# Intro to Data Exploration

Alan Perez

This has two parts:

- Part 1 uses R for data exploration
- Part 2 uses C++ for data exploration

## Part 1: RStudio Data Exploration

**Instructions:** Follow the instructions for the 10 parts below. If the step asks you to make an observation or comment, write your answer in the white space above the gray code box for that step.

### Step 1: Load and explore the data

- load library MASS (install at console, not in code)
- load the Boston dataframe using data(Boston)
- use str() on the data
- type ?Boston at the console
- Write 2-3 sentences about the data set below

Your commentary here: Upon running the code block I see that there are 506 observations(rows) and 14 variables(columns). I'm given the data pertaining to each row and column along with the help page. The "HELP PAGE" which is given to me after entering "?Boston" gives me detailed information regarding the housing values in the suburbs of Boston.

```
# step 1 code

# importing package
library(MASS)
data(Boston)
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 ...
```

```
?Boston
```

```
## starting httpd help server ... done
```

## Step 2: More data exploration

Use R commands to:

- display the first few rows
- display the last two rows
- display row 5
- display the first few rows of column 1 by combining head() and using indexing
- display the column names

```
# step 2 code
  # display the first few rows
head(Boston)
```

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

```
  # display last 2 rows
tail(Boston, n=2)
```

```
##        crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat
## 505 0.10959  0 11.93    0 0.573 6.794 89.3 2.3889   1 273      21 393.45  6.48
## 506 0.04741  0 11.93    0 0.573 6.030 80.8 2.5050   1 273      21 396.90  7.88
##     medv
## 505 22.0
## 506 11.9
```

```
  #display row 5
Boston[5,]
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio black lstat
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.9  5.33
##   medv
## 5 36.2
```

```
  #display first few rows of column 1 combining head
head(Boston[,1])
```

```
## [1] 0.00632 0.02731 0.02729 0.03237 0.06905 0.02985
```

```
  # display COLUM NAMES
colnames(Boston)
```

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

## Step 3: More data exploration

For the crime column, show:

- the mean
- the median
- the range

```
# step 3 code
# MEAN OF CRIME COLUMN
mean(Boston$crim)
```

```
## [1] 3.613524
```

```
# MEDIAN OF CRIME
median(Boston$crim)
```

```
## [1] 0.25651
```

```
#range of crim
range(Boston$crim)
```

```
## [1]  0.00632 88.97620
```

