Lab1: Exploratory Data Analysis

Claire Jellison

9/11/2019

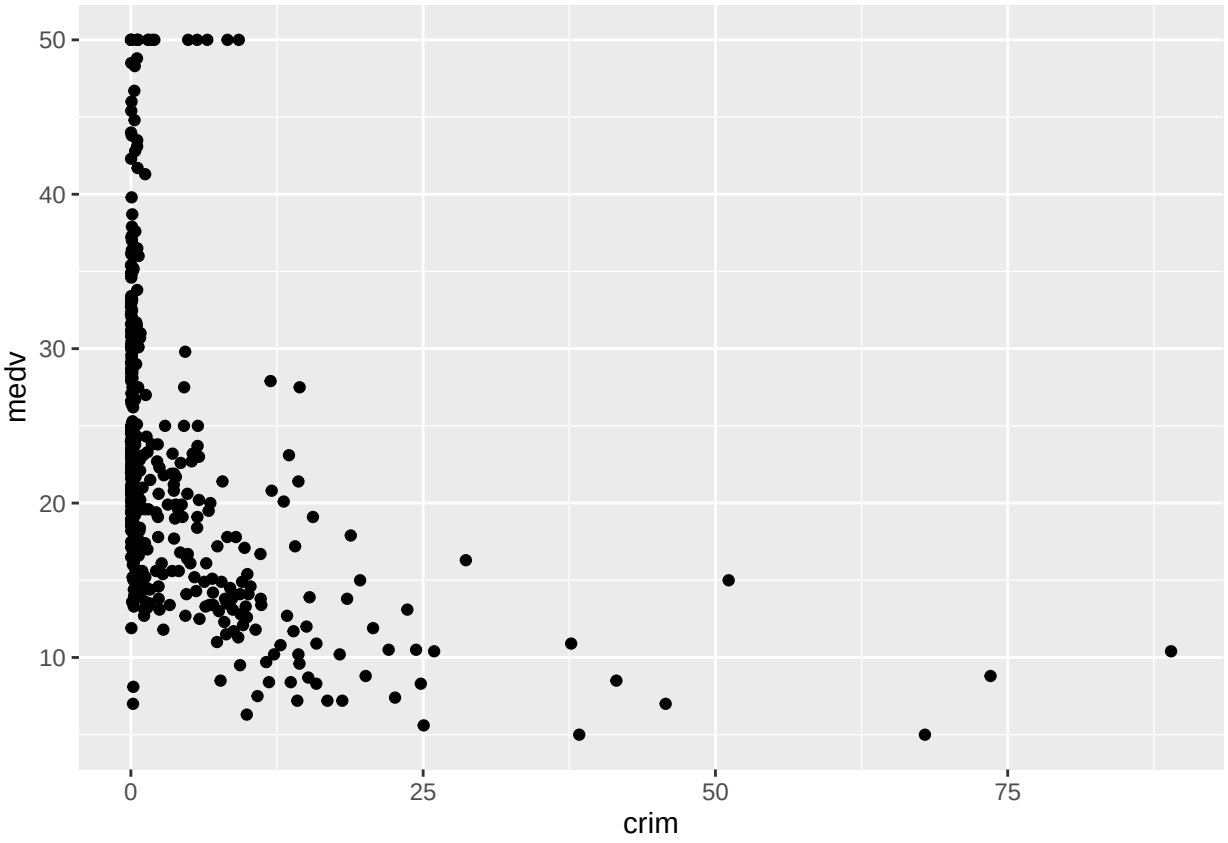
Exercise 1

```
library(MASS)
?Boston
```

