Homework 1

4375 Machine Learning with Dr. Mazidi

Sanika Buche

8/31/21

This homework has two parts:

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

This homework is worth 100 points, 50 points each for Part 1 and Part 2.

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: The Boston data set has to do with Housing values in Boston suburbs. After typing ?Boston into the console, we can see that it has 506 rows and 14 columns. We can also see that the study includes everything from how much homeowners earn, to taxes to the number of crimes in the neighborhood.

```
# step 1 code
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 ...
```

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
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
```

```
tail(Boston, 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
```

```
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
```

```
head(Boston$crim)
```

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

```
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(Boston$crim, na.rm=TRUE)
```

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

```
median(Boston$crim)
```

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

```
range(Boston$crim)
```

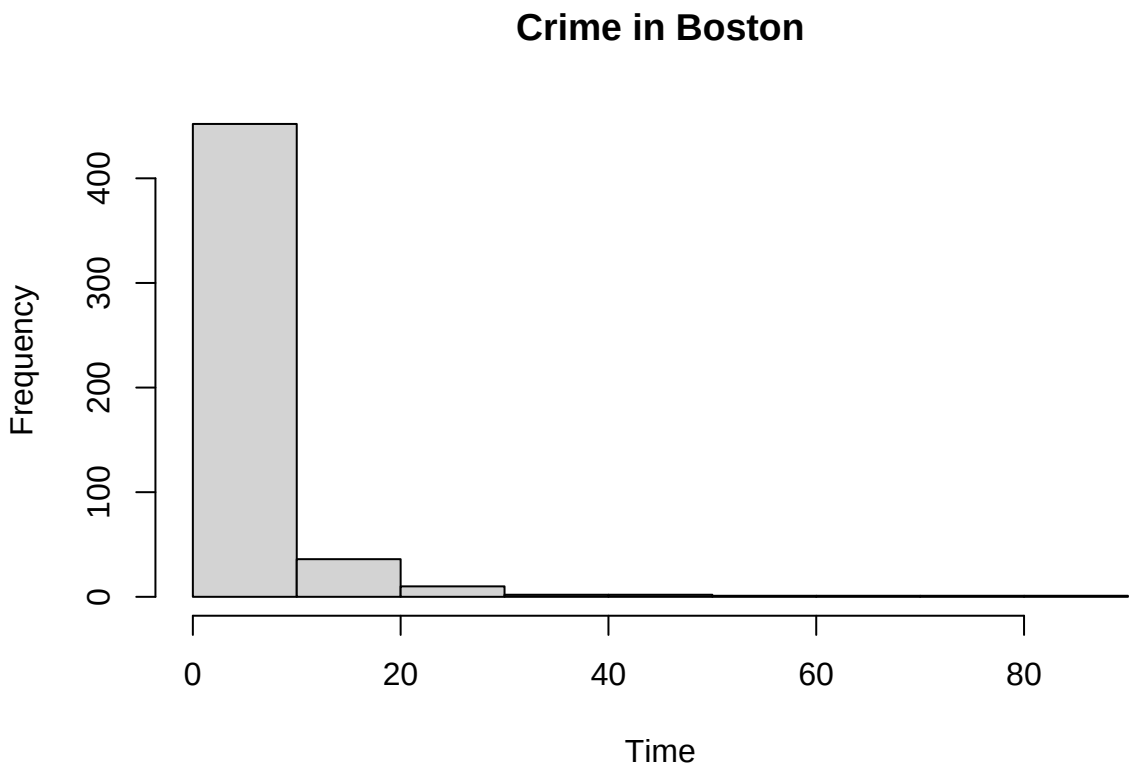
```
## [1]  0.00632 88.97620
```

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: Crime has gone down significantly in Boston as time has passed. Although the histogram may be hard to read since the intial bar in the graph is too big it is easy to see as time is progressing Boston is becoming a safer city to live in.

