Homework 3

Statistical Computing, 36-350

Due Wednesday Sept 23, 2015

Your homework must be submitted in R Markdown format. We will not (indeed, cannot) grade homeworks in other formats. Your responses must be supported by both textual explanations and the code you generate to produce your result. (Just examining your various objects in the "Environment" section of R Studio is insufficient—you must use scripted commands.)

The data set at http://www.stat.cmu.edu/~ryantibs/statcomp/homework/capa11.csv contains information about the housing stock of California and Pennsylvania, as of 2011. Information is aggregated into "Census tracts", geographic regions of a few thousand people which are supposed to be fairly homogeneous economically and socially.

General hint: see Recipes 10.1 and 10.2 in "The R Cookbook" for making scatterplots, and 10.18 and 10.26 for general plotting.

1. **Loading and cleaning**
    a. Load the data into a data frame called `ca_pa`. Remember to pass the full URL of the website containing the data file to `read.table()` (instead of calling `read.table()` with a path to a local version of the file that you downloaded).
    b. Let `ca_pa_mat` be the result of casting `ca_pa` as a matrix, as below. What happens to numerical columns? (Hint: now try to plot a histogram of the column named `Median_rooms`.) What difference between the two data types matrix and data frames does this highlight?

```
ca_pa_mat = as.matrix(ca_pa)
```

    c. Run this command, and explain, in words, what it returns (or does):

```
apply(ca_pa,c(1,2),is.na)
```

    d. Run this command, and explain, in words, what it returns (or does):

```
colSums(apply(ca_pa,c(1,2),is.na))
```

    e. Now, let's say we want to deal with these missing values. One way to do this is to *purge* the data set of rows with incomplete data. The function `na.omit()` takes a data frame and returns a new data frame, omitting any row containing an `NA` value. Using it, how many rows are eliminated? Use this new dataset to answer questions 2 through 5.
    f. Alternatively, implement `na.omit()` using your own control flow statements (for/while loops, or `apply()`). *Extra credit* for using `apply()` or its cousins, like `sapply()`.

2. **This very new house**
    a. The variable `Built_2005_or_later` indicates the percentage of houses in each Census tract built since 2005. Plot median house prices against this variable.
    b. Make a side-by-side plot that breaks this by state, with Pennsylvania on the left and California on the right. The plots should have the same axes limits, and should have the respective state names as titles. (Note that the state is recorded in the `STATEFP` variable, with California being state 6 and Pennsylvania state 42.) There is some plotting example code in the end.

3. **Nobody home**
   The vacancy rate is the fraction of housing units which are not occupied. The data frame contains columns giving the total number of housing units for each Census tract, and the number of vacant housing units.

   a. Add a new column to the data frame which contains the vacancy rate.
   b. How does vacancy rate differ for the two states? Produce histograms for vacancy rates in each of the two states.
   c. Plot the vacancy rate against median house value.
   d. Plot vacancy rate against median house value separately for California and for Pennsylvania. Is there a difference? Describe it!

4. **County investigation**
   The column `COUNTYFP` contains a numerical code for counties within each state. We are interested in Alameda County (county 1 in California), Santa Clara (county 85 in California), and Allegheny County (county 3 in Pennsylvania).

   a. Explain what the block of code at the end of this question is supposed to accomplish. Give a single line of R that produces the same output.
   b. For Alameda, and Allegheny Counties, what were the average percentages of houses built since 2005? Also produce histograms for the two counties, both using x axis limits between 0 and 60, using the `xlim` argument in `plot()`. What are some key differences in their distributions? (Hint: see next problem.)
   c. How many tracts in Alameda have percentages of houses built since 2005 higher than 30%? What are these values? Remove these (hint: six) tracts and take the average again, to see it drop drastically. Briefly comment on what you've learned.
   d. The `cor` function calculates the correlation coefficient between two variables. What is the correlation between median house value and the percent of housing built since 2005 in (i) the whole data, (ii) all of California, (iii) all of Pennsylvania, (iv) Alameda County, (v) Santa Clara County and (vi) Allegheny County?
   e. Make a single plot, showing median house values against median income, for Alameda, Santa Clara, and Allegheny Counties, clearly distinguishing the three counties. Add a plot legend using `legend()`. (This time, the function `points()`, and general plotting arguments like `pch`, `col`, `cex` are your friends!)

```
acca <- c()
for (tract in 1:nrow(ca_pa)) {
  if (ca_pa$STATEFP[tract] == 6) {
    if (ca_pa$COUNTYFP[tract] == 1) {
      acca <- c(acca, tract)
    }
  }
}
accamhv <- c()
for (tract in acca) {
  accamhv <- c(accamhv, ca_pa[tract,10])
}
median(accamhv)
```

Example plotting code
You may use following example code for plotting in 2b, 4b and 4e. The easiest way is to get started is to change some arguments slightly to see what happens. Detailed help can also be found in sections 10.18 or 10.26 in "The R Cookbook", or by typing `?plot` in the R console.

```r
# Tell R to draw plot in a 1 by 3 array
par(mfrow=c(1,3))
plot(x = 1:5, y = 1:5, col = "blue", cex = 1, xlim = c(0,6), ylim = c(0,6),
     type = 'p', main = "this title", xlab = "x", ylab = "y")
# Insert points on an existing plot
points(x = 1:3, y = (3:1)+.5, cex = 2, col = "green", pch = 17)
legend("topright", pt.cex = c(1,2), col = c("blue","green"), pch = c(1,17),
       legend = c("these blue points", "those green points"))
plot(x = 5:1, y = 1:5, col = "red", xlim = c(0,6), ylim = c(0,6), type = 'p',
     main = "that title", xlab = "x", ylab = "y")
legend("bottomleft", col = "red", pch=1, legend = c("more red points"))
hist(rnorm(100,1,2), xlim = c(-9,9), main = "Histogram example",
     xlab = "x label goes here", col = 'pink')
```