MSDS660 - Week 2: Simple Linear Regression

Christy PearsonMay 21, 2018

For this week's assignment we will be using the Boston housing dataset from the MASS package. We need to cover the following steps (in any order): 1. Describe the dataset? 2. Perform EDA 3. Perform pairwise scatterplots 4. Independent vs Dependent variable plot of choice 5. State both Ho and Ha 6. Show the linear regression 7) Explain how to read the results

Preparing the current environment

```
rm(list = ls()) # Prevents results from previous runs being carried over into new runs.
graphics.off()  # Clears the graphic plots window
cat("\014")     # Clears the console
```

```
setwd("C:/Users/Creat/OneDrive/Documents/MSDS660/Week 2")
getwd()
```

```
## [1] "C:/Users/Creat/OneDrive/Documents/MSDS660/Week 2"
```

Package Installations

```
# https://stackoverflow.com/questions/34739681/unable-to-move-temporary-installation-when-installing-de
# Sometimes needed when "cannot move temporary installation" error
#debug(utils:::unpackPkgZip)

library(MASS)# https://cran.r-project.org/web/packages/MASS/MASS.pdf
```

Simple Linear Regression Exercise:

1) Describe your data set dataset: Boston from package MASS 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. CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise) 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 centres RAD - index of accessibility to radial highways TAX - full-value property-tax rate per $10,000 PTRATIO - pupil-teacher ratio by town B - 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 number of observations expected: 506 documentation at https://www.cs.toronto.edu/~delve/data/boston/bostonDetail.html

```
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
```

```
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 ...
```

```
summary(Boston)
```

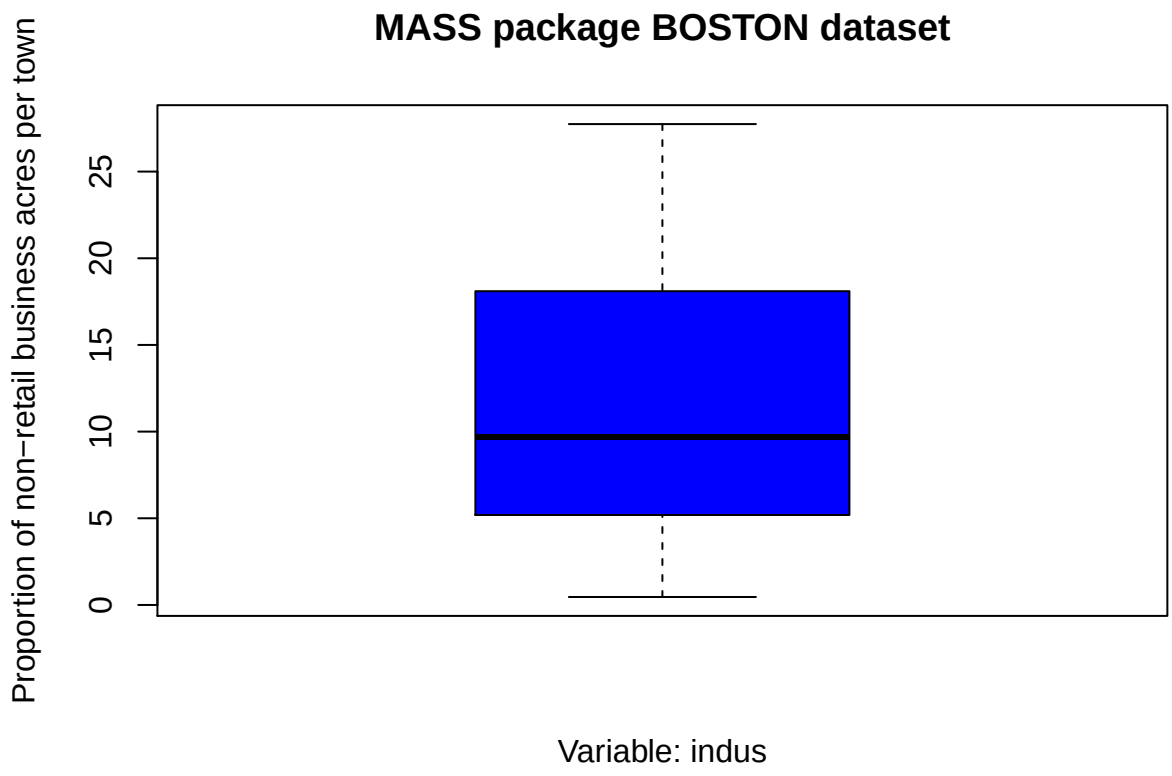
```
##       crim                zn              indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  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
```

