Homework 1

4375 Machine Learning with Dr. Mazidi

Tyler Echols

6/2/2022

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: # For the housing values of boston it is divided into a list with 506 row, and 14 columns. The data also contains various abriviated terms for eacch assingned value given to the terms of the houses representing the price value of the reported number

```
# 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, 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

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[,1])

## [1] 0.00632 0.02731 0.02729 0.03237 0.06905 0.02985

names(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

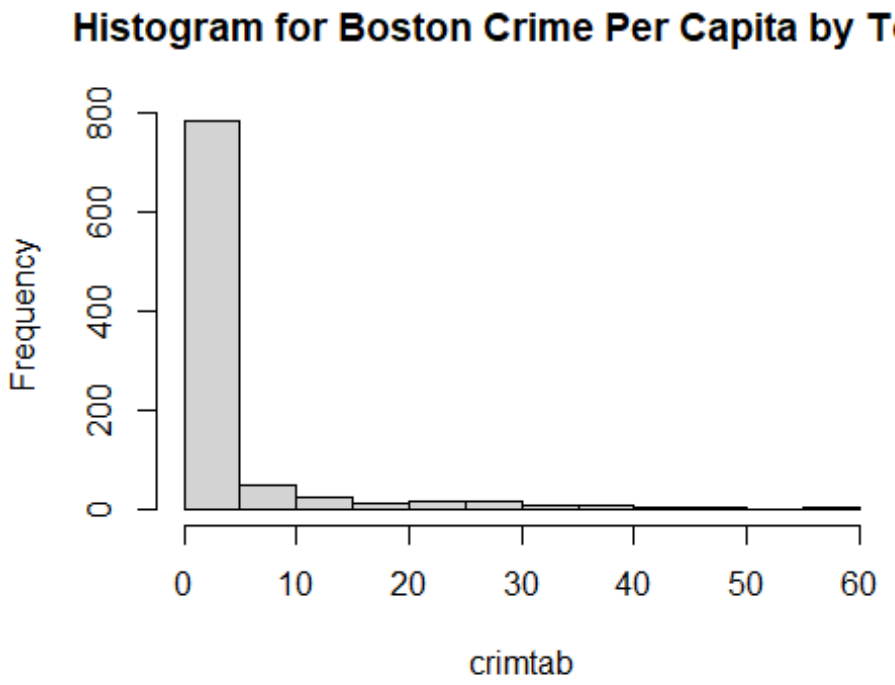
```
# step 3 code
mean(crimtab)

## [1] 3.246753

median(crimtab)

## [1] 0

range(crimtab)

## [1]  0 58
```

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: # With the hist feature it can create a histogram of any data it's reading in. If you want it to be more specific in how detailed you want the numbers to be displayed you can tell it to specifically read from 1 column.

```
# step 4 code
hist(crimtab, main="Histogram for Boston Crime Per Capita by Town" )
```



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 correlation between crim and medv is -0.30. meaning that the rate is a a negative linear relationship

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

## [1] -0.3883046
```

