Warm-Up 03 - Basic Data Manipulation

Stat 133, Fall 2018

Due date: Sep-18 (before midnight)

Introduction

The purpose of this assignment is to explore data collected for houses sold in Saratoga, New York in 2007. You will explore house prices, size of the house, and a few other variables such as the number of bedrooms and the number of fireplaces.

The main goal of this warmup is to give you practice working with differnt data types, subsetting, and creating vectors. A side benefit will be that you will gain practice thinking with data as you carry out your computations. The analysis of the data includes exmaning univariate distributions, exploring relationships between variables, and plotting.

General Instructions

- Write your narrative and code in an `Rmd` (R markdown) file.
- Name this file as `warmup03-first-last.Rmd`, where `first` and `last` are your first and last names (e.g. `warmup03-gaston-sanchez.Rmd`).
- Please do not use code chunk options such as: `echo = FALSE`, `eval = FALSE`, `results = 'hide'`. All chunks must be visible and evaluated.
- Submit your Rmd and html files to bCourses.
- If you have questions/problems, don't hesitate to ask us for help in OH or piazza.

Reading the Data Into R

The first task involves reading the data into R. We will provide the code to achieve this task. However, go through each of the following steps so that you begin to learn about the reading table functions in R.

- Look at the content of the data file, located in the github repository (see folder `data/`):

https://github.com/ucb-stat133/stat133-fall-2018/raw/master/data/saratoga.txt

- Examine the layout of the data. Do not download it to your computer.
  - Do the data have a header containing the variable names?
  - Are the values for an observation separated by a comma, blank, or tab?

- Read the documentation for `read.table()` and examine the parameters that are used to specify whether the data have a header and how the values in a row are separated.
- Without looking at the code below, try to write a command to import the data in R. If you can't do this, then use the commands shown below.
- After you imported the data, use `str()` to get a report of the dimensions of the data frame, as well as the class of each column.

```r
# assembling url so it fits on the screen
github <- 'https://github.com/ucb-stat133/stat133-fall-2018/'
repo <- 'raw/master/data/saratoga.txt'
house <- read.table(paste0(github, repo), sep = "\t")
```

Examine Price

We start by exploring house price. Make summary statistics of price. Remember that you can access a variable within a data frame with the `$` sign.

It's a bit surprising to see that the lowest price is only $5,000. Also notice that the mean price is about $25,000 more than the median price, which indicates that the distribution of price may be skewed right, possibly by having a long right tail.

Graph a histogram of house price with `hist()`. No need to give fancy labels at this point; we're just exploring the data. Be sure to specify enough bins to see the shape of the data. Indeed, we do find that the prices are skewed to the right with relatively few houses priced above $400K. That skewness explains why the mean is higher than the median. The few houses priced between $400K and $800K have pulled it away from the typical price.

Does a log transformation make the price distribution more symmetric? Log-transform price and graph another histogram. Now we see those unusually small values cropping up on the left. However the rest of the distribution looks symmetric and unimodal.

Let's use subsetting to dig a little deeper into the issue of the unusually inexpensive houses. We will answer the questions:

- How many are there?
- Do they have unusual values for the other variables?

To answer the first of these questions, create a logical vector that indicates if the log of the house price is under 10.5. Assign this logical vector to the variable named 'cheap'.

How many houses are what we call cheap?

Use logical subsetting to print the values of all of the variables for the cheap houses.

These houses do not appear to have unusual values for the other variables.

Use the variable `cheap` to (logically) subset the data frame house and remove these rows from the data frame. Call the new data frame `house`; that is, replace the existing data frame with the new one.

Explore Living Area

Next we explore the size of the house, i.e., the living area. Do you find a similar issue with the living area? Examine the summary statistics for living area and make a histogram of the values for living area and another histogra for the logarithm of these values.

We again find that the distirbution is skewed to the right. The log transformation of the values has a less peaked distribution that is roughly symmetric, but there appears to be an unusually large number of values at about 6.8. What value does this correspond to in the original units of measurement?

Transforming Variables into Factors

We next transform the number of bedrooms into a factor vector, i.e., into a categorical variable. We want to collapse the number of bedrooms into four categories, 2 or fewer, 3, 4, and 5 or more.

To do this, follow the following approach:

- Assign the variable *Bedrooms* in *house* to a vector called *BR*.
- Use subsetting to set all of the values in *BR* that are greater than 5 to 5.
- Use the same approach to set all of values in *BR* that are under 2 to 2.
- Use the *factor()* function to convert *BR* to a factor vector. Read carefully about the levels and labels parameters to the function. For the categories, use the following strings: “2- BR”, “3 BR”, “4 BR”, and “5+ BR”. Assign the factor to *BR*, i.e., over write the numeric *BR* with the factor *BR*. It will be helpful to use the “:” function and the *c()* function to create the arguments to the function call.

How does a summary of *BR* look different than a summary of *Bedrooms*?

Plotting Price against Living Area

We will now make a plot of price against living area (both on the log scale) and color the plotting symbols according to the number of bedrooms in the house. We will use the new vector *BR* to determine the color.

Begin by making a vector of 3 colors. Use the colors called `aquamarine3`, `darkgoldenrod2`, `coral2`, and `mediumorchid3`, in that order. Call this vector, `my_colors`.

Next make vector of colors that matches the length of *BR*, and where a value is "2- BR", the color is `aquamarine3`, where a value is "3 BR", the color is darkgoldenrod2, etc. Use subsetting by position of the vector `my_colors` to do this. Call this new vector, `br_colors`.

Lastly, we make the scatterplot as follows. Once you have created all of the necessary variables, i.e., BR, `br_colors` and `my_colors`.

```
plot( Price ~ Living.Area, data = house, log = "xy", main = "",
      xlab = "Living Area (log sq ft)", ylab = "Price (log $)",
      col = br_colors, pch = 19, cex = 0.4)

legend("bottomright", fill = my_colors, legend = levels(BR),
       title = "# Bedrooms", cex = 0.75)
```

Don't worry about the various arguments to the function call.

We see in the plot that price and living area have a roughly linear relationship and that houses with more bedrooms tend to command a higher price. This makes sense. It may be surprising the number of bedrooms is not more highly correlated with price and size of the house.

Number of Fireplaces

The original variable *Fireplaces* is the actual number of fireplaces in the house. Follow the approach in the previous section to create a new variable called *FP* that is a factor with just two levels: `"None"`,`"At least 1"`.

Create a new variable that contains the price per square foot of the house. Make a plot of price per square foot against price (determine if they should be on the log scale) and color the plotting symbols according to whether or not there is a fireplace in the house. To do this, create a vector of colors called `fp_colors`.

Are houses with fireplaces generally more expensive than houses without? What is the relationship between price per square foot and price? Can you explain why you might have this shape of relationship?