2) EDA by plotting the graphs, the distributions and so on. Then interpret.

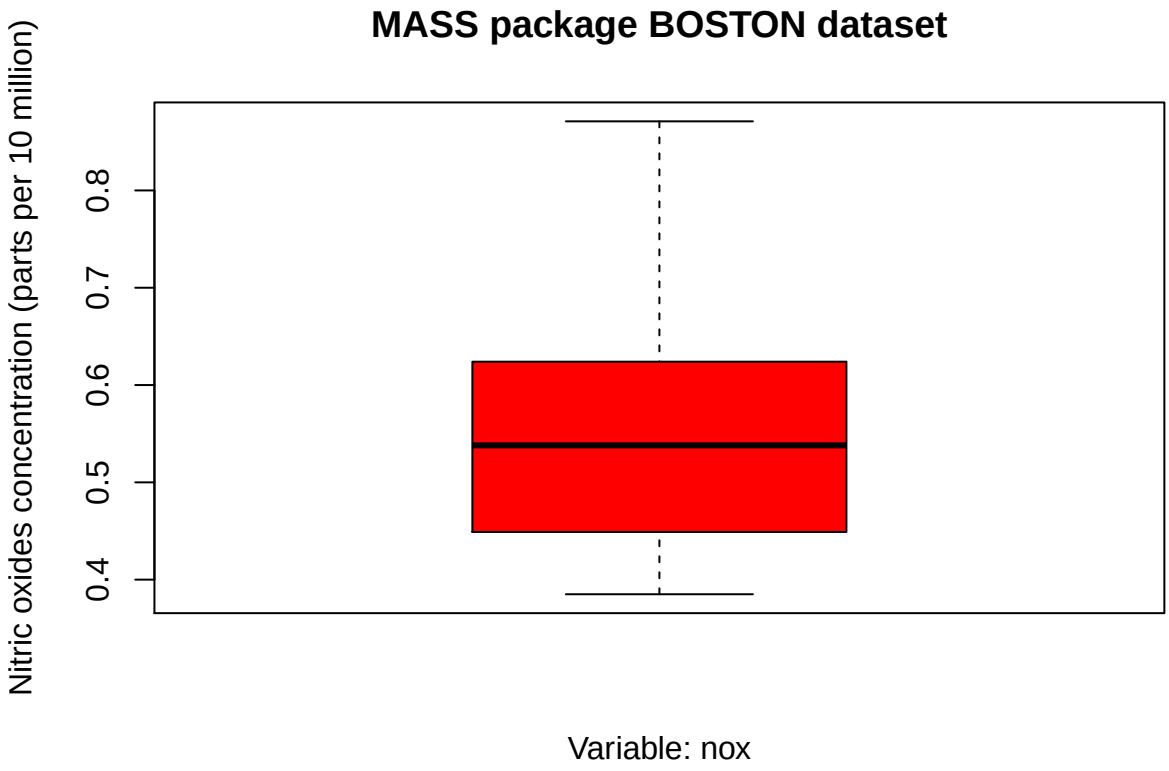
Let's look at variable: medv

```
attach(Boston)
boxplot(indus, col="blue",
```

```
        main="MASS package BOSTON dataset",
        xlab="Variable: indus",
        ylab="Proportion of non-retail business acres per town")
```