```
# step 4 code
hist(Boston$crim, main = "Crime in Boston", xlab = "Time")
```



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: The output number is negative which means that the two values have a negative impact on each other. Meaning as crime goes down the price of houses go up the same amount and as crime goes up the price of the houses go down. I think the crime column is useful in predicting median home value since the correlation value is pretty close to 0

```
# step 5 code
cor(Boston$medv, Boston$crim)
```

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

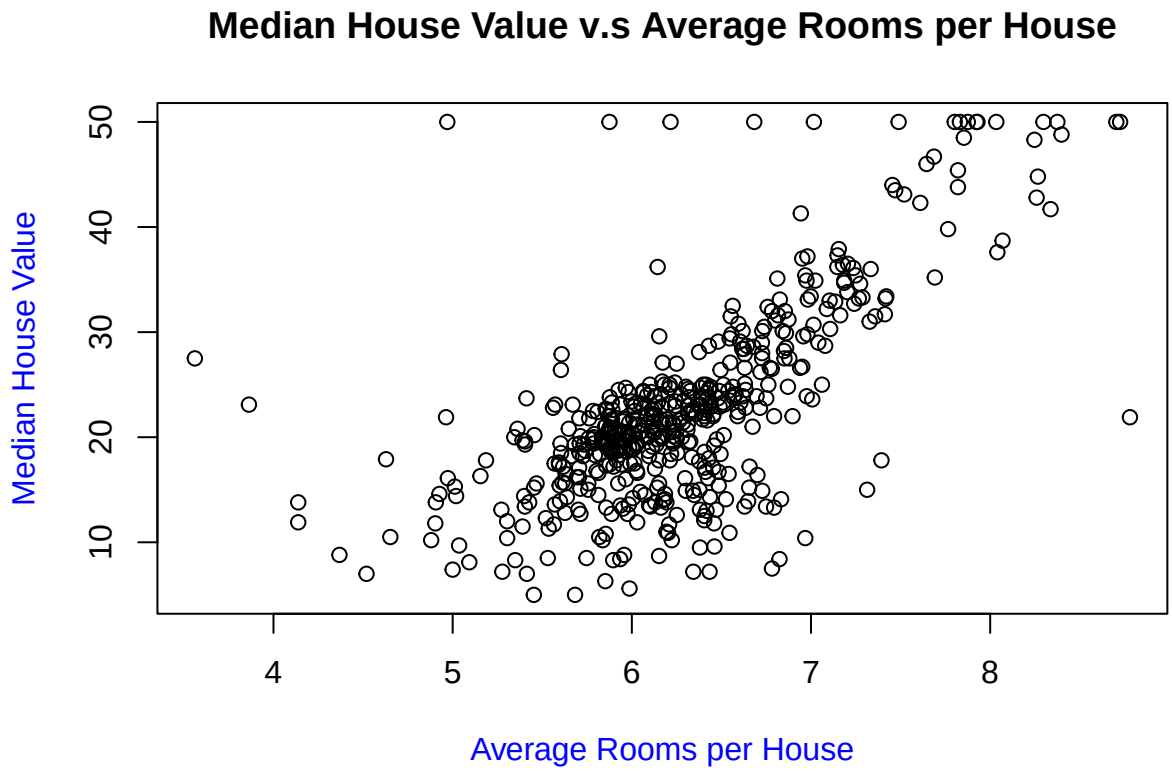
Step 6: Finding potential correlations

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: The correlation value of the number of rooms and the price of the house is positive and close to 1. This means that they both have a very strong linear relationship meaning if there is less rooms in the house the value of the house will fall and vice versa. The graph shows a clear positive linear relationship confirming what the correlation already said.