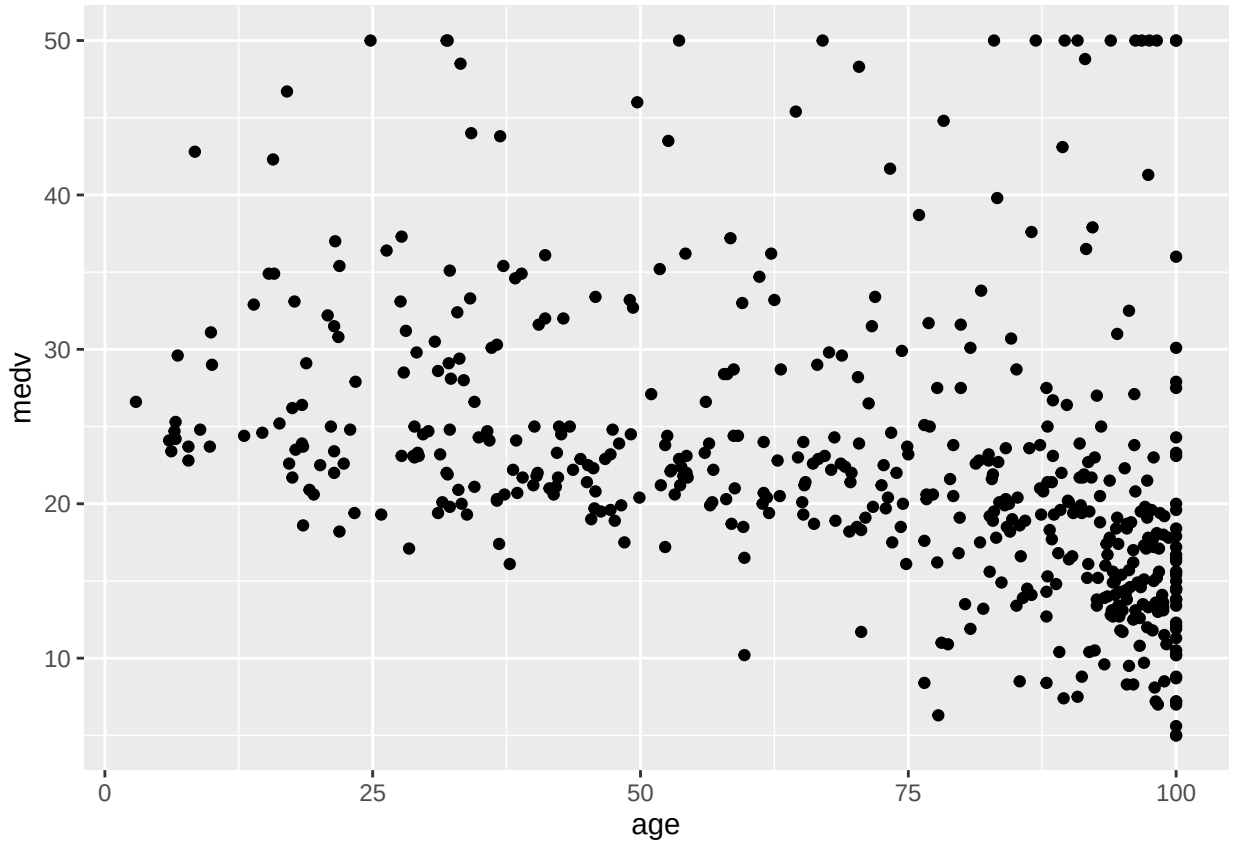
The data is on the housing values i the suburbs of Boston. There are 506 rows and 14 columns in this data set. The rows correspond to each of the observations and the columns include the crime rate, proportion of residential land zoned for lots over 25,000 sq ft, the proportion of non-retail business acres per town, a river dummy variable, the nitro oxide concentration, the average number of rooms per dwelling, the median value of owner occupied homes and several other characteristics of the observations (suburbs).

Exercise 2

```
library(MASS)
library(ggplot2)
# Basic scatter plot
ggplot(Boston, aes(x=crim, y=medv)) + geom_point()
```



```
ggplot(Boston, aes(x=age, y=medv)) + geom_point()
```



In the first graph, we see that when there is higher rates of crime the median value is lower, however this relationship does not appear to be entirely linear. There are also a lot more observations with low crime rates than higher crime rates.

In the second graph, we see that a very large proportion of owner occupied units build prior to 1940 seems to generally corresponds with a lower median value although the relationship is somewhat ambiguous and there is a large amount of variance in the data.

