COMP1 Your first R assignment for BUS212

Prof Carver

September 2018

## Introduction

This is an [R Markdown](http://rmarkdown.rstudio.com) Notebook. When you execute code within the notebook, the results appear beneath the code. This document has three purposes:

1. to demonstrate the use of RStudio’s *Notebook* capability
2. to illustrate/ review some useful r packages for preparing and plotting data
3. to explain the requirements of the first COMP assignment. You should be able to use this Notebook \*.rmd file as a guide to complete the assignment as well.

This document first uses the BostonHousing.csv data file from Chapter 3 of our text, and demonstrates a series of tasks and code chunks that you can use to complete the following tasks.

After examining the RMD file that created this document, YOU will complete an assignment using the data from Chapter 2.

## Your Assignment

After reviewing these examples you’ll again work with the WestRoxbury.csv data from Chapter 2 to answer 5 questions showsn here. This is an individual assignment; if you need help, talk to Prof. Carver or the TAs, but not to classmates.

Your assignment is to create an attractive, professional-quality report using an R Notebook in which you:

1. Read in the data
2. Create a titled and labeled scatterplot of Tax vs. Total Value, including a fitted line (simple regression with lm)
3. Report the formula that the City uses to compute the tax assessment on a property, AND explain why the Chapter 2 multiple regression model removed TAX from the model specification.
4. Create a side-by-side boxplot display showing variation in Total Value by the number of full bathrooms in a home (HINT: treat the number of bathrooms as a *factor*)
5. Comment on your impression of how property value varies with the number of bathrooms in a house.

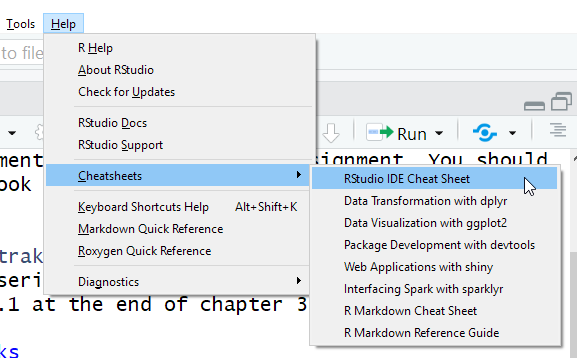
Note that your comments and written explanations are as important as the correctness of the two simple graphs.

One bit of background may be helpful: In Massachusetts, cities and towns raise much of their operating revenues by taxing real estate. Typically, the tax on a property is a fixed percentage of the market valuation. In the WestRoxbury dataset, the TAX column reflects the annual tax owed on each piece of real estate. In other words, TAX is a linear function of the property value.

The main goals of this assignment are to review your knowledge of R and to learn a bit about R Notebooks.

## R Markdown and Notebooks

Before going too far into this exercise, please review the online materials at [R Markdown](http://rmarkdown.rstudio.com). The pages at that site explain the general ideas, and also show specific syntax and commands for formatted text and for code chunks.

I highly recommend using RStudio’s “cheatsheets” for help with Markdown, Notebooks, ggplot2, dplyr and other useful packages. See the Help menu: 

Unlike the textbook, which calls packages as needed within a script, I recommend calling (library command) packages early on in a script. Either way works, but placing them all together allows a reader to find them easily.

Since library often generates messages and warnings that won’t add to my finished document, I recommend adding the warning = FALSE and message = FALSE in the chunk options:

library(tidyverse) # loads a number of helpful Hadley Wickham packages  
library(ggplot2) # way better than Base plotting  
library(readr) # allows to read csv files as "tibbles"  
library(tidyr) # newer replacement for package Reshape

## Boston Housing Example (adapted from Section 3.3 of our text)

# Use this as a guide for YOUR notebook code

In Chapter 3, Example 2 is about Amtrak ridership. We’ll reproduce the chapter 3 code chunks here, and add some more commands to illustrate why and how we might wish to “tidy” and plot the data. *NOTE*: In a markdown file, you *must* specify the entire file path when reading a csv file.

The BostonHousing data is published at the University of California, Irvine Machine Learning Repository [UCIMLR](https://archive.ics.uci.edu/ml/datasets/Housing); the original publication source is given in a footnote on p. 57, and each observation is one census tract in Boston.

All variables (columns) in the table are defined in *Table 3.1* on p. 58 of the text.