```
# step 6 code
plot(Boston$rm, Boston$medv, col.lab="blue", main = "Median House Value v.s Average Rooms per House", xl
```




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

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

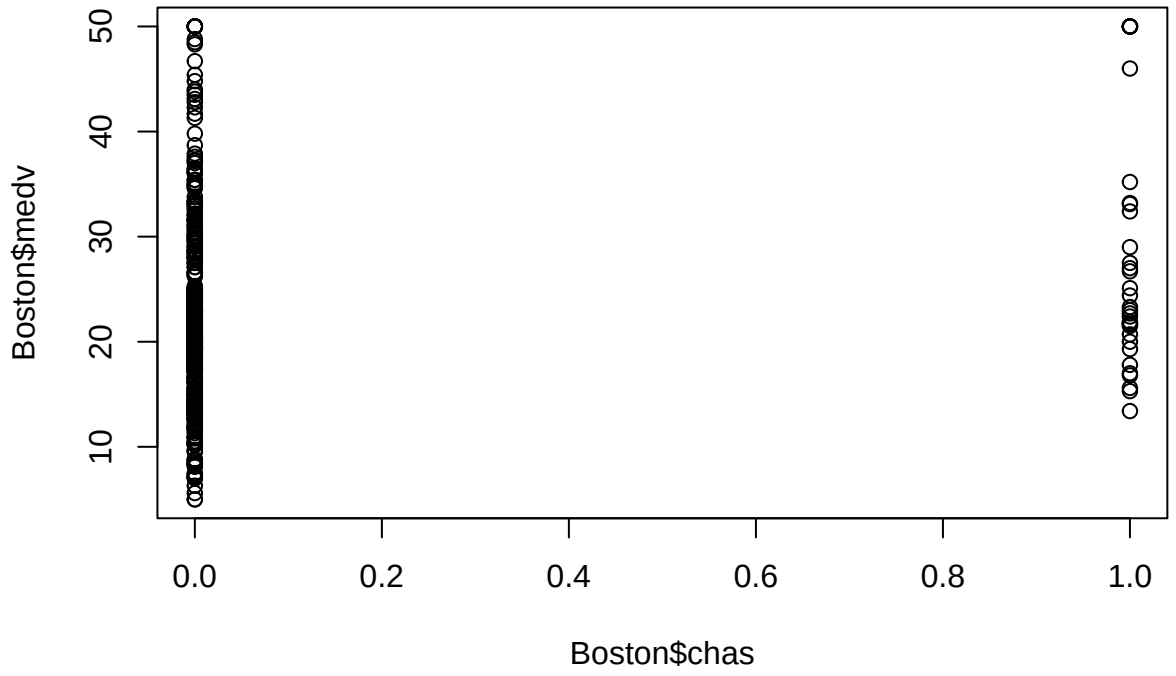
Step 7: Evaluating potential predictors

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: Chas stands for the Charles River. The 1 is for if the property is on the river and 0 is for if the property does not touch the river.

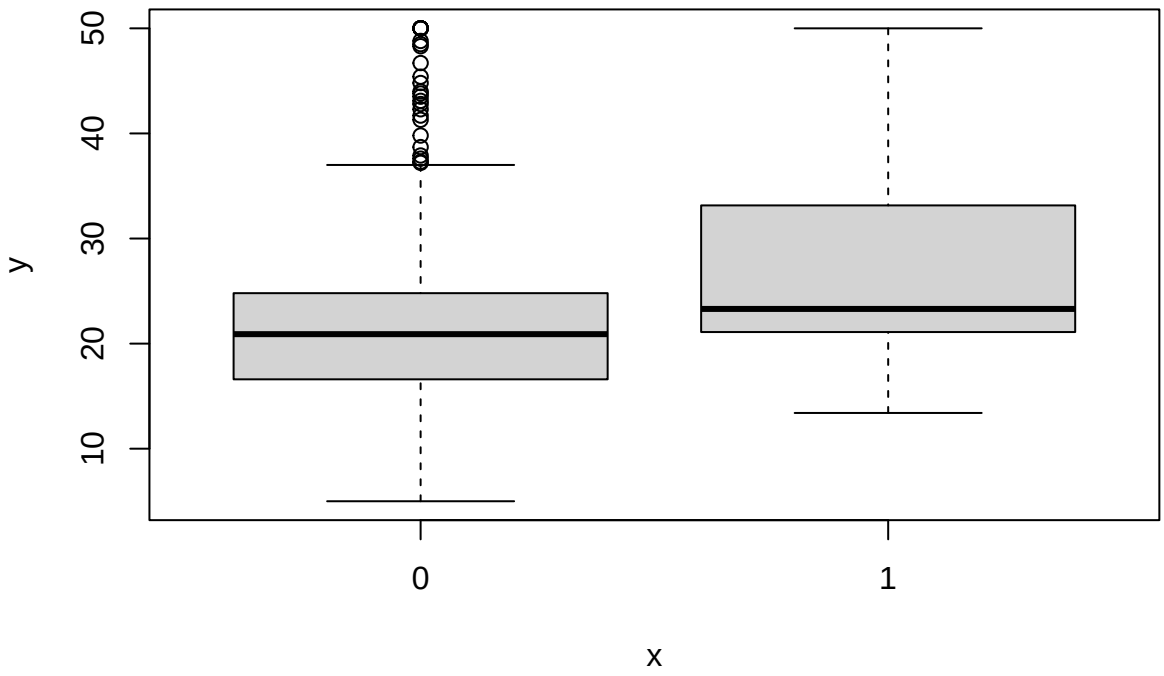
```r
# step 7 code
plot(Boston$chas, Boston$medv)
```




```r
#make it factor
Boston$chas <- as.factor(Boston$chas)
contrasts(Boston$chas)
```

```
##   1
## 0 0
## 1 1
```

```r
plot(Boston$chas, Boston$medv)
```



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: From the summary function rad looks like it is statistics of something after typing typeof(Boston$rad) we can see that it is an integer. From doing ?Boston and scrolling down I can tell that summary(Boston$rad) prints the distance from the houses to the highway. It prints the minimum distance as well as the median, mean and max distance each house is from the highway. The unique function doesn't really give any different information as it just outputs numbers. In order to find the proportion we have to take the sum of the neighborhoods with rad equal to 24 and divide it by the length of the rad column in the Boston dataset and then multiply the value by 100.