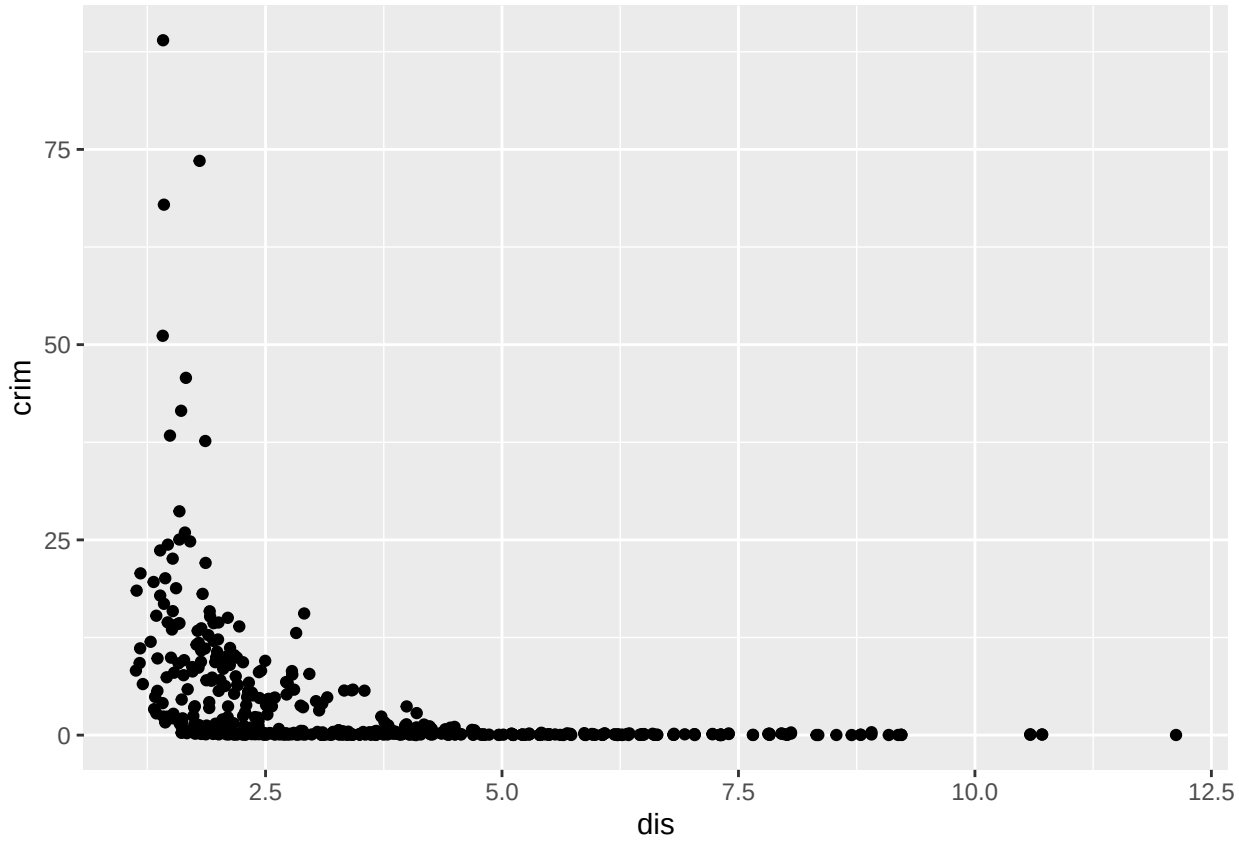
Exercise 3

```
lm <- lm(crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat + medv, d
summary(lm)
```

