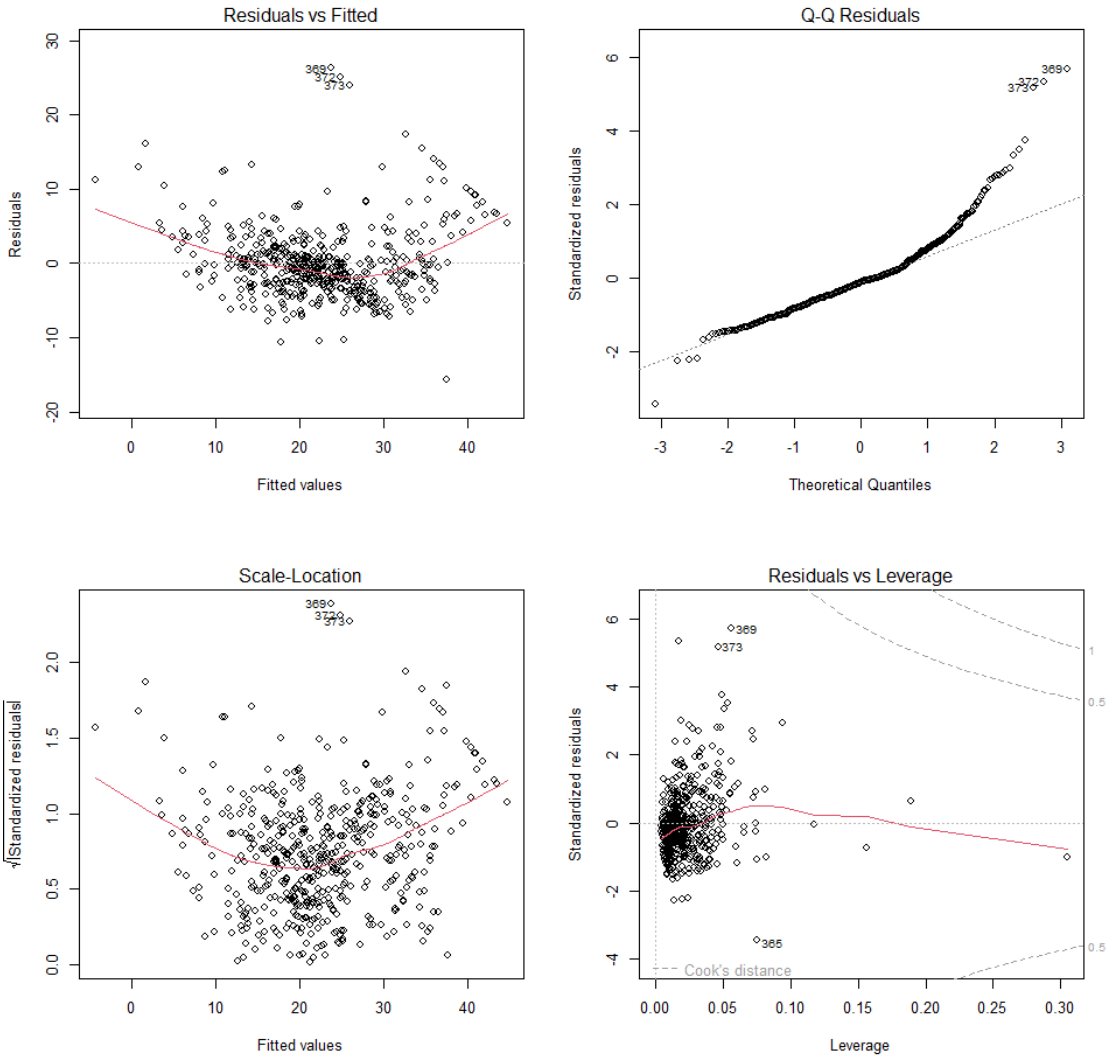
Yes, different selection algorithms find the same subset. The selected variables are: lstat, rm, ptratio, dis, nox, chas, black, zn, crim, rad and tax.

f) Does variable selection improve the R-square? How about the adjusted R-square?