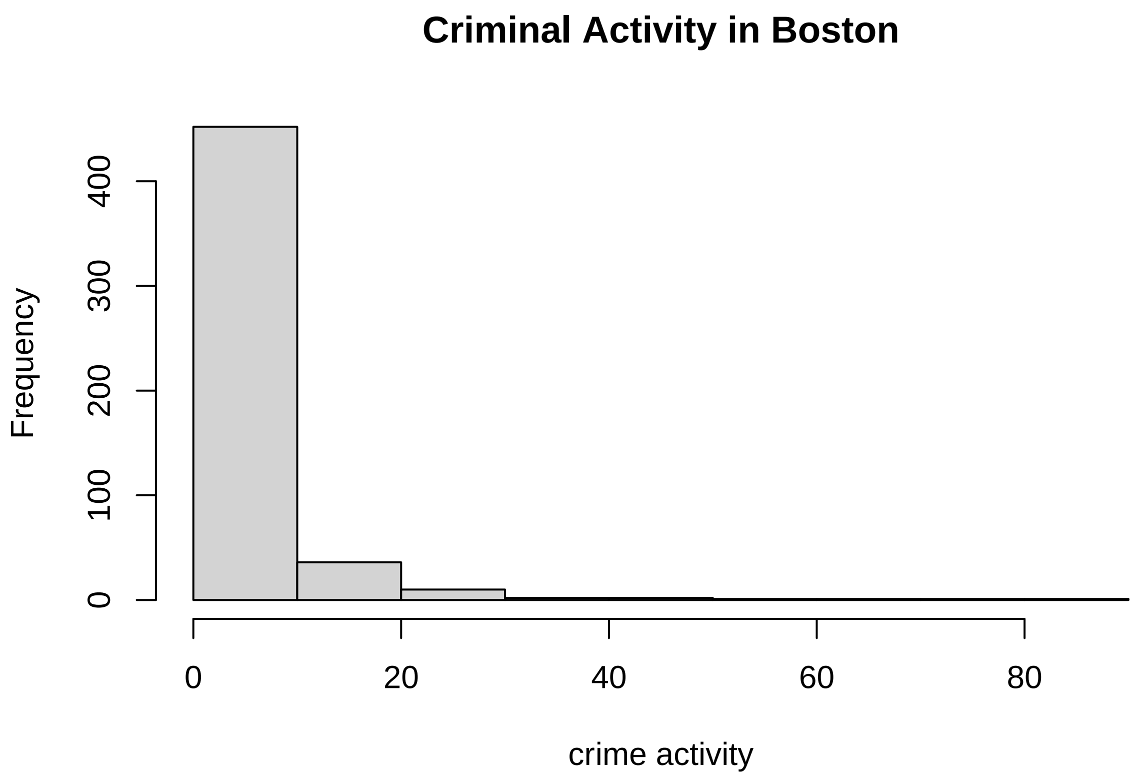
## Step 4: Data visualization

Create a histogram of the crime column, with an appropriate main heading. In the space below, state your conclusions about the crime variable:

Your commentary here: Surprised that we're able to see a visualization of the data that we have. We can see the criminal activity that it peaks instantly and drops off dramatically.

```
# step 4 code

# histogram of Boston, Crime Column
hist(Boston$crim, main="Criminal Activity in Boston", xlab="crime activity")
```



### Step 5: Finding correlations

Use the cor() function to see if there is a correlation between crime and median home value. In the space below, write a sentence or two on what this value might mean. Also write about whether or not the crime column might be useful to predict median home value.

Your commentary here: When correlating the crime and Home column I received a negative value. I believe the reason as to why the MEDIAN HOME VALUE will be dropping is due to the crime.

```
# step 5 code

# Correlation between crime and median home value

cor(Boston$crim, Boston$medv)
```

```
## [1] -0.3883046
```

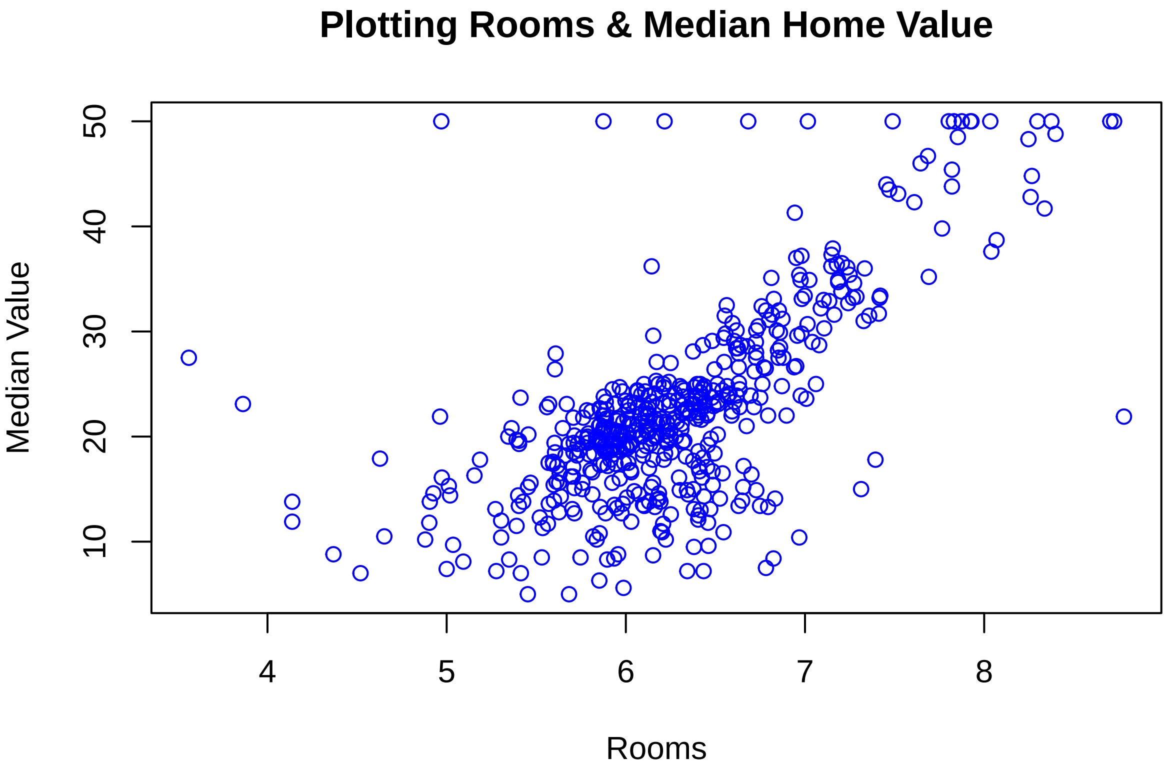
### Step 6: Finding potential correlations