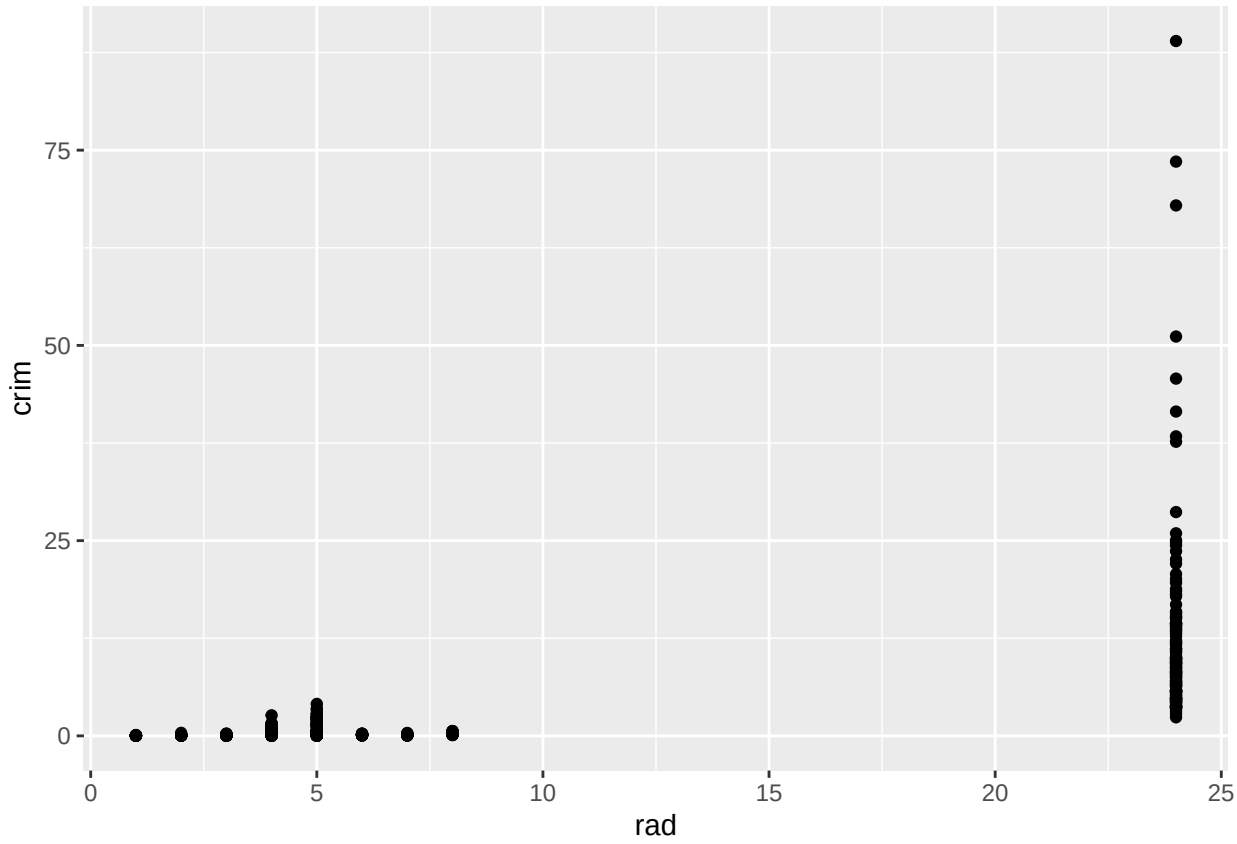
```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354 0.018949 *
```

```
## zn            0.044855   0.018734   2.394 0.017025 *
## indus        -0.063855   0.083407  -0.766 0.444294
## chas         -0.749134   1.180147  -0.635 0.525867
## nox         -10.313535   5.275536  -1.955 0.051152 .
## rm            0.430131   0.612830   0.702 0.483089
## age           0.001452   0.017925   0.081 0.935488
## dis          -0.987176   0.281817  -3.503 