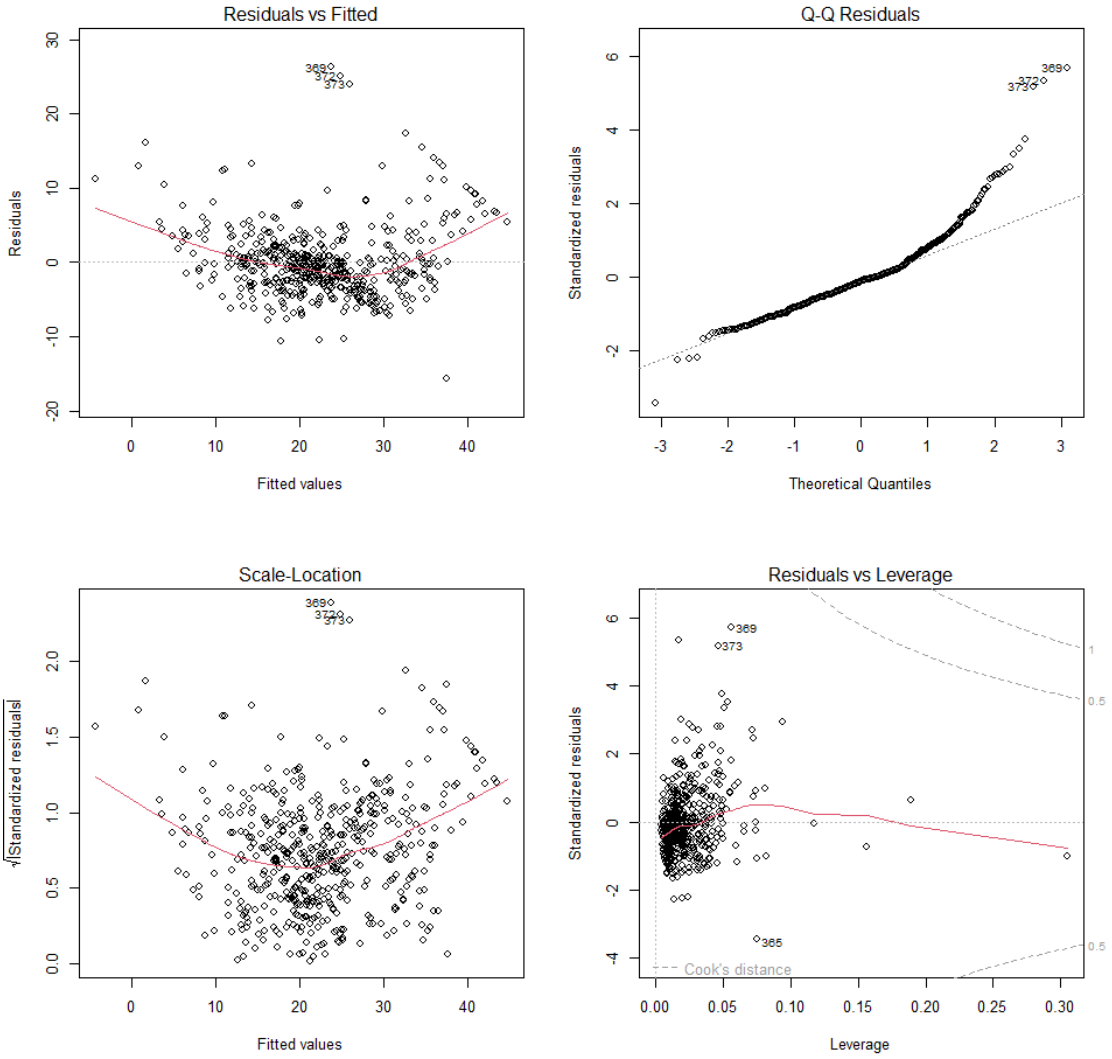
The variable selection yields a multiple R-squared of 0.7406 and an adjusted R-squared of 0.7348. This is compared to the model using all variables with a multiple R-squared of 0.7414 and an adjusted R-squared of 0.734, which indicates that variable selection decreases the multiple R-squared and increases the adjusted R-squared . This suggests that the selected variables have a stronger relationship with the outcome compared to including all variables in the model. The adjusted R-squared takes into account the number of variables in the model, and by selecting only relevant variables, it helps to improve the model's overall fit and reduce potential overfitting.