Create a plot showing the median value on the y axis and number of rooms on the x axis. Create appropriate main, x and y labels, change the point color and style. [Reference for plots(http://www.statmethods.net/advgraphs/parameters.html)

Use the cor() function to quantify the correlation between these two variables. Write a sentence or two summarizing what the graph and correlation tell you about these 2 variables.

Your commentary here: Based off the plotted graph I can tell that theres a lot more points landing between 6-7 rooms. I can also tell that more rooms increases the value, but that also depends wether or not if its in a good area. Overall the value of home correleates to the rooms it has.

```
# step 6 code
# x = rooms, y = medianValue
plot(Boston$rm, Boston$medv, pch=1, cex=1, col="blue", main="Plotting Rooms & Median Home Value", xlab=
```



```
# Cor room and medv
cor(Boston$rm, Boston$medv)
```

```
## [1] 0.6953599
```

### Step 7: Evaluating potential predictors

Use R functions to determine if variable chas is a factor. Plot median value on the y axis and chas on the x axis. Make chas a factor and plot again.

Comment on the difference in meaning of the two graphs. Look back the description of the Boston data set you got with the ?Boston command to interpret the meaning of 0 and 1.

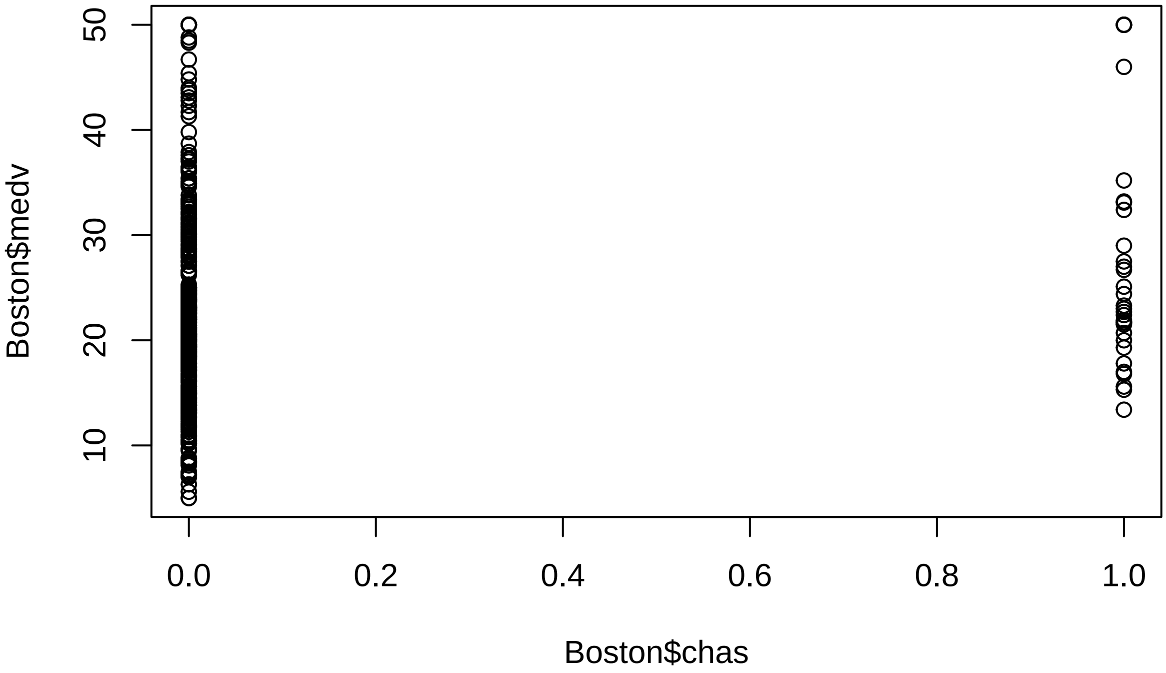
Your commentary here: The graphs came out the same after I converted the chas column to a factor. I noticed that the 1 is if the tract bounds river and 0 is otherwise.

```
# step 7 code

  # checks if the column CHAS is a factor and returns boolean.
is.factor(Boston$chas)
```