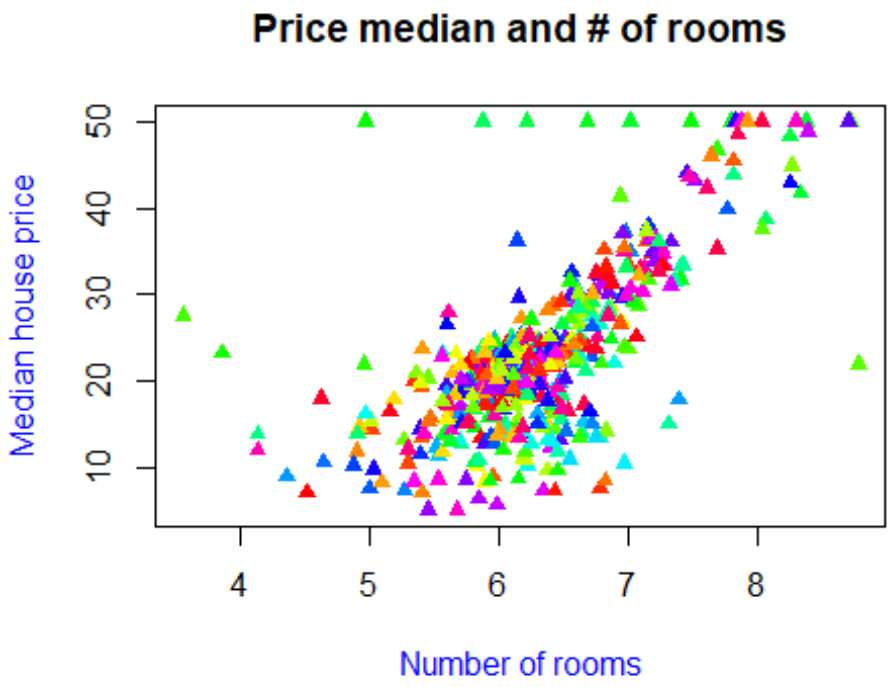
Step 6: Finding potential correlations

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: # It shows that a majority are in the 20 range where there are houses with 6 of less rooms.

```
# step 6 code
plot(medv ~ rm, data = Boston, main = " Price median and # of rooms ", xlab =
"Number of rooms", ylab = "Median house price" , col = rainbow (69) , pch =
17 , col.lab = "Blue")
```




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

## [1] 0.6953599
```

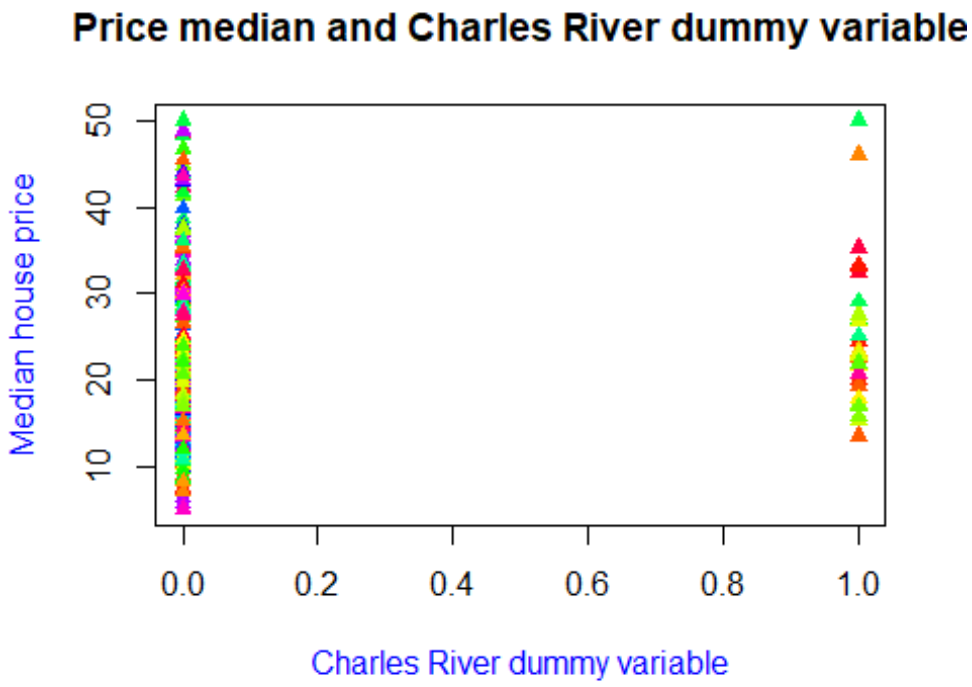
Step 7: Evaluating potential predictors

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 fact it seems that both are at perfect points on the graph with no in the middle valuse. They are all either 0.0 or 0.1.

```
# step 7 code

plot(medv ~ chas, data = Boston, main = " Price median and Charles River
dummy variable ", xlab = "Charles River dummy variable", ylab = "Median house
price" , col = rainbow (69) , pch = 17 , col.lab = "Blue")
```




```
is.factor(Boston$chas)

## [1] FALSE

?Boston

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

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: # You get information about how accessible it is to each radial highway. Also listing the various numbers of neighbors

```
# step 8 code
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)

## [1] 4832
```

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?