Q6 Residual plot. Please plot the residual plots of multiple linear regression model fitted in Problem 5 and answer the following question.

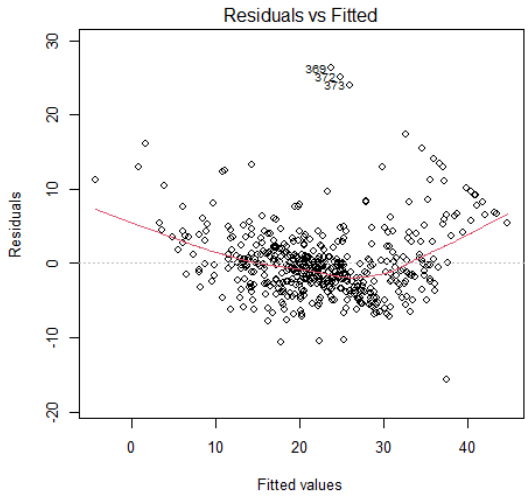
```
par(mfrow=c(2,2))
plot(select.forward)
```

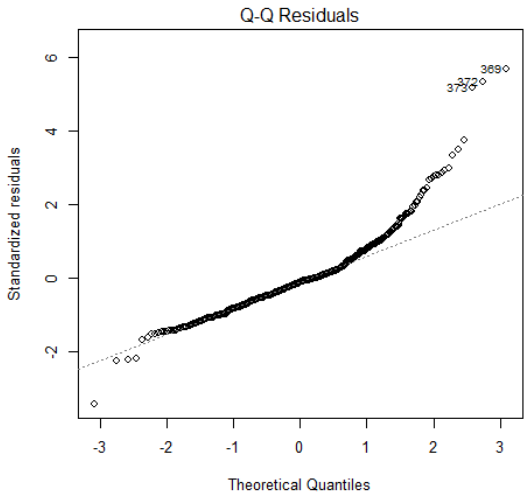


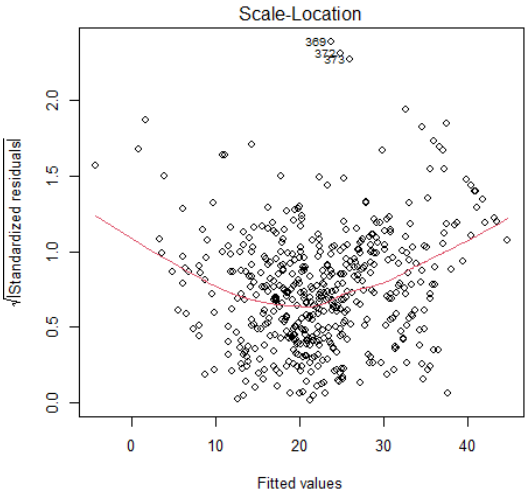
```
par(mfrow=c(2,2))
plot(select.backward)
```