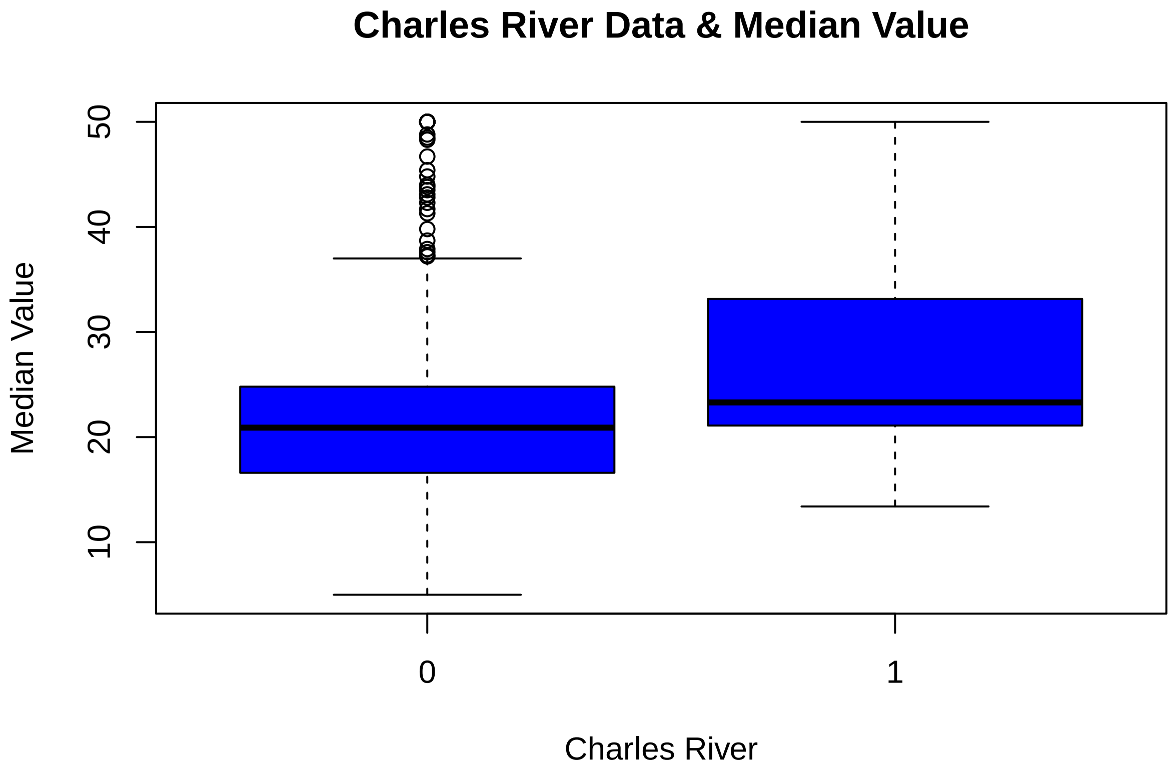
```
## [1] FALSE
```

```
# PLOTTING BEFORE CONVERTING IT TO A FACTOR
plot(Boston$chas, Boston$medv)
```



```
    # CONVERT CHAS TO A FACTOR
Boston$chas <- as.factor(Boston$chas)

  #PLOTTING AFTER CONVERTING CHAS INTO A FACTOR
plot(Boston$chas, Boston$medv, pch=1, cex=1, col="blue", main="Charles River Data & Median Value", xlab=
```