0.000502 ***
## rad           0.588209   0.088049   6.680 6.46e-11 ***
## tax          -0.003780   0.005156  -0.733 0.463793
## ptratio      -0.271081   0.186450  -1.454 0.146611
## black        -0.007538   0.003673  -2.052 0.040702 *
## lstat         0.126211   0.075725   1.667 0.096208 .
## medv         -0.198887   0.060516  -3.287 0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

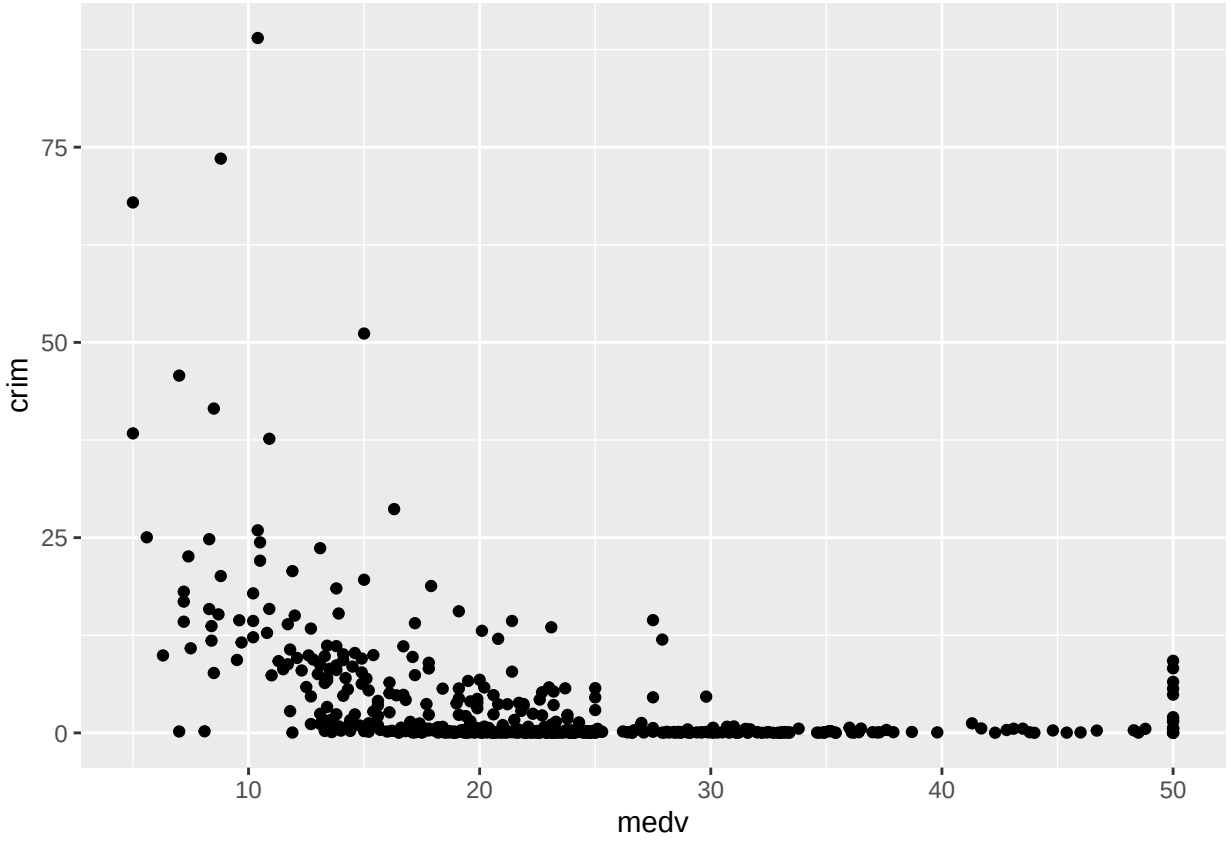
```
ggplot(Boston, aes(x=dis, y=crim)) + geom_point()
```