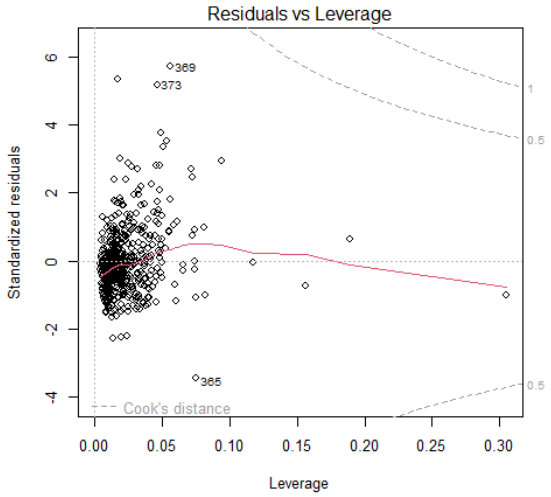



```
par(mfrow=c(2,2))
plot(select.mixed)
```

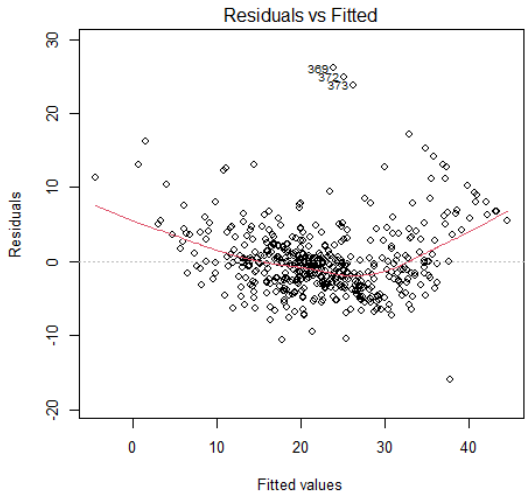


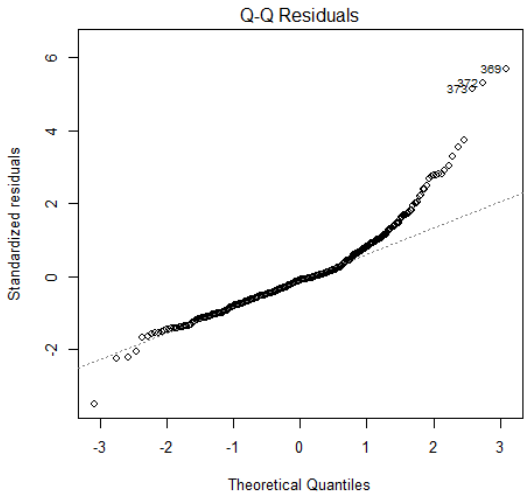


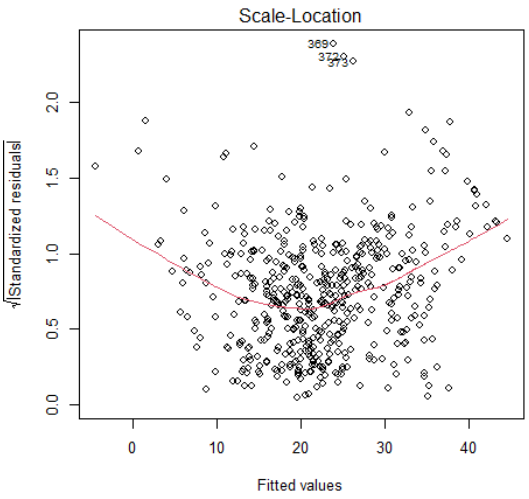


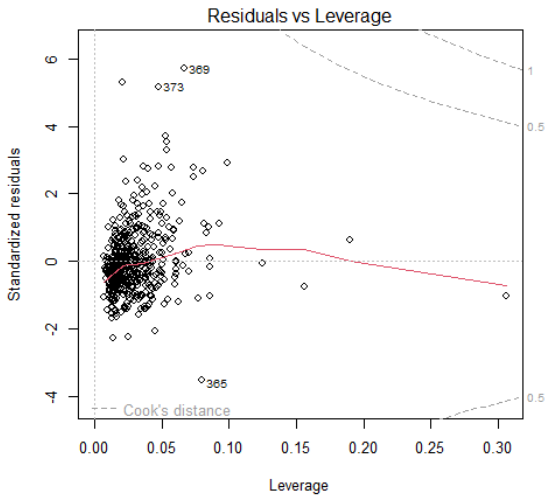



```
par(mfrow =c(2,2))
plot(fit.multiple)
```