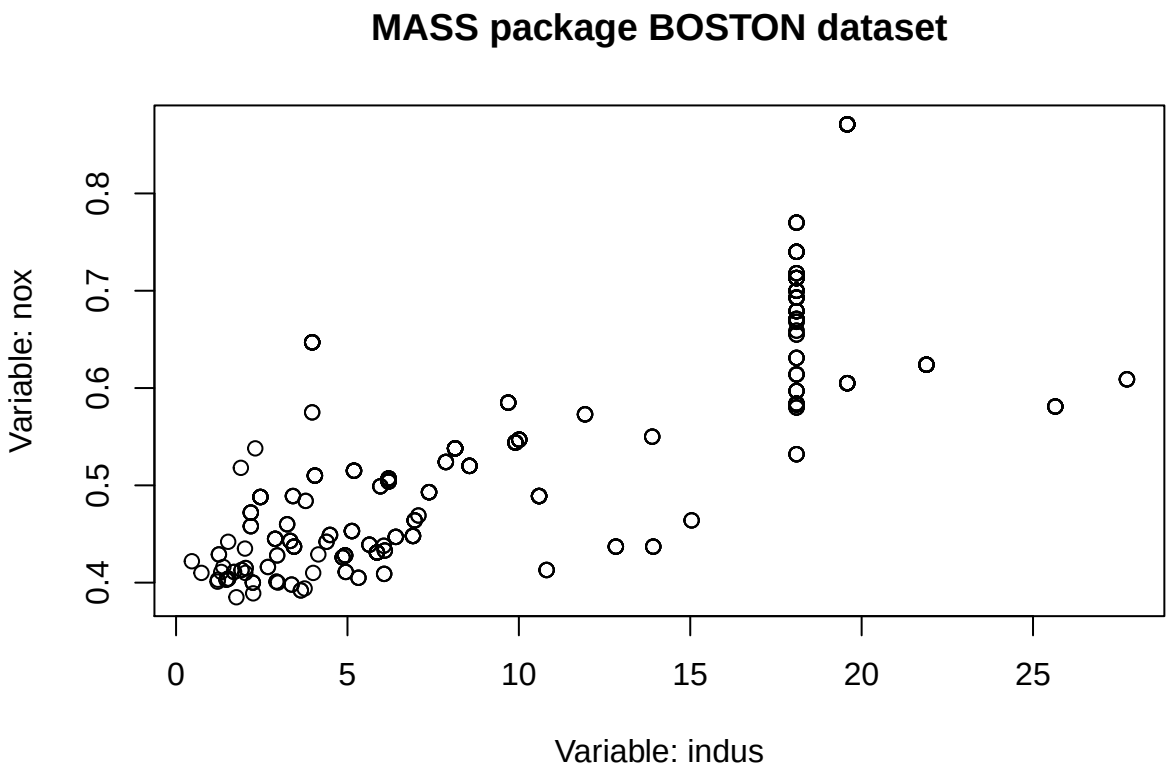



```
boxplot(nox, col="red",
        main="MASS package BOSTON dataset",
        xlab="Variable: nox",
        ylab="Nitric oxides concentration (parts per 10 million)")
```



Now let's look at how these two variable interact with each other. Looking at scatterplots: indus vs nox