```
ggplot(Boston, aes(x=rad, y=crim)) + geom_point()
```




```
ggplot(Boston, aes(x=medv, y=crim)) + geom_point()
```



From this it appears that a greater distance to a Boston employment center is correlated with lower per capita crime rate, a higher rad index to a higher crime rate, and a higher median value to a lower crime rate. These relationships can also be seen in the scatterplot. Some other variables also may have an effect on the crime rate but these are the ones with the lowest p values.

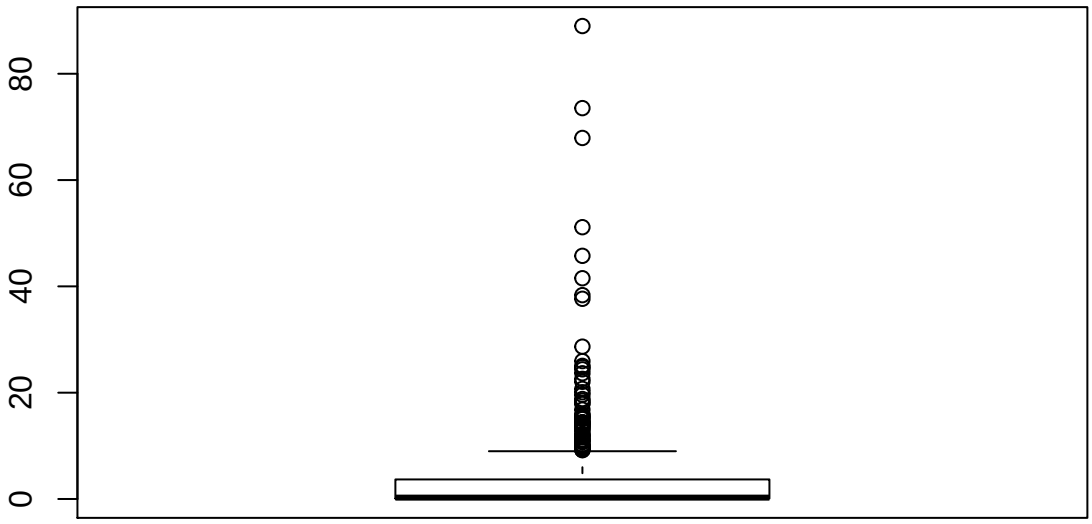
Exercise 4

Are there any suburbs of Boston that appear to have particularly high crime rates? Tax rate? Pupil-teacher ratios? Comment on the range of each predictor.

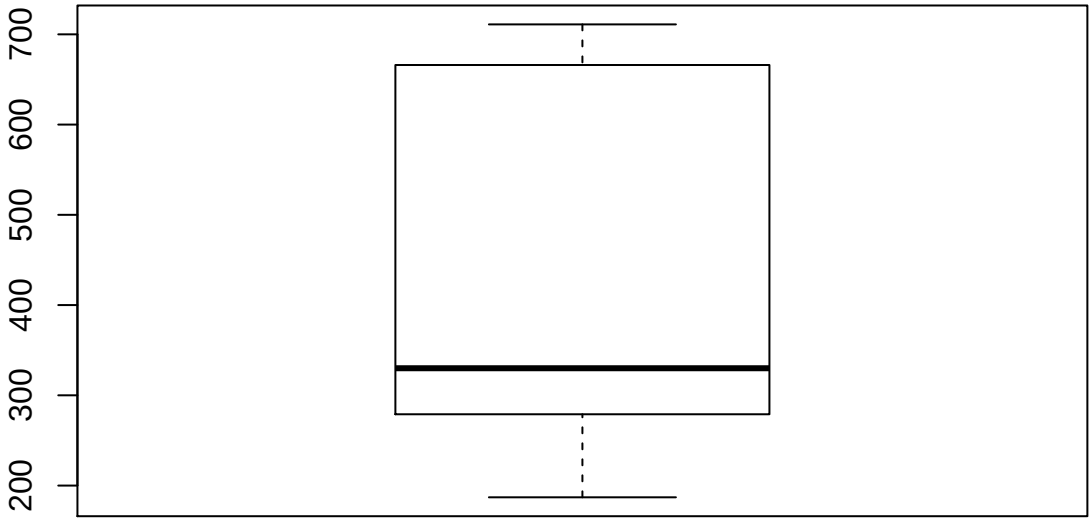
```
library(plyr)
t(sapply(Boston, range))
```