### Step 8: Evaluating potential predictors

Explore the rad variable. What kind of variable is rad? What information do you get about this variable with the summary() function? Does the unique() function give you additional information? Use the sum() function to determine how many neighborhoods have rad equal to 24. Use R code to determine what percentage this is of the neighborhoods.

Your commentary here: Rad type is an integer and Rad is the index of accessibility to radial highways. When running the summary function on the rad column I receive the min and max along with median/mean.Unique returns a vector of the data frame while removing duplicates. Upon running the sum function to find the number of neighborhoods equivalent to 24 is 132.

```
# step 8 code
typeof(Boston$rad)
```

```
## [1] "integer"
```

```
summary(Boston$rad)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   1.000   4.000   5.000   9.549  24.000  24.000
```

```
unique(Boston$rad)
```

```
## [1]  1  2  3  5  4  8  6  7 24
```

```
# initially was setting the values to variables.
#sum_of_rad <- sum(Boston$rad == 24)
sum(Boston$rad == 24)
```

```
## [1] 132
```

```
# store num of rows into variable
#num_rows <- nrow(Boston)

#percentage_rad <- Boston$rad/(sum_of_rad)

sum((Boston$rad == 24) / nrow(Boston) * 100)
```

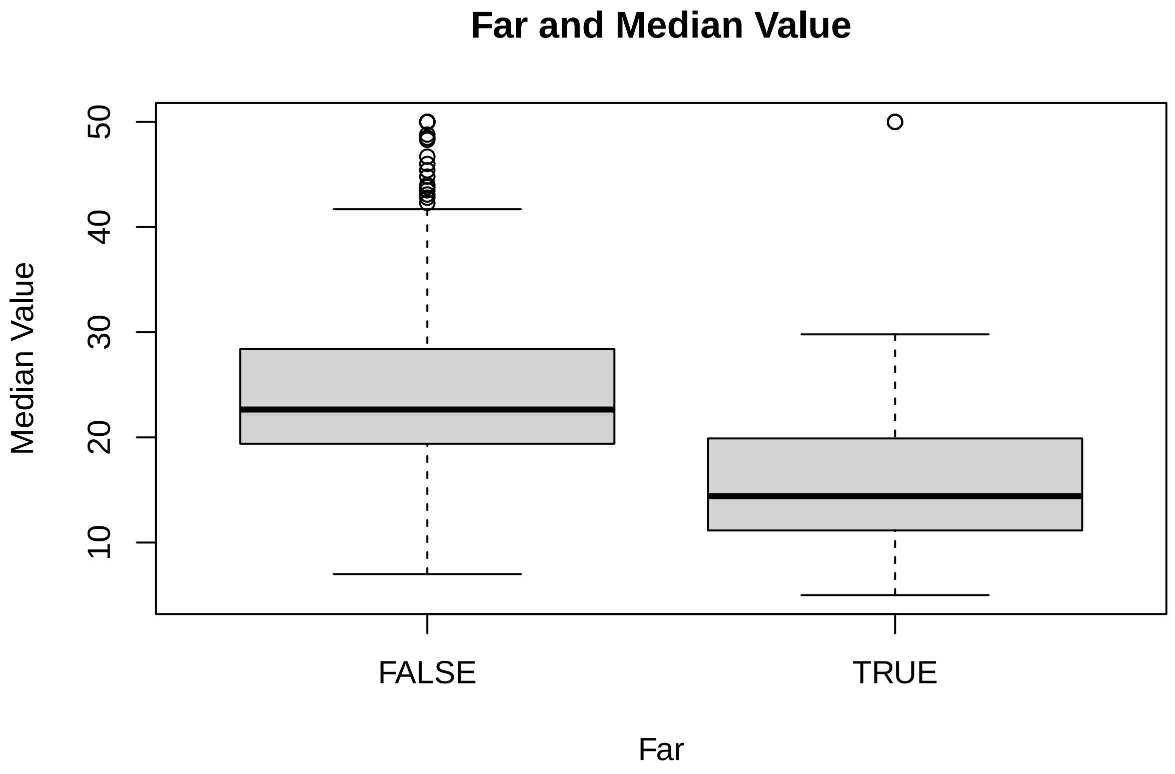
```
## [1] 26.08696
```

### Step 9: Adding a new potential predictor

Create a new variable called "far" using the ifelse() function that is TRUE if rad is 24 and FALSE otherwise. Make the variable a factor. Plot far and medv. What does the graph tell you?

Your commentary here: Based on the graph I can tell that the median value is cheaper if you're further away from the highways versus being closer to them. closer were are and have access to highways the value increases probably due to the areas.

```
# step 9 code
Boston$far <- ifelse(Boston$rad == 24, TRUE, FALSE)
Boston$far <- as.factor(Boston$far)
plot(Boston$far, Boston$medv, main="Far and Median Value", xlab="Far", ylab="Median Value")
```



### Step 10: Data exploration

- Create a summary of Boston just for columns 1, 6, 13 and 14 (crim, rm, lstat, medv)
- Use the which.max() function to find the neighborhood with the highest median value. See p. 176 in the pdf
- Display that row from the data set, but only columns 1, 6, 13 and 14