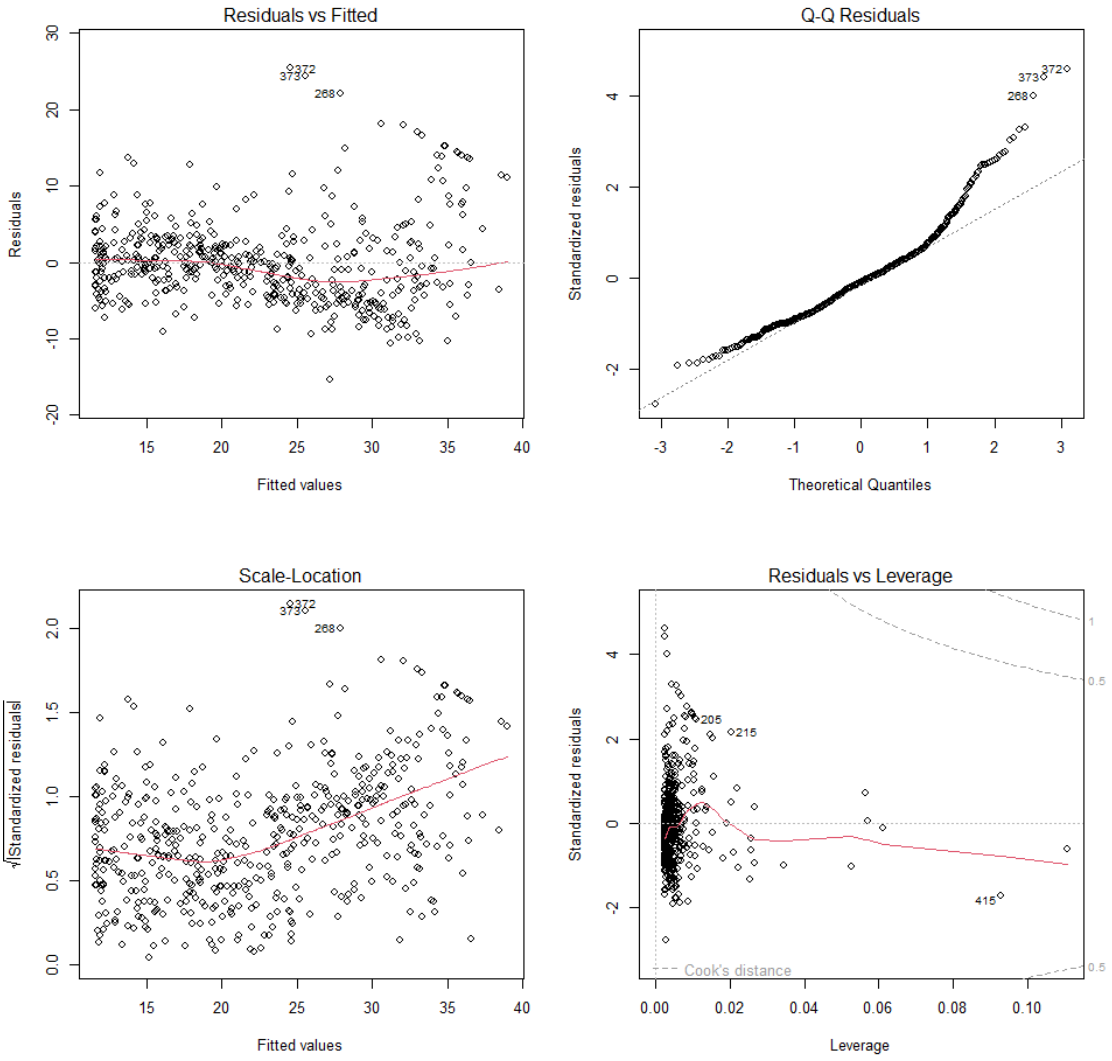


Q7 Use non-linear transformation to include lstat^2.

```
fit.nonlinear <- lm(medv~lstat + I(lstat^2) ,data = data)
summary(fit.nonlinear)

##
## Call:
## lm(formula = medv ~ lstat + I(lstat^2), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.2834  -3.8313  -0.5295   2.3095  25.4148
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 42.862007   0.872084   49.15   <2e-16 ***
## lstat       -2.332821   0.123803  -18.84   <2e-16 ***
```

```
## I(lstat^2)   0.043547   0.003745   11.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.524 on 503 degrees of freedom
## Multiple R-squared:  0.6407, Adjusted R-squared:  0.6393
## F-statistic: 448.5 on 2 and 503 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(fit.nonlinear)
```



a) Is the model improved?

Compare to fit.simple model with a multiple R-squared of 0.5441 and adjusted R-squared of 0.5432. The fit.nonlinear performed better with an R-squared of 0.6407 and an adjusted R-squared of 0.6393.

The Residual standard error also significantly decreased from 6.216 to 5.524.

b) Is the nonlinear effect significant?

The nonlinear effect is not significant because its corresponding coefficient is very small (0.043547).

c) Use the residual plot to see if the nonlinear relationship is solved.

From the residual plot we can see that the residuals are not evenly distributed, thus the nonlinear relationship is not solved. We can confirm the result with the Breusch-Pagan test.

```
##
##  studentized Breusch-Pagan test
##
bptest(fit.nonlinear)

##
##  studentized Breusch-Pagan test
##
## data:  fit.nonlinear
## BP = 48.74, df = 2, p-value = 2.608e-11
```

Using the Breusch-Pagan test, we obtain the p-value of 2.608e-11 which is less than 0.01. Thus, we reject Ho (The variance of the errors is constant -homoscedasticity) and conclude that there is heteroscedasticity between error terms. Since there is still heteroscedasticity between residuals, the nonlinear relationship is not solved.