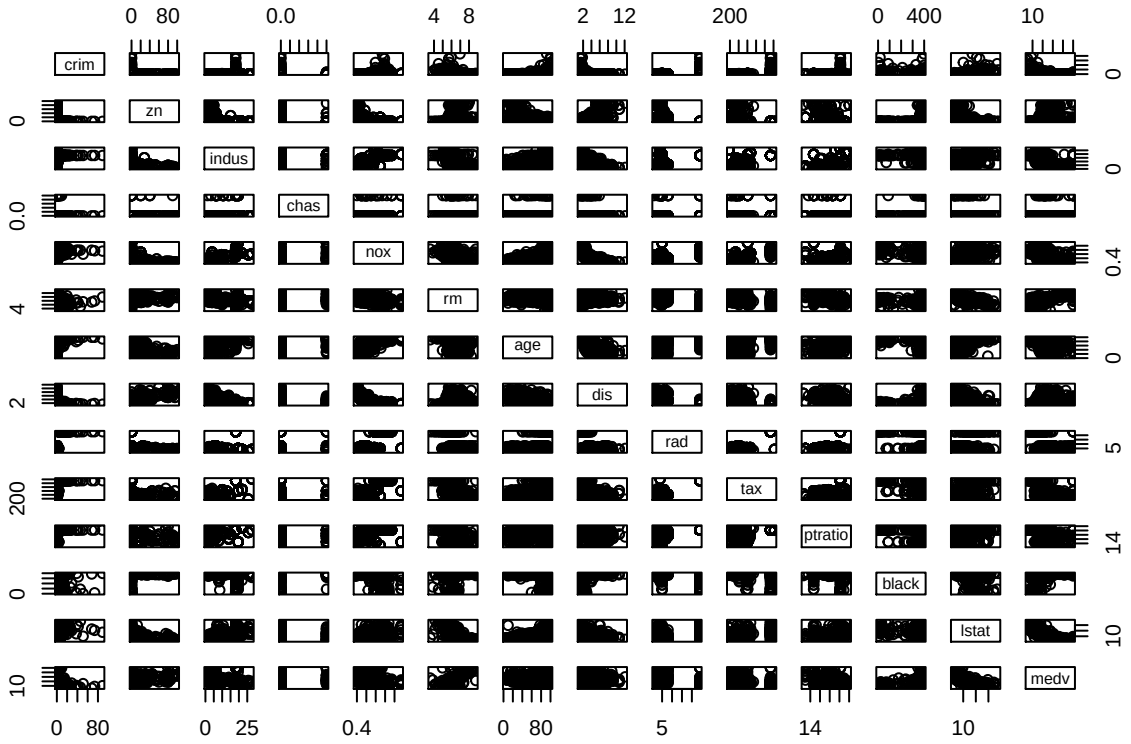
```
plot(indus, nox,
     main="MASS package BOSTON dataset",
     xlab="Variable: indus",
     ylab="Variable: nox")
```



Rough interpretation of this graph, the lower values of indus also