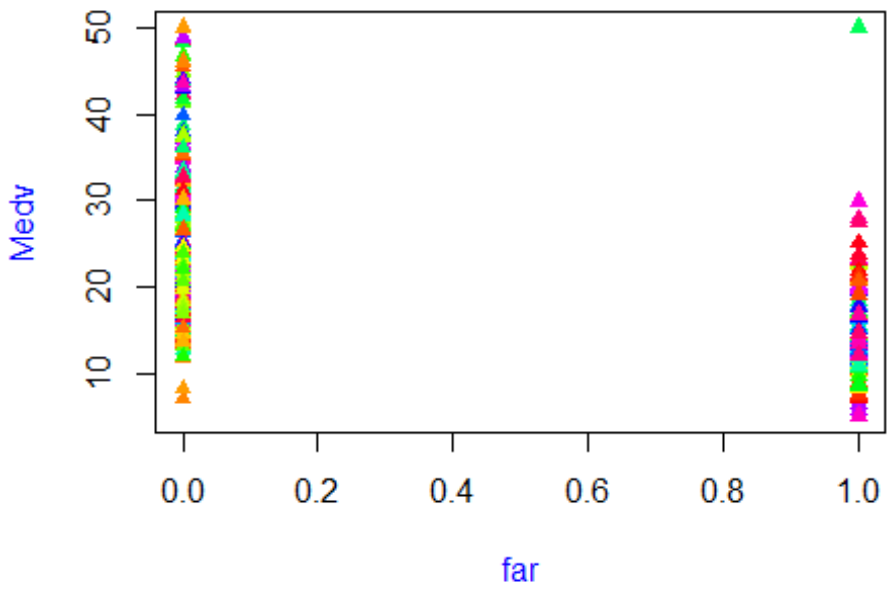
Your commentary here: # This graph tells me that everything is high on when you look at 0 for far. But when you look at the one for 1.0 it seeems like it is more twoard the lower 30's and 20's.

```
# step 9 code
ifelse(Boston$rad == 24, far<- TRUE, far<-FALSE)

##   [1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [25] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [37] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [49] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [61] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [73] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [85] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
##  [97] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [109] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [121] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [133] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [145] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [157] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
```

```
FALSE
## [169] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [181] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [193] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [205] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [217] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [229] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [241] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [253] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [265] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [277] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [289] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [301] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [313] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [325] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [337] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [349] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE  TRUE
TRUE
## [361]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [373]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [385]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [397]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [409]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [421]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [433]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [445]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [457]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

```
TRUE
## [469]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
TRUE
## [481]  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE
FALSE
## [493] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
FALSE
## [505] FALSE FALSE
```

```
far <- ifelse(Boston$rad == 24, far<- TRUE, far<-FALSE)
plot(medv ~ far, data = Boston,  xlab = "far", ylab = "Medv" , col = rainbow
(69) , pch = 17 , col.lab = "Blue")
```



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: # it looks as if the crime rate is low starting out but arounf the 3rd quater it picked up. the number of rooms per house had a slight increase but has seemed to hit a high value with the max number of rooms in houses to be high. with the increase of of

population maxing out at 37.97. So that about have the ammount of people living in this city which must mean this city is to expensive to live in.

```
# step 10 code
summary(Boston$crim)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##  0.00632  0.08204  0.25651  3.61352  3.67708 88.97620

summary(Boston$rm)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.561   5.886   6.208   6.285   6.623   8.780

summary(Boston$lstat)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.73    6.95   11.36   12.65   16.95   37.97

summary(Boston$medv)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    5.00   17.02   21.20   22.53   25.00   50.00

x <- which.max(Boston$medv)
print(Boston[x,c("crim", "rm", "lstat","medv" )])

##        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. See the reading in cpp picture in Piazza.

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.