Q8 Include the interaction term lstat X black.(so what are these processes where we add non linearity to a lineary model.why are they important)

The formula lstat * black is equivalent to lstat + black + lstat:black, so it includes the main effects of lstat and black as well as their interaction term.

Here's what the terms represent:

- The main effect of lstat represents the change in the response variable for a one-unit change in lstat, assuming black is 0.
- The main effect of black represents the change in the response variable for a one-unit change in black, assuming lstat is 0.
- The interaction term lstat:black represents the additional change in the response variable when both lstat and black are simultaneously nonzero. The presence of an interaction term suggests that the effect of lstat on the response variable is modified by the level of black.

The interaction term are important because they allow us to capture more complex relationships between variables. By including interaction terms, we can account for situations where the effect of one variable on the outcome depends on the value of another variable. This helps to improve the model's predictive accuracy and better represent real-world scenarios where relationships may not be strictly linear. Additionally, including

interaction terms can also help to detect and account for potential confounding or moderation effects in the data.

```
fit.interact <- lm(medv~lstat*black,data = data)
summary(fit.interact)

##
## Call:
## lm(formula = medv ~ lstat * black, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.666  -3.918  -1.146   1.936  24.830
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.3344653  3.7527308   4.886 1.39e-06 ***
## lstat       -0.2607734  0.1772904  -1.471 0.141949
## black        0.0426663  0.0098329   4.339 1.73e-05 ***
## lstat:black -0.0018059  0.0004758  -3.796 0.000165 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.109 on 502 degrees of freedom
## Multiple R-squared:  0.5614, Adjusted R-squared:  0.5588
## F-statistic: 214.2 on 3 and 502 DF,  p-value: < 2.2e-16
```

a) Is the model improved?