have a lower value for nox. However, there is something odd that happens at indus = 18 that results in a wide range of nox values.

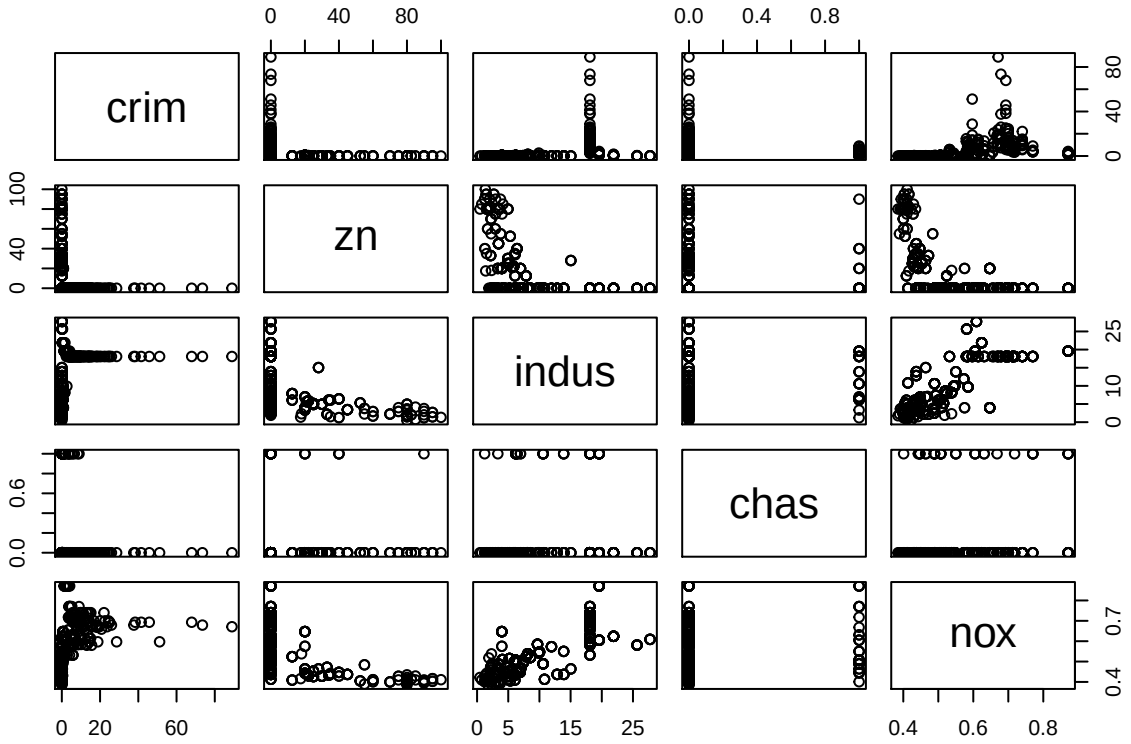
3) Perform pairwise scatterplots

```
pairs(Boston)
```