- Write a few sentences comparing this neighborhood and the city as a whole in terms of: crime, number of rooms, lower economic percent, median value.

Your commentary here: Retrieved the highest median house value and noticed that it's crime rate compared to the others is a lot lower which makes sense. The amount of rooms the houses have are pretty huge compared to the other data. Based off that I can deduce that this neighborhood is on the wealthy side which explains the lowered crime rate.

```
# step 10 code

# Summary of boston with selected columns (1,6,13,14)
summary(Boston[,c(1,6,13,14)])
```

```
##       crim                rm            lstat             medv      
##  Min.   : 0.00632   Min.   :3.561   Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 0.08205   1st Qu.:5.886   1st Qu.: 6.95   1st Qu.:17.02  
##  Median : 0.25651   Median :6.208   Median :11.36   Median :21.20  
##  Mean   : 3.61352   Mean   :6.285   Mean   :12.65   Mean   :22.53  
##  3rd Qu.: 3.67708   3rd Qu.:6.623   3rd Qu.:16.95   3rd Qu.:25.00
##  Max.   :88.97620   Max.   :8.780   Max.   :37.97   Max.   :50.00
```

```
# which.max returns the position of the lement with the max value.
# assign which to var
max_var = which.max(Boston$medv)
Boston[max_var,c(1,6,13,14)]
```

```
##        crim    rm lstat medv
## 162 1.46336 7.489  1.73   50
```

## Part 2: C++

In this course we will get some experience writing machine learning algorithms from scratch in C++, and comparing performance to R. Part 2 of Homework 1 is designed to lay the foundation for writing custom machine learning algorithms in C++.

To complete Part 2, first you will read in the Boston.csv file which just contains columns rm and medv.

---

In the C++ IDE of your choice:

1 Read the csv file (now reduced to 2 columns) into 2 vectors of the appropriate type.

2 Write the following functions:

- a function to find the sum of a numeric vector
- a function to find the mean of a numeric vector
- a function to find the median of a numeric vector
- a function to find the range of a numeric vector
- a function to compute covariance between rm and medv (see formula on p. 74 of pdf)
- a function to compute correlation between rm and medv (see formula on p. 74 of pdf); Hint: sigma of a vector can be calculated as the square root of variance(v, v)

3 Call the functions described in a-d for rm and for medv. Call the covariance and correlation functions. Print results for each function.