ML 4375 – Intro to Machine Learning – Summer 2018 – Mazidi

Homework 1: Data Exploration

Objective: Learn how to use R for data exploration and become familiar with the R environment.

**Turn in:** an “.Rmd” Rstudio file to eLearning

**Important**: Start the file with comments for your name, Homework 1; Have prominent headings for the steps like “# 1” so the TA can find your code.

For each step below, place your code in the grey {r} boxes and comments and answers to questions in the white space above the code which you will label with a step number comment like # 1.

1. Load library MASS. The first time you use a library you will have to install it but do that at the console, not in your R code. After you load the library, look at the Environment pane in the upper right corner of RStudio. Notice that Boston is listed as <Promise>. When you load the package, R will be aware of the datasets in the package but won’t waste memory loading them unless you request the data. We want to use the Boston data set so load that into memory with data(Boston). Use the str() function to get an overview of the data. Type ?Boston *at the console* (not in your code) and you will see a description of the data set in the lower right hand corner of Rstudio. Write a brief 2-3 sentence description of the data set in the white text portion of your answer for #1.
2. Use R commands to:
   1. display the first few rows
   2. display the last 2 rows
   3. display row 5
   4. display the first few rows of column 1 by combining head() and indexing
   5. display the variable names
3. Use R statistical functions to find the mean, median, range of the crime column.
4. Create a histogram of the crime column, with an appropriate main heading. What does the histogram tell you about this variable?
5. Use the cor() function to see if there is a correlation between crime and the median house value. Comment on what this value might mean. How useful might the crime column be for predicting median value?
6. 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: <http://www.statmethods.net/advgraphs/parameters.html> Use the cor() function to quantify the correlation between these two variables. Write a sentence summarizing what the graph and correlation tell you about these 2 variables.
7. 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 the description of the Boston data set you got with the ?Boston command to interpret the meaning of 0 and 1.
8. 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.
9. 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?
10. 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. Display that row from the data set, but only colums 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.

|  |  |
| --- | --- |
| Element | Points |
| R script runs correctly | 40 |
| Appropriate comments and white space | 10 |
| Steps 1-10 | 5 points each |