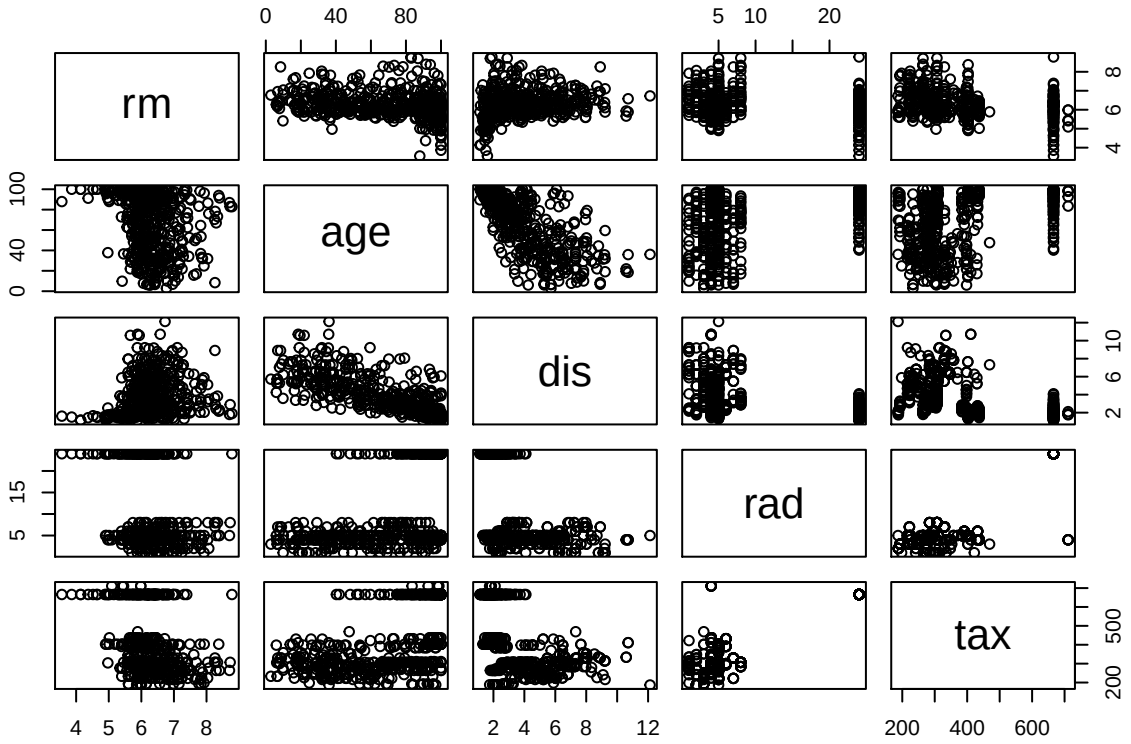


That's really hard to understand, let's break this down a bit

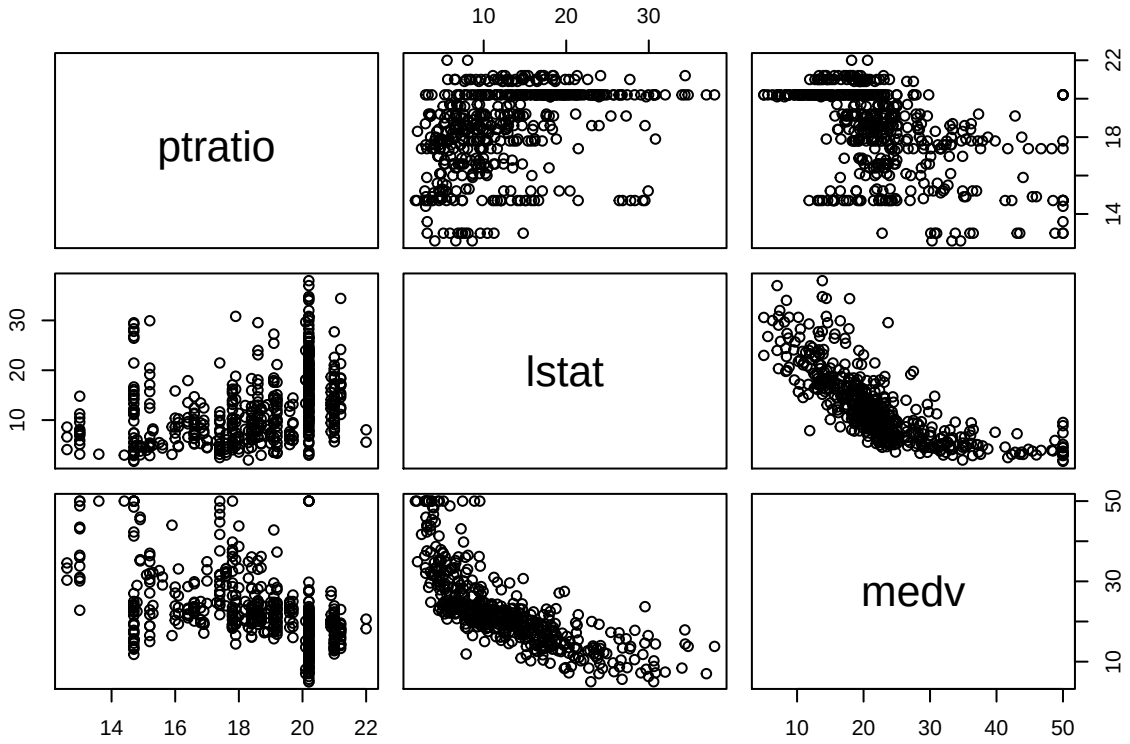
```
pairs(~crim+zn+indus+chas+nox,data=Boston)
```




```
pairs(~ rm+age+dis+rad+tax,data=Boston)
```