With a multiple R-squared of 0.5614 and an adjusted R-squared of 0.5588. Compared to fit.nonlinear model, the fit.interact model does not improved the performance.

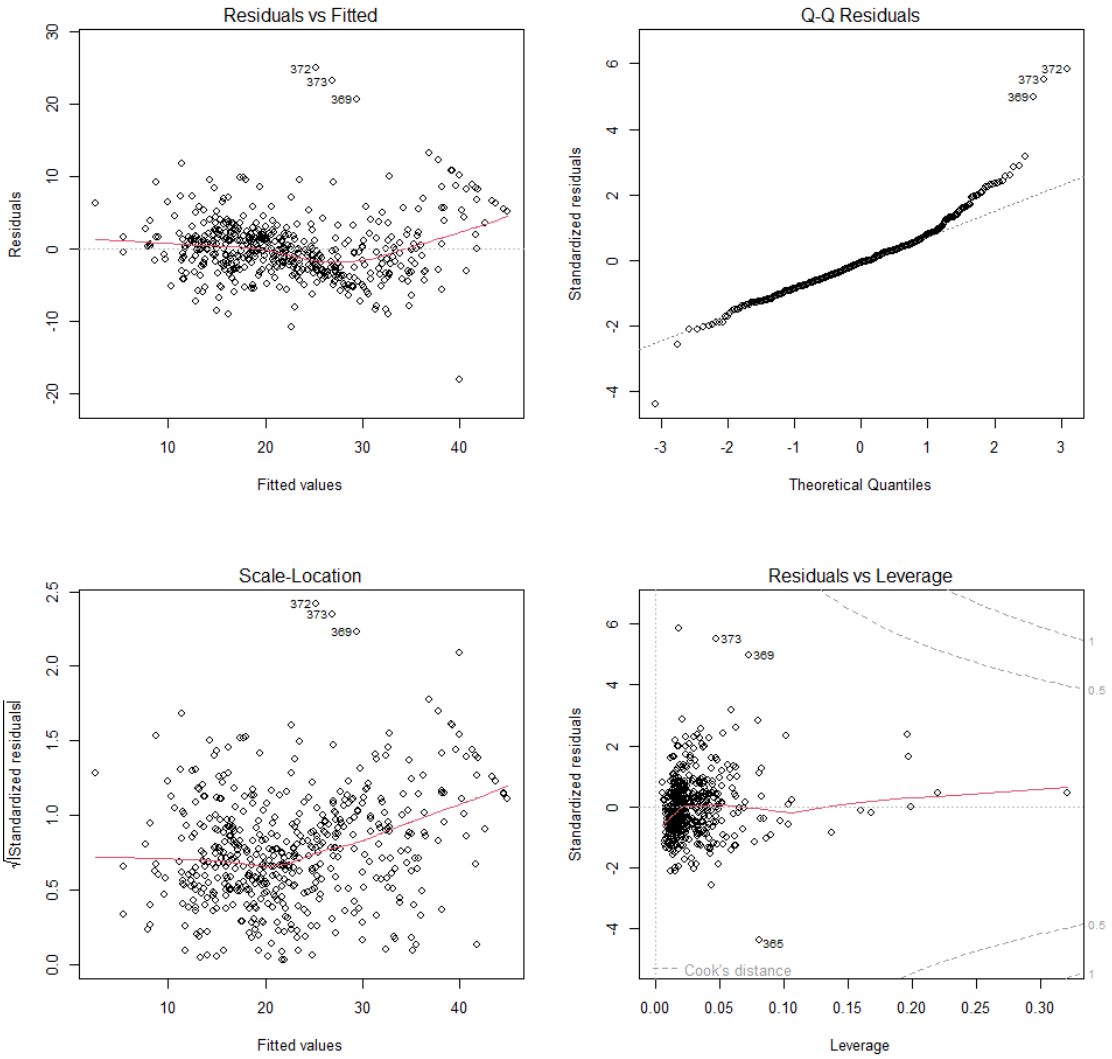
b) Is the interaction effect significant?

