Web Scraping of Cherry Blossom Runs

Shon Mohsin

June 1, 2019

## R Markdown

## Introduction

The results of the Cerry Blossom Ten Mile Run a available at <http://www.cherryblossom.org/>. For this paper we do a step-by-step analysis of how to web-scrape the run time data from this website. We outline the steps to get the data from the website, format the data for initial data manipulation. Afterwards, we answer Question 17 from Chapter 2 of Data Sciene in R; Nolan et al.

## Background

Question 17 of Data Science in R asks the following: “In Section 2.7, we discovered that the HTML ï¬le for the male 2000 results was so poorly formatted that htmlParse() was unable to ï¬x it to allow us to extract the text table from the

tag. In this exercise, we programmatically edit this HTML ï¬le so that we can use htmlParse() as desired. To do this, begin by reading the HTML ï¬le located at <http://www.cherryblossom.org/cb003m.htm> using readLines(). Carefully examine the HTML displayed in Section 2.7 and come up with a plan for correcting it. Consider whether you want to drop s or close them properly. Once you have ï¬xed the problem so that the

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 ï¬le to disk and reading it in."

TO accomplish this, we must fist scrape the data from the website. Then we edit the HTML file.

We then consider if we want to drop s or close them.

After this decision, we pass the corrected HTML to the htmlParse() function by using text connection.

## Method

We observe the data for the

## Results

## Conclusions

## Future Work

## References