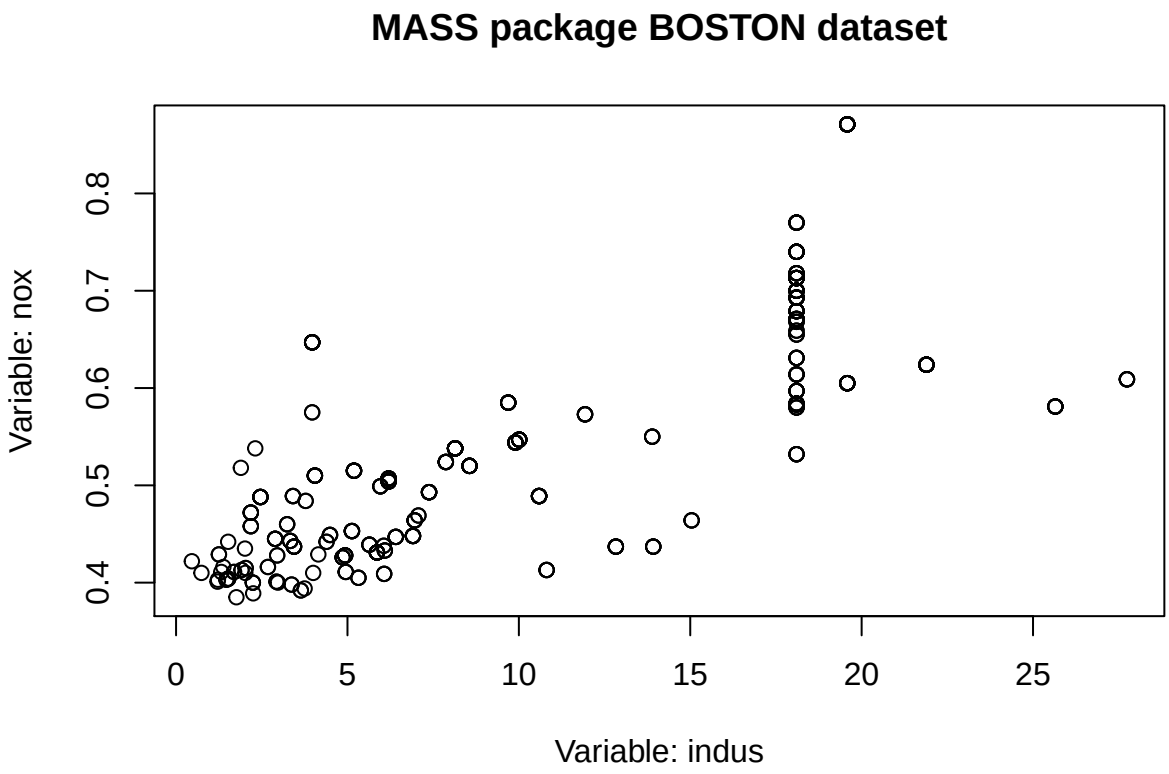

```
pairs(~ ptratio+lstat+medv,data=Boston)
```



The smaller plots are easier to look at and they don't represent all of the possible pariings for this dataset, but for our exercises this week we have enough to work with. For example, the variables lsat and medv show a negative correlation while the variables indus and nos show a positive corrleation. Finally, the variables chas and zn is an example of no correlation.

4) Independent vs Dependent variable plot of choice Let's stick with indus and nox as our two variables, here's the plot again

```
plot(indus, nox,
     main="MASS package BOSTON dataset",
     xlab="Variable: indus",
     ylab="Variable: nox")
```