```
# 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
```

```
sum(Boston$rad==24, na.rm=TRUE)
```

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

```
prop <- sum(Boston$rad==24, na.rm=TRUE)/length(Boston$rad)
per <- prop*100
```

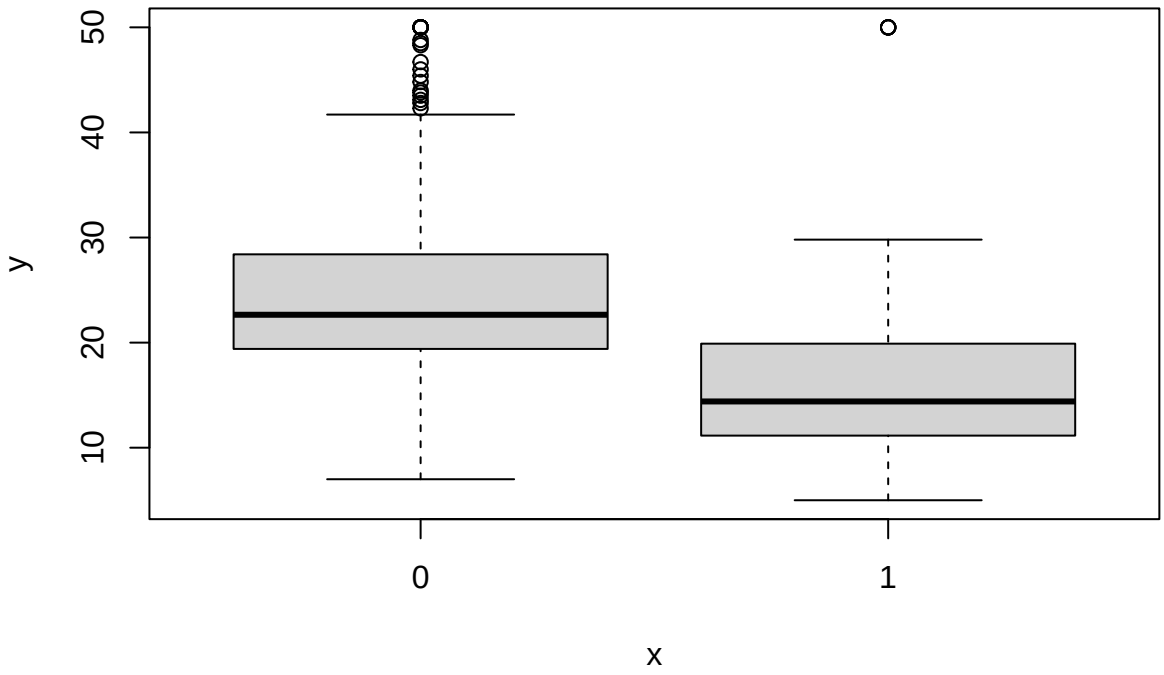
Step 9: Adding a new potential predictor

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: This graph tells me that

```
# step 9 code

far <- ifelse(Boston$rad==24, 1, 0)
far <- factor(far)
plot(far, Boston$medv)
```



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: Based on the data it looks like there is a lot of crim in boston but only in specific parts of the city since the median is much less than the max.The average number of rooms is more consistant with this neigborhood compared to the rest of Boston. The same can't be said for the lower economic percent and median value. The max numbers and median/mean numbers are much different.

```
# step 10 code
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
```

```
i <- which.max(Boston$crim)
print(Boston[i,1])
```

```
## [1] 88.9762
```

```
i <- which.max(Boston$rm)
print(Boston[i,6])
```

```
## [1] 8.78
```

```
i <- which.max(Boston$lstat)
print(Boston[i,13])
```

```
## [1] 37.97
```

```
i <- which.max(Boston$medv)
print(Boston[i,14])
```

```
## [1] 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.