ALSO: here we demonstrate the readr command read\_csv and compare the resulting dataframe to the more conventional results of read.csv'. Packagereadr` is part of the “tidyverse” family of packages. Think of these as updates and improvements to some of the older packages.

housing.df <- read.csv("C:/Users/Rob/Box Sync/My R Work/BUS212/Data/BostonHousing.csv")  
  
head(housing.df,9) # top 9 rows of data, as in Table 3.2

## CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO LSTAT  
## 1 0.00632 18.0 2.31 0 0.538 6.575 65.2 4.0900 1 296 15.3 4.98  
## 2 0.02731 0.0 7.07 0 0.469 6.421 78.9 4.9671 2 242 17.8 9.14  
## 3 0.02729 0.0 7.07 0 0.469 7.185 61.1 4.9671 2 242 17.8 4.03  
## 4 0.03237 0.0 2.18 0 0.458 6.998 45.8 6.0622 3 222 18.7 2.94  
## 5 0.06905 0.0 2.18 0 0.458 7.147 54.2 6.0622 3 222 18.7 5.33  
## 6 0.02985 0.0 2.18 0 0.458 6.430 58.7 6.0622 3 222 18.7 5.21  
## 7 0.08829 12.5 7.87 0 0.524 6.012 66.6 5.5605 5 311 15.2 12.43  
## 8 0.14455 12.5 7.87 0 0.524 6.172 96.1 5.9505 5 311 15.2 19.15  
## 9 0.21124 12.5 7.87 0 0.524 5.631 100.0 6.0821 5 311 15.2 29.93  
## MEDV CAT..MEDV  
## 1 24.0 0  
## 2 21.6 0  
## 3 34.7 1  
## 4 33.4 1  
## 5 36.2 1  
## 6 28.7 0  
## 7 22.9 0  
## 8 27.1 0  
## 9 16.5 0

Now let’s see the structure and a glimpse of the tibble:

housing.tbl <- read\_csv("C:/Users/Rob/Box Sync/My R Work/BUS212/Data/BostonHousing.csv")

## Parsed with column specification:  
## cols(  
## CRIM = col\_double(),  
## ZN = col\_double(),  
## INDUS = col\_double(),  
## CHAS = col\_integer(),  
## NOX = col\_double(),  
## RM = col\_double(),  
## AGE = col\_double(),  
## DIS = col\_double(),  
## RAD = col\_integer(),  
## TAX = col\_integer(),  
## PTRATIO = col\_double(),  
## LSTAT = col\_double(),  
## MEDV = col\_double(),  
## `CAT. MEDV` = col\_integer()  
## )

head(housing.tbl,9) # Note additional metadata

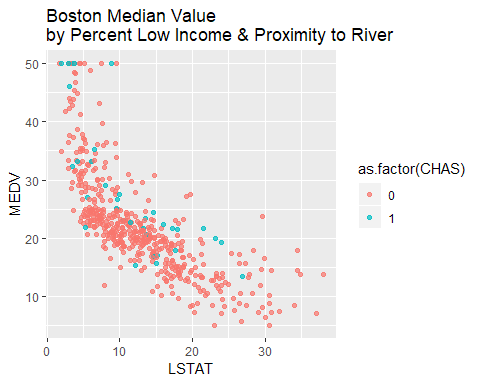
## # A tibble: 9 x 14  
## CRIM ZN INDUS CHAS NOX RM AGE DIS RAD TAX PTRATIO  
## <dbl> <dbl> <dbl> <int> <dbl> <dbl> <dbl> <dbl> <int> <int> <dbl>  
## 1 0.00632 18 2.31 0 0.538 6.58 65.2 4.09 1 296 15.3  
## 2 0.0273 0 7.07 0 0.469 6.42 78.9 4.97 2 242 17.8  
## 3 0.0273 0 7.07 0 0.469 7.18 61.1 4.97 2 242 17.8  
## 4 0.0324 0 2.18 0 0.458 7.00 45.8 6.06 3 222 18.7  
## 5 0.0690 0 2.18 0 0.458 7.15 54.2 6.06 3 222 18.7  
## 6 0.0298 0 2.18 0 0.458 6.43 58.7 6.06 3 222 18.7  
## 7 0.0883 12.5 7.87 0 0.524 6.01 66.6 5.56 5 311 15.2  
## 8 0.145 12.5 7.87 0 0.524 6.17 96.1 5.95 5 311 15.2  
## 9 0.211 12.5 7.87 0 0.524 5.63 100 6.08 5 311 15.2  
## # ... with 3 more variables: LSTAT <dbl>, MEDV <dbl>, `CAT. MEDV` <int>

# glimpse is a tidyr function to examine the structure of a tibble  
glimpse(housing.tbl)

## Observations: 506  
## Variables: 14  
## $ CRIM <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.029...  
## $ ZN <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 1...  
## $ INDUS <dbl> 2.31, 7.07, 7.07, 2.18, 2.18, 2.18, 7.87, 7.87, 7....  
## $ CHAS <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...  
## $ NOX <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0...  
## $ RM <dbl> 6.575, 6.421, 7.185, 6.998, 7.147, 6.430, 6.012, 6...  
## $ AGE <dbl> 65.2, 78.9, 61.1, 45.8, 54.2, 58.7, 66.6, 96.1, 10...  
## $ DIS <dbl> 4.0900, 4.9671, 4.9671, 6.0622, 6.0622, 6.0622, 5....  
## $ RAD <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4,...  
## $ TAX <int> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, ...  
## $ PTRATIO <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15...  
## $ LSTAT <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, ...  
## $ MEDV <dbl> 24.0, 21.6, 34.7, 33.4, 36.2, 28.7, 22.9, 27.1, 16...  
## $ `CAT. MEDV` <int> 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...

