STA242 Assignment 1

Race Times - The Cherry Blossom Run
Due: Friday, April 10th 4pm.
Send an electronic version to Duncan, and put a printed version in Nick's box in the Statistics department main office.
Most recent version at bitbucket.org

This assignment involves reading data into R from a slightly non-standard format, exploring it (both numerically and graphically), identifying possible errors, summarizing the data, and finding and communicating aspects of interest. As part of the work, you will also work on matching individuals across years to create longitudinal data. This matching process occurs often and is challenging and fuzzy.

The data are results from the Cherry Blossom 10-mile running race, http://cherryblossom.org/. We have these separately for women and men, for years 1999 through 2010. (You can also get the data for 2011 and 2012, but 2013 and 2014 are in a different and less accessible format.) The data are in files named, e.g., women10Mile_1999 and men10Mile_2010. These are available within the course git repository, within the data/ directory.

Your assignment is to a) read the data into R, and b) explore different aspects of the data and offer insights. This is intentionally open-ended and not prescriptive. As statisticians and data scientists, I want you to be able to identify questions we can answer with this data and then explore some of these. The focus is exploratory data analysis and summarizing the major interesting aspects of the data. If you are drawn to fitting statistical models, do consider whether this is the most effective way to identify and summarize the insights, and also if they are appropriate (e.g., the assumptions for the models are satisfied).

You are to use git to manage your work. Put this in a private bitbucket.org repository so that it is not visible to others. You can share it with Nick and myself by setting the access rights to include us.

The Report

Submit a succinct report that summarizes your most interesting findings. This should explain why they are interesting and provide evidence for any conclusions, typically in the form of graphical displays. Take time to design these displays so that they are well-labeled, uncluttered, and effectively convey your conclusions. Also, discuss the limitations of your conclusions, i.e., other possible explanations and how likely they are, as well as how you might explore these further. The report should be 3 to 5 double-sided pages (no more), including tables and figures, but not including code. Reports that are disorganized or excessively lengthy will be penalized. Less is more - communicate the essential information, and do not elaborate for its own sake.

You are to present your code in the appendices as R scripts. Cutting-and-pasting direct input to the R console is not acceptable. Use short functions with comments and avoid repeating yourself. Appendix 1 should contain the code for reading the data into R, transforming the text to numbers, etc. and also validating the data.

This should include small functions for reading individual files. Appendix 2 should contain code for creating derived variables, cleaning the data and creating the summaries you included.

Getting Help

It is important to get started immediately on this project. There are several parts to it and many details.

You must write the code and perform the data analysis yourself. You may discuss ideas with each other, preferably via Piazza, but not exchange code. Please do not post large segments of code to Piazza that you are using to address a question. If you want help with problems with your code, simplify the code and post that. Alternatively, post it privately so that Nick and I can help. However, it is a very important skill to create a minimal example that mimics the problem. The process of doing this typically teaches you a lot and helps you to solve the problem yourself, so this is an essential part of the learning process.

Duncan Temple Lang <duncan@r-project.org>