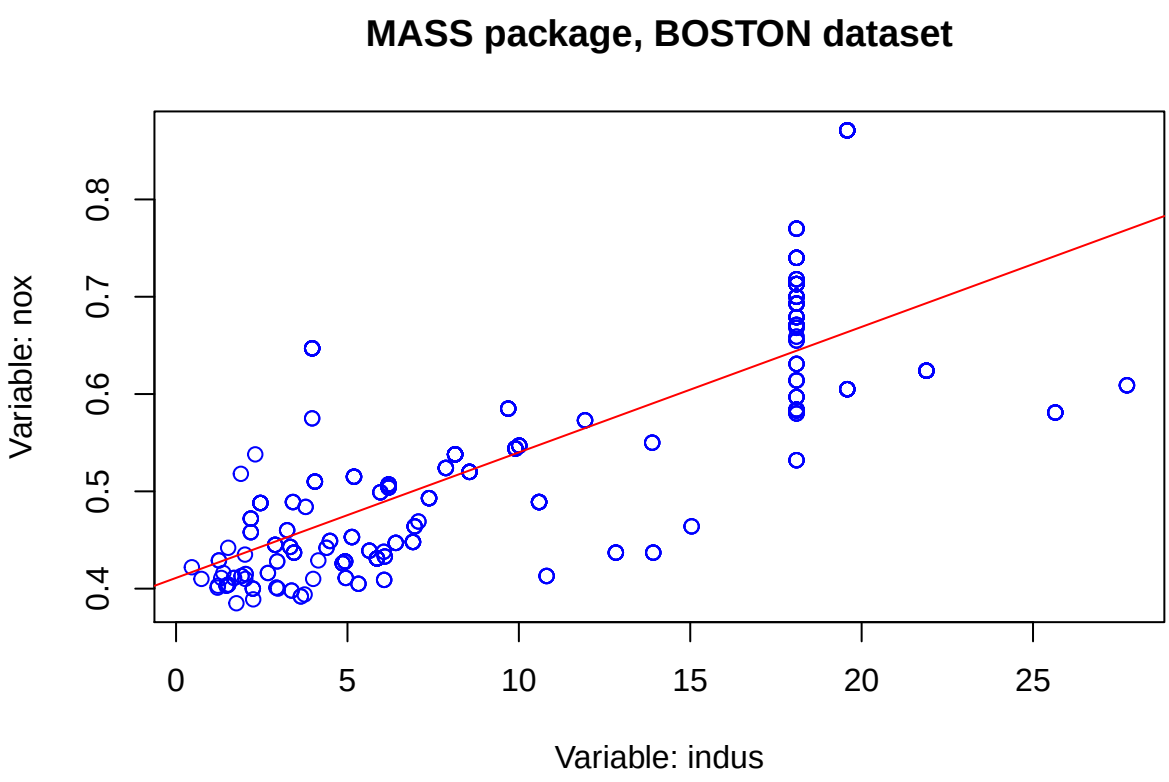


Linear regression Hypothesis statements: Ho: There is no relationship between indus and nox, meaning that as indus increases, there is no relationship to the values for nox. Ha: A relationship does exist between indus and nox, meaning that as indus increases, the values for nox increase too. Calculate the regression

```
boston.lm <- lm(nox ~ indus, Boston)
```

Plot indus vs nox again and add the lm line this time

```
plot(indus, nox,
     main="MASS package, BOSTON dataset",
     xlab="Variable: indus",
     ylab="Variable: nox", col="blue")
abline(boston.lm, col="red")
```



According to this version fot ehscatter plot, there is a positive relationship between indus and nox. Let's look at the numbers now.

```
boston.lm
```

```
##
## Call:
## lm(formula = nox ~ indus, data = Boston)
##
## Coefficients:
## (Intercept)        indus
##      0.4110       0.0129
```

The y-intercept is .4110 and the slop of the lm is .0129. Not much of a slope

```
summary(boston.lm)
```

```
##
## Call:
## lm(formula = nox ~ indus, data = Boston)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.160898 -0.052175 -0.001458  0.037011  0.207398
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4110442  0.0063521   64.71   <2e-16 ***
## indus       0.0128988  0.0004858   26.55   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.07489 on 504 degrees of freedom
## Multiple R-squared:  0.5832, Adjusted R-squared:  0.5823
## F-statistic: 705.1 on 1 and 504 DF,  p-value: < 2.2e-16
```

The p-value for boston.lm is really low and the assignment said to assume a level of significance of .05. Since p-value $<$ level of signficance, we should reject the null hypothesis(Ho). So, we accept Ha, meaning that there is a realtionship between indus and nox.

Let's see how strong that realtionship is with cor() function

```
vars <-cbind(indus,nox)
cor(vars, use = "all.obs", method = "pearson")
```

```
##           indus       nox
## indus 1.0000000 0.7636514
## nox   0.7636514 1.0000000
```

With a correlation coefficient of .7636, this means that the relationship between indus and nox is fairly strong.