In the chapter, Figures 3.1 and 3.2 shows some simple graphs, and include both the base r plotting functions and code using ggplot2. Here we just use ggplot2, which I will encourage us to adopt as a standard for most of the course. This code chunk is adapted from the textbook, first creating a scatterplot similar to the one in the upper right of Figure 3.1. For added information, I’ve colored the points to indicate which tracts border the Charles River.

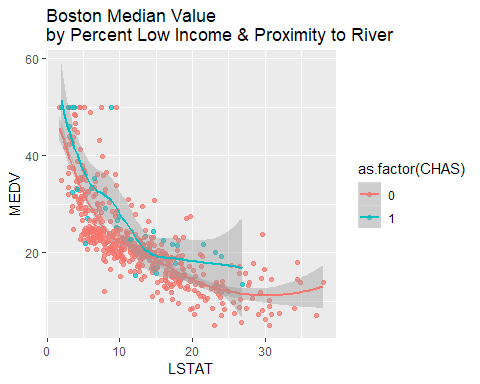
# Create the plot object, p, in layers and then display  
p <- ggplot(housing.tbl, aes(x=LSTAT, y=MEDV, color=as.factor(CHAS))) +  
 geom\_point(alpha = 0.7) # alpha controls the transparency of the points   
p + ggtitle("Boston Median Value \nby Percent Low Income & Proximity to River")



We might also add a smoother to the plot:

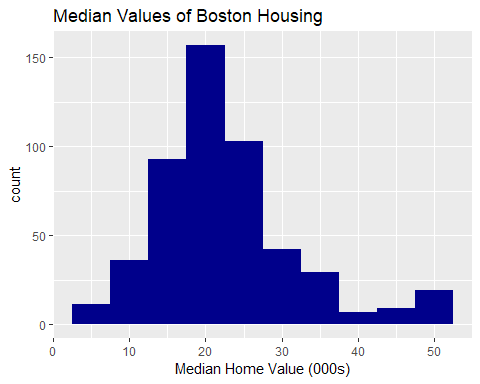
p + geom\_smooth() +  
 ggtitle("Boston Median Value \nby Percent Low Income & Proximity to River")

## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



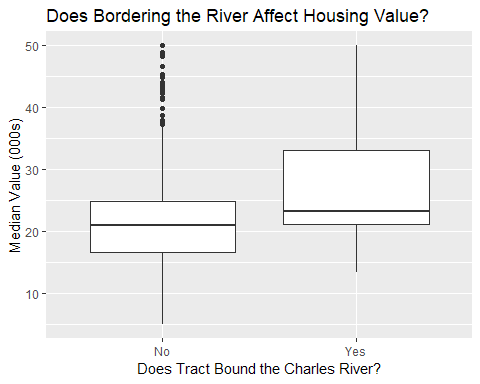
In Figure 3.2, we find two distributional graphs of MEDV. We start with a histogram:

hist <- ggplot(housing.tbl) +  
 geom\_histogram(aes(x=MEDV), fill="darkblue", binwidth = 5) +  
 ggtitle("Median Values of Boston Housing") +   
 xlab("Median Home Value (000s)")  
hist



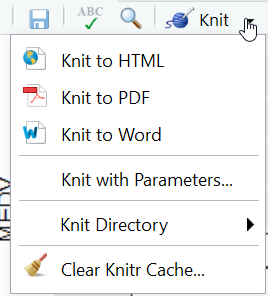
And then, the side-by-side boxplots of Median Values for properties that do and do not bound the Charles River. For improved labels, let’s create a new factor for the second variable, and assign descriptive level names:

housing.tbl$river <- factor(housing.tbl$CHAS, labels=c("No", "Yes"))  
bp <- ggplot(housing.tbl) +  
 geom\_boxplot((aes(x=river, y=MEDV))) +   
 xlab("Does Tract Bound the Charles River?") +  
 ylab("Median Value (000s)") +  
 ggtitle("Does Bordering the River Affect Housing Value?")  
bp



## Conclusion

There you are: an example of an R Notebook that weaves together text, graphs, and R Code using tidy tibbles and ggplot2!

After writing and testing the code, choose the knit button 

to knit it to a *Word* document, which you should then upload to LATTE along with your RMD file. In other words, upload TWO (2) files: your RMD and your Word.