The interaction effect is not significant because its corresponding coefficient is very small (-0.0018059).

c) Include both nonlinear term in Q7 and interaction term, answer a) and b).

```
fit.interact.nonlinear <- lm(medv~.+I(lstat^2)-indus-age+lstat*black,data =
data)
summary(fit.interact.nonlinear)

##
## Call:
## lm(formula = medv ~ . + I(lstat^2) - indus - age + lstat * black,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0007  -2.6072  -0.2572   1.8976  24.8574
##
```

```
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.461e+01  5.297e+00   8.422 4.11e-16 ***
## X           -2.459e-03  1.865e-03  -1.319 0.187923
## crim        -1.513e-01  3.003e-02  -5.040 6.54e-07 ***
## zn           2.300e-02  1.257e-02   1.829 0.067961 .
## chas         2.569e+00  7.753e-01   3.314 0.000989 ***
## nox         -1.368e+01  3.232e+00  -4.233 2.75e-05 ***
## rm           3.294e+00  3.762e-01   8.756  < 2e-16 ***
## dis         -1.357e+00  1.691e-01  -8.024 7.61e-15 ***
## rad          2.934e-01  5.942e-02   4.937 1.09e-06 ***
## tax         -9.041e-03  3.117e-03  -2.901 0.003891 **
## ptratio     -7.793e-01  1.184e-01  -6.583 1.19e-10 ***
## black        1.239e-03  7.873e-03   0.157 0.874970
## lstat       -1.834e+00  2.079e-01  -8.822  < 2e-16 ***
## I(lstat^2)   3.424e-02  3.431e-03   9.980  < 2e-16 ***
## black:lstat  3.504e-04  3.778e-04   0.927 0.354175
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.294 on 491 degrees of freedom
## Multiple R-squared:  0.788,  Adjusted R-squared:  0.782
## F-statistic: 130.4 on 14 and 491 DF,  p-value: < 2.2e-16

par(mfrow=c(2,2))
plot(fit.interact.nonlinear)
```



d) Is the model improved?

With a multiple R-squared of 0.788 and an adjusted R-squared of 0.782. Compared to fit.nonlinear model, the fit.interact.nonlinear model improved the performance.

e) Is the interaction effect significant?

The interaction effect is not significant because its corresponding coefficient is very small (3.504e-04).

Q9 Apply K-nearest neighbor regression model on this dataset and find the optimal K.

Step-1: Randomly separate the dataset into training and test data

we used sample function to randomly reorder the samples and use first 400 samples as training and the remaining samples as test.

```
# write comments on what is happening below
randid <- sample(c(1:nrow(data)))
Boston.train <- data[randid[c(1:400)],]
Boston.test <- data[randid[c(401:506)],]
lstat.train <- Boston.train['lstat']
medv.train <- Boston.train['medv']
lstat.test <-Boston.test['lstat']
medv.test <- Boston.test['medv']
```

Step-2: Use training data to predict medv values at test data and select the best K

Predict the medv under K=1,5,10,50,100,250

```
library(FNN)
ks <- c(1,5,10,50,100,250)
model_list <- list()

#repeat or automate till you get an optimum k value
for (k in ks){
  model <- knn.reg(train = lstat.train, test = lstat.test, y = medv.train, k
= k)
  model_list <- append(model_list, list(model))
}
```

Step-3: Find the best K

Calculate the mse under each K value

```
mse_list = list()

for (model in model_list){
  mse = sum((model$pred-medv.test)^2)/106
  mse_list <- append(mse_list, list(mse))
}

k_best_indx <- which.min(mse_list)
k_best <- ks[k_best_indx]
mse_best <- mse_list[k_best_indx]

glue('The best K is {k_best} with min MSE of {mse_best}')

## The best K is 50 with min MSE of 34.6540052830189
```