```
##             [,1]     [,2]
## crim      0.00632  88.9762
## zn        0.00000 100.0000
## indus     0.46000  27.7400
## chas      0.00000   1.0000
## nox       0.38500   0.8710
## rm        3.56100   8.7800
## age       2.90000 100.0000
## dis       1.12960  12.1265
## rad       1.00000  24.0000
## tax     187.00000 711.0000
## ptratio  12.60000  22.0000
## black     0.32000 396.9000
## lstat     1.73000  37.9700
## medv      5.00000  50.0000
```

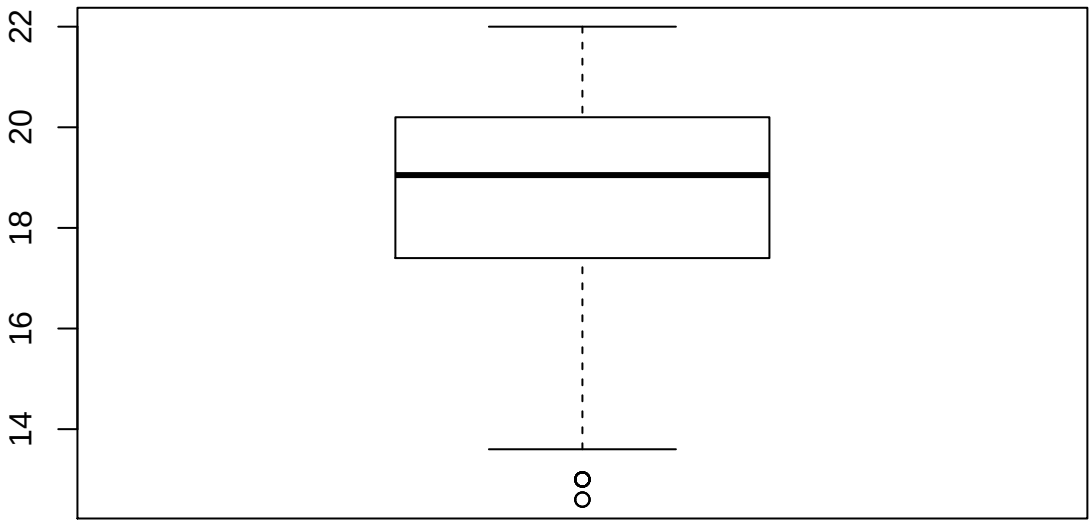
```r
boxplot(Boston$crim)
```




```r
boxplot(Boston$tax)
```




```r
boxplot(Boston$ptratio)
```



The crime rate per capita has a very large range from close to nothing all the way to 89. The full-value property-tax rate goes from 187 to 711 per $10,000. The pupil to teach ratio goes from around 13 pupils to 1 teacher to 22 pupils per teacher. From the boxplots it looks as though there are some areas that have unusually high crime rates while there are some places with unusually low pupil teacher ratios.

Exercise 5

How many of the suburbs in this data set bound the Charles river?

```
sum(Boston$chas == '1', na.rm=TRUE)
```

```
## [1] 35
```

```
sum(Boston$chas == '0', na.rm=TRUE)
```

```
## [1] 471
```

This shows that 35 of the suburbs are bound by the Charles River while the other 471 are not.

Exercise 6

```
median(Boston$ptratio, na.rm = TRUE)
```

```
## [1] 19.05
```

The median pupil to teach ratio among the towns in this data set is around 19 pupils to 1 teacher.

Exercise 7

If you want to build a model to predict the average value of a home based on the other variables, what is your output/response? What is your input?

```
lm <- lm(crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio + black + lstat + medv, d
lm
```

```
##
## Call:
## lm(formula = crim ~ zn + indus + chas + nox + rm + age + dis +
##     rad + tax + ptratio + black + lstat + medv, data = Boston)
##
## Coefficients:
```

```
## (Intercept)          zn       indus        chas         nox
##   17.033228    0.044855   -0.063855   -0.749134  -10.313535
##          rm         age         dis         rad         tax
##    0.430131    0.001452   -0.987176    0.588209   -0.003780
##     ptratio       black       lstat        medv
##   -0.271081   -0.007538    0.126211   -0.198887
```

The output respose variable would be the average value of a home with the input being a given set of characteristics. So you could give the characteristics of a certain home and the model would predict a value of the home based on the relationships it has seen with previous data.