## 2 Modeling Runners' Times in the Cherry Blossom Race

1. Follow the approach developed in [the section called “Reading Tables of Race Results into R”](http://rdatasciencecases.org/exercises.html#sec:CB_ReadData) to read the files for the female runners and then process them using the functions in [the section called “Data Cleaning and Reformatting variables”](http://rdatasciencecases.org/exercises.html#sec:CB_CreateVariables) to create a data frame for analysis. You may need to generalize the *createDF()* and *extractVariables()* functions to handle additional oddities in the raw text files.
2. We have seen that the 1999 runners were typically older than the 2012 runners. Compare the age distribution of the runners across all 14 years of the races. Use quantile-quantile plots, boxplots, and density curves to make your comparisons. How do the distributions change over the years? Was it a gradual change?
3. Normalize each male runner's time by the fastest time for the runner of the same age. To do this, find the fastest runner for each year of age from 20 to 80. The *tapply()* function may be helpful here. Smooth these times using *loess()* , and find the smoothed time using *predict()* . Use these smoothed times to normalize each run time. Use density plots, quantile-quantile plots, and summary statistics to compare the distribution of the age-normalized times for the runners in 1999 and 2012. What do you find? Repeat the process for the women. Compare the women in 1999 to the women in 2012 and to the men in 1999 and 2012.
4. In [the section called “Scraping Race Results from the Web”](http://rdatasciencecases.org/exercises.html#sec:CB_WebScrape), we discovered that the **HTML** file for the male 2000 results was so poorly formatted that *htmlParse()* was unable to fix it to allow us to extract the text table from the <pre> tag. In this exercise, we programmatically edit this **HTML** file so that we can use *htmlParse()* as desired. To do this, begin by reading the **HTML** file located at <http://www.cherryblossom.org/cb003m.htm> using *readLines()* . Carefully examine the **HTML** displayed in [the section called “Scraping Race Results from the Web”](http://rdatasciencecases.org/exercises.html#sec:CB_WebScrape) and come up with a plan for correcting it. Consider whether you want to drop <font>s or close them properly. Once you have fixed the problem so that the <pre> tag contains the text table, pass your corrected **HTML** to *htmlParse()* . You may want to use a text connection to do this rather than writing the file to disk and reading it in.

He discusses the questions at 00:51:54 during the 9-11-19 class.